CBESS Discussion Paper 09-16

Professional interpretation of the standard of proof: an experimental test on merger regulation

by Bruce Lyons,* Gordon Douglas Menzies and Daniel John Zizzo*****

* CCP, University of East Anglia
** University of Technology Sydney
*** CBESS, University of East Anglia

Abstract
There is considerable debate about the alternative economic approaches to merger control taken by competition authorities. However, differences in economic analysis are not the only reason for alternative decisions. We conduct an experiment in decision making in the context of merger appraisal, identifying the separate influences of different standards of proof, volumes of evidence, cost of error and professional training. The experiment was conducted on current practitioners from nine different jurisdictions, in addition to student subjects. We find that legal standards of proof significantly affect decisions, and identify specific differences due to professional judgment. We are further able to narrow the range of explanations for why professionalization matters.

JEL classification codes
L33, L40, L50, C91

Keywords
standard of proof; experiment; merger control

Acknowledgements:
The support of the Economic and Social Research Council is gratefully acknowledged. Thanks for excellent research assistance by Kei Tsutsui and Oindrila De. We would particularly like to thank agency heads at the following jurisdictions for their cooperation: Austria, Canada, Denmark, European Union, France, Hungary, Japan, Netherlands, Norway, Spain and the UK. This paper has benefitted greatly from comments by Bill Kovacic, and participants at a CCP seminar and the EARIE conference in Slovenia, September 2009. None of them can be held responsible for the views we express.

University of East Anglia
CBESS
Centre for Behavioural and Experimental Social Science

CBESS
University of East Anglia
Norwich NR4 7TJ
United Kingdom
www.uea.ac.uk/ssf/cbess

Contact details:
Bruce Lyons, School of Economics and the ESRC Centre for Competition Policy, University of East Anglia, Norwich, NR4 7TJ, United Kingdom.
b.lyons@uea.ac.uk

Gordon Douglas Menzies, School of Finance and Economics, University of Technology Sydney, PO Box 123, Broadway NSW 2007, Australia
gordon.menzies@uts.edu.au

Daniel John Zizzo, School of Economics, University of East Anglia, Norwich, NR4 7TJ, United Kingdom.
d.zizzo@uea.ac.uk

1. Introduction

The legal literature contains a long debate about how the civil standard of proof should be interpreted but very little evidence on how it is applied in practice. For example: are practitioners actually influenced by the formal standard, or does the wording matter little relative to 'gut feeling'? Inasmuch as wording matters, how are standards interpreted in the face of limited and conflicting evidence? A theoretical literature has related the standard of proof to other dimensions of the rules of evidence such as default rules, burden of proof and accuracy of evidence. This has been approached both from an economic perspective (e.g. Posner, 1973; Rubinfeld and Sappington, 1987; Davis, 1994) and that of the law (Vesterdorf, 2005). Much of the economic theory relates to finding rules that minimize the cost of error. The legal research has focused on sufficiency of evidence. Neither has provided a systematic empirical appraisal of what matters in practice.

The aim of this paper is to understand how legal decision making is influenced by professionalization and alternative standards of proof in the face of limited and conflicting evidence. We also want to understand how such decisions are affected by the volume of evidence, cost of error and experience. Merger control provides a flexible context because it has a natural two-phase setting. In the first phase, the merger may either be cleared or referred for further investigation.[1] In the second phase, the merger has a second chance of clearance or it may be prohibited with the consequence of long-term loss of efficiency if the prohibition decision was wrong. The latter cost of Type 1 error (i.e. cost of an adverse finding when there is, in truth, no competitive harm) is greater than the short term cost of unnecessary referral (also a Type 1 error).[2] The second phase also provides more time for qualitative and quantitative evidence to be gathered and processed. This institutional setting allows us to investigate differences in cost of error, amounts of evidence and standards of proof in a consistent framework. We analyze whether alternative standards of proofs make any practical difference to decision making; whether the interpretation of standards of proof is professionally determined; whether there is a connection between experience and toughness;

[1] In the USA, this takes the form of issuance of a Second Request. In the EC, this is called Phase II. In the UK, it takes the form of a reference by the OFT to the Competition Commission.
[2] The cost of Type 2 error (i.e. allowing an anticompetitive merger) is the same in both phases, apart from a few weeks' delay in implementation.

whether people take the volume of evidence and cost of errors into account; and whether there are systematic cognitive biases in their decision making.

We employ an experimental approach designed to understand how decisions are made in a legal context. Suppose someone is given the task of judging, in the absence of any specific context, whether A is true or false 'on the balance of probabilities'. They are given 100 pieces of evidence, each in a form that suggests A is either true or false. Each piece of evidence is of equal weight and subject to error. In a non-judicial context, it seems clear that the judgment should be that A is judged to be true if at least 51 pieces of evidence suggest this, and false if 49 or fewer pieces suggest truth. However, in many specific contexts, the problem changes if the decision maker starts from a prior belief or presumption. In the legal context, the burden of proof may imply a particular presumption. For example, if the burden is to prove that A is true, it is no longer obvious that 51 pieces of evidence suggesting truth will be sufficient to overturn the presumption that A is false even 'on the balance of probabilities'.

An experimental approach is innovative in our context and is particularly appropriate as there are severe limits to what can be learnt by investigating case data. Economic quantification of a large number of real cases is not possible because of their complexity, confidentiality of key data and the lack of natural experiments (e.g. changes in applicable standard of proof). Legal research can clarify the basis for reaching a specific decision and appeal courts reflect on cases where mistakes may have been made, but the complexity and idiosyncrasy of a few isolated appeals make it difficult to draw wider implications. This leaves the experimental approach as one which allows the systematic collection of data and potentially allows for careful control of complicating factors.

Experiments in economics most frequently use students as subjects. This is fine for understanding underlying human traits but it may introduce bias if professional training, organizational culture or self-selection in the relevant profession influence decisions taken in a judicial or quasi-judicial environment. The technical nature of the decisions to be made in our setting makes a professional sample especially useful to enhance what experimental economists label the *external validity* of our experiment, i.e. the extent to which the experimental results are likely to generalize to real world decisions. Because of this, we use practitioners from a number of competition authorities as subjects, supplemented by an

identical frame using student subjects, thus enabling us to verify whether the two subject pools behave differently.[3]

A general aspect of interest from our approach is that subjects exhibit conservative beliefs, i.e. people appear to hold beliefs in much the same way as scientists hold to a null hypothesis unless there is enough evidence to reject it (e.g., Edwards, 1968; Menzies and Zizzo 2007, 2009). In the presence of belief conservatism, the implied test significance level, which we call α, can be quite small even for apparently low standards of proof; i.e. there is a reluctance to reject the null hypothesis. In our decision tasks, belief conservatism takes the form of the strength of the presumption that the merger is 'innocent', i.e. harmless for competition.[4] By comparing the extent of this effect between students and professionals, our experiment tests the extent to which choices by professionals may be driven by a general psychological effect towards belief conservatism as opposed to a professionally informed evaluation of standards of proof. If, for example, both subject pools behave the same, this would be evidence for a general psychological mechanism rather than professional formation.

Our experiment is also of interest for understanding international differences in merger decisions taken by different jurisdictions. The debate usually focuses on the alternative economic approaches to merger control taken by competition authorities (e.g. Fox, 2002).[5] However, economic analysis is not the only reason for agencies to come to different decisions, as there may be differences in the law, discretion, organizational culture, experience, professional mix or standards of proof applied in different jurisdictions. Our methodology abstracts from all possible differences in economic analysis in order to focus on these other possible sources of difference. While our set of practitioners is not large, it is drawn from a wide range of competition authorities from 9 different jurisdictions.

Among our results, we find that practitioners are, on average, as belief conservative as students; i.e. they require similar amounts of evidence to shift them from the belief that the merger is harmless. Practitioners are, however, more discriminating in considering standards

[3] For a general discussion of the use of professional samples in economic experiments, see Friedman and Sunder (1994).

[4] Some of the behavioural evidence in other settings points in the opposite direction of under-valuing prior information (e.g. Goodie and Fantino 1996). This would yield to the opposite prediction in our context.

[5] This debate has typically been prompted by a few cases where judgement has been different on different sides of the Atlantic. Most recently, the high profile GE-Honeywell proposed merger was allowed by the US Department of Justice with only minor remedies, but prohibited by the European Commission. This spawned a large literature, such as Patterson and Shapiro (2001) and Ordover and Reynolds (2002).

of proof. Both practitioners and students are, other things being equal, less likely to decide against a merger if the decision is to prohibit it than if it is to refer it for further investigation, but the effect is much greater for practitioners. Practitioners are also particularly able to take into account the volume of the evidence they have, albeit not quite to the full extent of a professional statistician. Age makes practitioners slightly tougher, but the evidence is only marginally significant, and there is no evidence of other individual characteristics mattering (including economics versus legal training, or background legal system). Our results are strongly consistent with predicted high, medium or low standards of proof. However, we find that the 'accurate, reliable, consistent and sufficient evidence' standard emerging from a European Court of Justice decision[6] (ECJ, 2005) places a high standard of proof on merger decisions, indeed almost as a high as the 'beyond reasonable doubt' standard used in criminal courts.

The remainder of the paper is organized as follows. Section 2 provides the legal background and motivation for our selection of standards of proof. In section 3, we set out our experimental design, and our results are summarized and discussed in sections 4 and 5. Section 6 concludes.

2. Standards and Burden of Proof

2.1 Six Standards of Proof

Appendix A provides the legal background and phrasing of standards of proof used in the EU, USA and UK in the context of merger regulation. In relation to final decisions which might involve prohibition (which we shall refer to as stage 2 decisions) exact wordings differ but they boil down to whether it is expected that the merger is more likely than not harmful or, in evidentiary terms, where the preponderance of evidence lies. This is the common law civil standard that is applied to all commercial disputes. A lower standard is applied for deciding whether to refer a merger for an in-depth (stage 2) investigation. This stage 1 standard of proof is much less well developed, but is clearly meant to be lower than is applied in stage 2.

With this background we wished to select six alternative wordings for the standard of proof to be used in our experiment. Our choice was limited to six by the practicalities of

[6] ECJ. 2005. Tetra Laval Case C-12/03 P 15th February 2005 http://curia.europa.eu/jcms/jcms/j_6/.

experimental design.[7] We wanted half of the standards to be phrased in essentially probabilistic language (e.g. doubt, prospect, likelihood) and the other half to be phrased as weights of evidence. This is because probability concepts may be consider more difficult to understand than weights of evidence. We wanted at least one probabilistic and one evidentiary standard to reflect the practice used in major jurisdictions in each of stage 1 and stage 2 decisions. We also wanted to include the familiar criminal standard of 'beyond reasonable doubt' as a yardstick. This is expected to provide a higher threshold than the civil standard because of the seriousness of criminal penalties for individual freedom. A final requirement was that we wanted a wording that should be interpreted as closely as possible to 'half of the evidence plus one' in order to include a question against which we could provide an incentive to subjects (as discussed in section 3.2 below).

On this basis, subjects were asked to take action if and only if the merger would harm competition according to one of the following six standards: (1) it is proved beyond reasonable doubt; (2) it is likely; (3) there is a realistic prospect; (4) there is accurate, reliable, consistent and sufficient evidence; (5) the balance of the evidence is; (6) or the evidence raises a concern. We expect these to represent high, medium and low standards respectively, first expressed in probabilistic language and then as evidentiary burdens. We included standard 4 ('accurate, reliable, consistent and sufficient evidence', or ARCSE for short) because it was suggested in a landmark European legal judgment[8] as the appropriate stage 2 standard, whereas the view of some commentators was that it points to a high rather than medium burden of proof (see appendix A). Table 1 summarizes and indexes the six standards of proofs for future reference.

(Insert Table 1 about here.)

2.2 Burden of Proof

The essential approach of antitrust law is that firms are allowed to compete as they see fit unless they contravene competition law. For example, a firm may set whatever price it likes unless it has agreed that price (or market allocation, or another strategic variable) with one or more rivals, or possibly if it is acting in a predatory manner against an actual or

[7] The time we could ask of competition policy practitioners, and still get a sufficient number of responses, was limited, and so we aimed for the experiment to be completed in an average 45 minutes.
[8] See footnote 6.

potential rival.[9] This suggests a basic presumption in antitrust (as distinct from economic regulation) that the firm is free to choose unless the evidence suggests it has acted illegally. Put another way, the burden is on the agency to provide the proof, to the requisite standard, that the parties have acted illegally.[10]

Similarly, firms are allowed to merge unless this would lead to a substantial lessening of competition (or some other substantive test). It is not always obvious that this is the same as a presumption that a merger should be allowed unless the evidence suggests harm. For example, there may be a strong argument to reverse this presumption for merger to monopoly. However, the vast majority of proposed mergers are cleared without serious challenge.[11] This is consistent with a predominant presumption that a merger should be cleared unless the evidence suggests it will harm competition.[12] It is also consistent with a wider legal principle according to which 'he who asserts must prove' the affirmative (in this case, that the merger would be harmful).[13] We therefore build this burden of proof into our experimental design. In the language of hypothesis testing, the legal framework corresponds to a null hypothesis that the merger is harmless.[14]

3. Experimental Design

3.1 Experimental Treatments

[9] More controversially, in some jurisdictions a dominant firm may also be admonished for charging an exploitative price against its customers.

[10] Lawyers distinguish between the legal burden of proof and the evidential burden, but there is no difference in the context of this paper.

[11] Based on figures available on the DG Competition website, 88% of mergers qualifying for EC scrutiny (i.e. mergers of firms with combined turnover of €5b that do not sell at least 2/3 of their output in only one Member State) are cleared in Phase 1 without any required remedy or further investigation. With limited agency resources and the number of notifications being between 211 and 402 pa over the last decade, it is clear that the predominant presumption must be that mergers are allowed unless sufficient evidence suggests otherwise. USA statistics are not strictly comparable but they convey a similar message (FTC/DOJ, 2008). A lower reporting threshold (and possibly also greater propensity to merge) has meant between 1,014 and 2,201 mergers pa were reported to either the FTC or DOJ Antitrust Division 2002-07 of which only between 2.5% and 4.3% in each year were issued with a Second Request (equivalent to a Phase 2 investigation by the EC). An unknown proportion of notified mergers in the USA will have been modified to head off a Second Request. For the same years, the percentages referred to Phase 2 by the EC ranged between 3.3% and 4.0%.

[12] Nevertheless, the European Court of Justice recently argued that there was the same standard of proof required to approve as there is to prohibit a merger (see Bertelsmann AG and Sony Corporation of America v Independent Music Publishers and Labels Association (Impala), Case C-413/06 P, 10th July 2008). This is difficult to reconcile with any presumption, or any decision rule other than 51% of the evidence. However, the merging firms normally present evidence that competition will not be harmed and this could be seen as sufficient evidence of 'no harm' in the absence of basic evidence to the contrary being collected by the agency. This European Court of Justice decision was made public after our experiment had been completed.

[13] See e.g. Emson (2008).

[14] The equivalence of hypothesis testing with the presumption of 'innocence' is also proposed in Davis (1994).

The experiment was conducted between March and July 2008.[15] It was fully computerized and had three experimental treatments. Treatment A was a standard laboratory experiment with university students. Subjects were randomly seated in the laboratory. Computer terminals were partitioned to avoid communication by facial or verbal means. Subjects read the experimental instructions and answered a preliminary questionnaire, to check understanding of the instructions, before proceeding with the tasks. If they answered any questions incorrectly in the 'understanding check' questionnaire, they received feedback from the computer screen and could ask for additional help from the experimental supervisors. We had 57 students participating to Treatment A.

Treatment B was a purely online version of the experiment, run again with university students (different from those who participated to Treatment A). Subjects could log in the experiment remotely from their own workstations and do the experiment in their own time. They did the same understanding check questionnaire as in Treatment A and could email the experimental supervisors if anything was unclear (very few did). We had 153 students completing Treatment B online.

Treatment C was the most interesting treatment, as it was run online with competition policy practitioners who we had approached through the heads of their principal merger regulation competition agencies.[16] Practicioners who completed the experiment came from Austria, Canada, Denmark, the European Commission, France, Hungary, Japan, Netherlands, Norway, Spain and the UK (both the Competition Commission and the Office of Fair Trading). The experimental protocol was exactly the same as in Treatment B. We had 67 practitioners completing Treatment C online.

By comparing Treatment C with Treatment B, we can examine the effect of being a policy practitioner as opposed to a student in making merger decisions. If a difference is found, this could be due to a combination of agency culture, training, familiarity with the task, experience and self selection into the job. The comparison is also significant in verifying the usefulness of having professionals as opposed to the usual sample of students mostly found in

[15] The experimental instructions are provided in appendix B.

[16] Most competition policy agency heads expressed a concern that only a limited number of their staff should participate because of the opportunity cost of staff time. Most agency heads nominated a member of their staff to forward an e-mail approach from us so we could invite them to log on to a confidential secure website in order to participate in the experiment. We do not know how many staff were approached by each agency in this way and so we cannot calculate a meaningful response rate.

laboratory experiments (as discussed in the introduction). By comparing Treatment B with Treatment A, we can further identify whether moving from a controlled laboratory environment to a less controlled, but potentially less pressured, online environment makes any difference. There is evidence suggesting that there may be a difference (Shavit et al., 2001), and so this is a useful check.

3.2 The Decision Tasks

In all treatments, after subjects answered some demographic questions, subjects did 24 decision tasks in random order.[17] The general problem was set up for them in the instructions:

> "This is an experiment on decision making in relation to merger control by a competition authority. You are assumed to be working for an agency which is required to make a decision on the basis of a 'competition test' (discussed below). **A particular merger should be allowed unless it fails this competition test.** We use the terminology that a merger should not 'harm' competition; in practice, the formal law uses phrases like 'substantially lessens competition' or 'significantly impedes effective competition'. 'Harm' to competition should be interpreted as the same as one of these phrases."

In order to simplify the problem and maximize the interpretability of the results, evidence was presented as a number of 'signals' which might indicate that the merger would or would not harm competition. There was no separation of signals into the normal merger control issues of market definition, market share, cross-elasticities of demand, ease of entry, etc, though these were given as examples of the type of evidence contained in signals. Instead, subjects were told that signals may indicate either that the merger would be harmful or that it would not. They were also told that all signals were subject to error. More specifically, they were told that any one signal would give the incorrect answer 1/3 of the time and the correct answer 2/3 of the time.[18]

---

[17] By presenting the decision tasks in random order across subjects we were able to control for order effects and within experiment learning effects. Subjects did not receive any feedback on the 'correctness' of their choices after each round.

[18] If evidence was always relevant, true and incapable of misinterpretation, even one piece of evidence would be sufficient for an accurate decision. However, that is not the nature of most real evidence. For example, a high market share is not usually relevant to merger appraisal if the merger does not enhance that share in any one market; and if it does, that may still not indicate enhanced market power if the market definition is inappropriate; and, even if the merger substantially enhances market share in a relevant market, that may not matter if entry is easy or customers have buyer power. In general, therefore, any single piece of evidence is a noisy signal. The issue is how to decide on the basis of noisy and often conflicting evidence. This is what we model in the experiment (in a simplified way) with our noisy signals.

For each decision task, subjects were told the number of signals received on whether the merger would impede competition. They were then asked to report the minimum number of adverse signals, out of those received, needed for them to decide against the merger. Figure 1 contains example screens with the kind of decisions that subjects had to face.[19]

(Insert Figure 1 about here.)

There were 24 decision tasks provided by all combinations of 6 alternative standards of proof × 2 volumes of evidence × 2 decision types.[20] The standards of proof are discussed in the previous section. We now focus on the other dimensions.

Volumes of evidence. Each question had one of two volumes of evidence: 25 or 100 signals. Subjects were told that each signal was equivalent to data gathering or processing equivalent to one day's work by the competition agency. The aim was to provide a calibration that bore some similarity to the actual time taken to investigate a merger in each phase (bearing in mind some diminishing returns to further investigation).[21] Having two levels of volume of evidence also allows us to check whether subjects, or at least practitioners, understand the importance of volume in evaluating the evidence.

Decision types. Each question also required one of two levels of decision: to clear or refer the merger for further investigation; or to clear or prohibit the merger. No explicit costs were associated with errors in either type of decision, but practitioners in particular will have been very aware that the implied costs of erroneous referral versus prohibition could be very different for firms. More generally, it is possible that subjects may evaluate the moral cost of rejecting a merger outright as larger than simply referring it for possibly someone else to decide.[22]

Payments. Subjects were informed that the computer would take two 'rounds' out of 24. They were not told which rounds these would be. Subjects were told that, for each of these rounds, their rule would be compared with the available evidence where evidence relating to a merger would be randomly proposed by the computer. If, based on the given standard of

[19] The screens contain a pdf file link. This link appeared in Treatment B and C and, as explained in the instructions, it enabled subjects to download a pdf file with the instructions. In the lab version of the experiment of Treatment A, instructions were provided both online and in print.

[20] In experimental jargon, this corresponds to a 6 × 2 × 2 full factorial design, allowing us to separately identify the effect of each factor individually and in combination.

[21] For example, the European Commission has a statutory requirement to complete its phase 1 investigation in 25 working days and phase 2 allows an additional 90 days.

[22] For a moral cost decision making framework, see Levitt and List (2007).

proof and the other available information, their rule correctly decided in favour of or against the merger, the subject earned 1 point. If a merger which should be decided against was approved, or if a merger which should be approved was decided against, the subject earned 0 points. Subjects were not given any feedback on their decisions during the experiments; they only learned how many points they had earned at its end. This incentive compatible mechanism ensured that subjects would maximize their chances of earning points by making the choice which they genuinely felt was best.[23]

In Treatment A, each point earned was worth 5 pounds, and students also earned a participation fee of 3 pounds. However, apart from the logistics of paying a large number of subjects online, this kind of incentive structure would not have been motivating to (most) practitioners. Instead, points acted as lottery tickets to earn a large prize, in the form of a 250 pounds Amazon gift voucher.[24] To ensure comparability, in Treatment B we used the same lottery prize incentive system as in Treatment C.

4. Experimental Results

4.1 Paying Attention to the Task

Before presenting our main results, information can be gleaned on how generally thoughtful and attentive subjects were in making a decision by looking at the average response times taken by subjects to answer each question (see Figure 2).[25]

(Insert Figure 2 about here.)

Students took an average of 30 seconds and a median of 28 seconds to make a decision in the laboratory (treatment A) and an average of 25 seconds and a median of 17 seconds on

[23] The mechanism also required us (as any alternative incentive mechanism would also have) to specify a 'correct answer'. This does not matter from the perspective of our experimental data, since no feedback was provided during the experiment, but it was a choice to make nevertheless on defensible grounds. The four questions for which such an answer is in our view normatively least controversial are the 'balance of evidence' questions. We took the correct answer to be half the signals plus one: i.e., it was to state 13 negative signals about the merger when there were an overall 25 signals, and 51 negative signals when there were an overall 100 signals. For example, if the subject required 55 negative signals out of 100, she would be wrong if the computer generated 51, 52, 53 or 54 adverse signals. The computer selected two of these four 'balance of evidence' questions to determine points earned.

[24] Practitioners and Treatment B students earned up to 2 points as described above, and got an extra point as participation fee, and so they ended up with between 1 and 3 points. They were told that, on 1st July 2008, we would determine randomly four winners among all participants in treatments B and C. Each winner was awarded a gift voucher equivalent to 250 British pounds on www.amazon.co.uk or www.amazon.com or other Amazon.com, Inc. retail website. Participants who had scored 2 points had double the chance of winning relative to participants who had scored 1 point; participants who scored 3 points had three times the chance of winning relative to participants who had scored 1 point.

[25] Response times are measured by the software as the time taken to make a decision. As usual in experiments, response times were longest towards the beginning and decreased with time.

the web (treatment B); in neither case is there any particular pattern of average response rates across standards of proof, though the drop in decision time in moving to the less controlled web environment is significant (Mann Whitney $p < 0.001$). In moving from treatment B to treatment C, the average response time increased to 47 seconds, and the median response time increased to 29 seconds. Therefore, practitioners seemed to be at least as thoughtful using the online interface as students were in the more controlled laboratory environment,[26] and clearly spent more time than students doing the experiment online (Mann Whitney $p < 0.001$).

RESULT 1. *Practitioners appeared to pay at least as much as attention (if not more) to the decision tasks, when doing the experiment online, as students did when doing the experiment in the laboratory*.

Figure 2 also shows seemingly more time spent on average by practitioners to consider the ARCSE test (standard 4) and 'the evidence raises a concern' test (standard 6), but median response times are largely in line with those spent with others.[27] These two wordings are likely to have been the least familiar to practitioners (see Appendix A). That being said, within the class of evidence based standards, practitioners found it harder to process ARCSE (Wilcoxon $p = 0.05$) than the 'balance of evidence' standard; the comparison between 'evidence raises a concern' and 'balance of evidence' is less clear (Wilcoxon $p = 0.1$).

4.2 Measuring Experimental Responses

In turning to what decisions subjects made, we measure experimental responses in two ways. Our first measure is simply the fraction (*minimum number of adverse signals / number of signals received*), i.e. the proportion of adverse signals required before coming to an adverse decision: φ. While this standardizes for the total number of signals, it does not take explicit account of how subjects might adjust this proportion in the light of higher volumes of evidence reducing sample variation. Our second measure, α, takes explicit account of this and

[26] An obvious qualification is in order: in the online treatments, measurement of response time does not, as such, enable us to identify whether a subject was actually paying attention as opposed to, say, working on the experiment alongside working on another task. However, the across standards differences in response times by practitioners, which we consider shortly, cannot be explained by this alternative interpretation.

[27] Median response times for practitioners were 28, 26, 31, 30, 22 and 29 seconds respectively for standards 1, 2, 3, 4, 5 and 6 (as defined by Table 1). This reflects a few practitioners finding standards 4 and 6 particularly difficult.

frames the problem as one of classical hypothesis testing. Given the null hypothesis that the merger is not harmful to competition; we should still observe on average 1/3 of the signals being adverse, since 1/3 of times the signals will be wrong. The alternative hypothesis is that the merger is harmful: in this case, we would expect the proportion of adverse signals to be *sufficiently* greater than 1/3, where 'sufficiently' is determined by the test size α. The lower is α, the greater the proportion of adverse signals required. α is then the p-value (test size) of an alternative hypothesis set against the null hypothesis that the merger is not harmful (see Figure 3). Put another way, it is the probability of making a Type 1 error (i.e. referring or prohibiting a merger that is not anticompetitive). The calculation of α is set out in appendix C.[28]

(Insert Figure 3 about here.)

Since a larger sample size reduces the variance of the distribution of adverse signals, the same α would be associated with a lower φ. This is because more signals imply lower standard errors in the hypothesis test: if, to take an extreme example, we observe 40% of adverse signals with 5 observations, this is not significantly greater than 33.3% at a 0.05 significance level (i.e., with $\alpha = 0.05$), while it would be statistically significantly greater than 33.3% if there were 500 observations. A larger sample size implies a lower proportion φ of adverse signals is required to reject the null hypothesis, for any given α.

Given the above, if experimental responses fully take account of volumes of evidence (as if conducting a classical hypothesis test), the implied α would not change with the volume of evidence, but φ would. If instead subjects entirely neglect volumes of evidence in making their choices, then φ would not change with the volume of evidence, but α would; α would become lower with a larger volume of evidence. If, finally, subjects *partially* take account of volumes of evidence, then we would expect *both* a lower φ (to the extent that subjects do take volumes of evidence into account) and a lower α (to the extent that they do not) with a larger volume of evidence.

4.3 Analyzing Experimental Responses

[28] Rubinfeld (1985) provides a detailed discussion of appropriate significance levels for hypothesis testing in the quite different context of specific items of economic evidence.

We first provide some basic experimental results and then move on to employ regression analysis. Table 2 and Figure 4 provide information on the fraction φ of adverse signals required to decide against the merger, and on the corresponding implied average α values (i.e. assuming implicit statistical tests are conducted to determine φ).

(Insert Figure 4 and Table 2 about here.)

Professional and educational background. First, we compare students with practitioners. Average α and φ values are not statistically significantly different across treatments: practitioners are neither tougher, nor less tough, than students. With an average φ value at 0.55 and an average α value at just 0.22, there is consistent across-treatment evidence of belief conservatism regardless of professional background.

RESULT 2. *Both practitioners and students are belief conservative, i.e. they tend to retain the null hypothesis that the merger is harmless for competition*.

Equally, if we consider average α and φ values among practitioners with 2 or more years of work experience, it is not statistically different from that among inexperienced practitioners, although inexperienced practitioners appeared a little more cautious on average in point estimates (average $\varphi = 0.579$ and $\alpha = 0.187$ among inexperienced, and $\varphi = 0.557$ and $\alpha = 0.202$ among experienced practitioners). Professional training as either a lawyer or economist does not seem to affect how tough practitioner subjects are (see Table 3).[29]

(Insert Table 3 about here.)

RESULT 3. *Training and experience have no significant effect on average φ and α across practitioners*.

Age may be a better proxy for wider experience than stated antitrust work experience.[30] Age is also a factor that can be analyzed in the regression analyzed for both students and practitioners, though with students it tends to have low variance and so is less interesting. Following Garside et al (2009), we hypothesize that older subjects are tougher. There is no support for this hypothesis in Treatment A, but it receives some limited support in Treatment

[29] Studying economics or law also did not seem to make a difference in the student samples.
[30] We also checked gender effects, and found none.

B (in relation to φ, Spearman ρ = - 0.144, $p = 0.038$; in relation to α, Spearman ρ = 0.128, $p = 0.057$) and with Treatment C practitioners (in relation to φ, Spearman ρ = - 0.167, $p = 0.089$; in relation to α, Spearman ρ = 0.151, $p = 0.112$). We shall review this effect in the regression analysis, and find it statistically significant with practitioners. We therefore postpone summarizing the age effect as a Result.

Type of decision. Table 2 shows that, in all treatments, when the decision is to prohibit as opposed to refer the merger, subjects are less tough. To put it differently, the prospect of prohibition makes people more conservative in their beliefs: average φ goes up by about 0.04 and average α goes down by about 0.04. This difference is statistically significant (at $p < 0.05$ or better) in Wilcoxon tests for all treatments and for both φ and α measures. Therefore, decisions appear sensitive to the likely costs of Type 1 error (i.e., cost of wrongly coming to an adverse decision), which is likely to be greater if the merger is prohibited outright than if it is referred for further investigation.

RESULT 4. *Practitioners and students alike are less tough if the decision is to prohibit the merger than if it is just to refer the merger for further investigation. The effect is quantitatively larger for practitioners (who may be more aware of the potential cost of a Type 1 error in the merger context).*

Standards of proof. For practitioners (Treatment C), Figure 4 and Table 2 show a clear pattern, in terms of φ values, broadly consistent with our characterization of high, medium and low standards of proof, both in relation to probabilistic and to evidence based standards. The implied differences in standards are typically not only statistically significant but also quantitatively large.[31] Among probabilistic standards, average φ drops from 0.714 for the hypothesized high burden of proof required by standard 1 ('beyond reasonable doubt') to 0.527 for the hypothesized medium burden of proof required by standard 2 ('it is likely';

[31] That being said, the figure does not show much of an average difference in φ and α between the set of probabilistic standards and that of evidence based standards (φ = 0.564 and 0.535, α = 0.208 and 0.232 for probabilistic and evidence based standards respectively). While these differences are statistically significant ($P < 0.001$), they are quantitatively small and in different directions (in terms of implied degree of conservatism) depending on whether we use φ and α; furthermore, they do not hide any treatment specific large difference driving the results.

Wilcoxon $p < 0.001$) and 0.466 for the hypothesized low burden of proof required by standard 3 ('there is a realistic prospect'; Wilcoxon $p < 0.001$). Among evidence based standards, the ARCSE standard (standard 4) is almost as high in terms of burden proof ($\varphi = 0.663$) as 'beyond reasonable doubt' is among probabilistic ones. This is despite the fact that ARCSE was suggested in the context of civil merger control and not in the criminal context. ARCSE is higher than the hypothesized medium burden of proof 'balance of evidence' (standard 5, $\varphi = 0.532$, Wilcoxon $p < 0.001$), which in turn is higher than the hypothesized low burden of proof 'raises a concern' (standard 6, $\varphi = 0.39$, Wilcoxon $p < 0.001$).

Confirming this picture, average α values for practitioners for 'beyond reasonable doubt' and ARCSE are also very low, hovering around 0.1. This can be interpreted as needing to be 90% certain that they will not incur a Type 1 error (i.e. wrongly find against the merger). Interestingly, we find the estimated α value for 'balance of evidence' also low (0.128), confirming a significant degree of belief conservatism. The probabilistic standard 'it is likely' is confirmed, comparatively speaking, as a medium burden of proof. It has higher α than the above three higher standards (0.222; Wilcoxon $p < 0.001$), and lower α than the hypothesized low burden of proof 'there is a realistic prospect' (0.288; Wilcoxon $p = 0.004$) and particularly 'raises a concern' (0.443; Wilcoxon $p < 0.001$). The latter is unequivocally the lowest standard of proof.

RESULT 5. *ARCSE stands out as placing almost as high a standard of proof as the 'beyond reasonable doubt'.*

RESULT 6. *'It is likely' places a medium standard of proof, though possibly one assessed higher, according to α measures, than the 'balance of evidence' standard.*

RESULT 7. *'The evidence raises a concern' is the lowest standard of proof, followed by 'there is a realistic prospect'.*

In moving from practitioners to the students of Treatments A and B, Figure 4 and Table 2 illustrate that, broadly speaking, the same pattern occurs, but, equally, that there is

considerably less variability. While practitioners are no less or no more belief conservative than students on average, practitioners do appear more discriminating in their choices. For example, while average φ values span 0.39 and 0.714 for practitioners, they span only half the range for both Treatment A (between 0.454 and 0.613) and Treatment B (between 0.443 and 0.630). The same is true for ranges of α values.

RESULT 8. *Practitioners were more discriminating in their responses than students, i.e. they were better able to make their response a function of each standard of proof.*

As an aside on experimental methodology, note that Treatments A and B follow each other closely, showing that having an online experiment, rather than a more controlled laboratory one, does not seem to have had a material effect on subjects' decisions.

RESULT 9. *An online environment did not change how subjects comparatively treated different standards of proof.*

Volume of evidence. As discussed in section 4.2, we consider both φ and α measures in our analysis to allow people to take into account the volume of available evidence. Table 2 shows that φ went down slightly as a result of having 100 rather than 25 signals. This is statistically significant in all treatments (Wilcoxon $p = 0.02$, 0.03 and < 0.001 for Treatments A, B and C respectively). Estimated α *also* goes down with students, though the difference is statistically significant only with Treatment B ($p = 0.002$). Practitioners instead appear to take the volume of evidence fully into consideration and their α values are virtually identical with 25 and 100 signals (Wilcoxon *n.s.*). Since the regression analysis, which we consider next, points to a slightly different picture, we defer condensing these findings into a Result until then. That being said, the comparative stability of α values, especially and crucially with practitioners, points to the usefulness of using α values at least as a complementary measure of the toughness of subjects in appraising evidence.

Regression analysis. We now employ tobit random effects regressions to analyze the data in more depth while controlling for covariates. Tobit regressions are used because of the censoring of the data between 0 and 1, and the random effects control for the non-independence of choices made by the same subject. Panel (a) of Table 4 displays the regressions on φ, and panel (b) those on α.[32]

(Insert Table 4 about here.)

The independent variables are: *Round*, which is the round number from 1 to 24,[33] and its squared value *Round Squared* to enable us to check whether initial learning adjustment slows down as the experiment progresses; *Decision Type*, equal to 1 if the decision is whether to prohibit and to 0 otherwise; *Nsignals*, equal to 1 (0) if there are 100 (25) signals; standard of proof dummy variables (*Standard1*, equal to 1 if the standard of proof is 1 as defined by Table 1, and equal to 0 otherwise; *Standard2*, *Standard3*, *Standard4* and *Standard5*, similarly defined) which operate against the baseline of the standard that requires the lowest burden of proof, 'the evidence raises a concern' (see Figure 4); *Age*, equal to the age of the subject; *Gender*, equal to 1 (0) if the subject is a man (woman); and, for Treatment C, *ECN Authority*, equal to 1 if the practitioners works in a European (European Competition Network) competition authority; and to 0 otherwise; *CmL Authority* equal to 1 if the respondent's agency is from a country with a predominantly common law tradition (UK and Canada) as distinct from civil law tradition equal to 0.[34] Non-national individuals sometimes work in national competition agencies, so we also test for the effect original culture may have, with *ECN National* and *CmL National* mirroring the agency dummies but defined by individual.[35]

The *Round* and *RoundSquared* variables give qualitatively the same dynamic picture across all regressions, and are statistically significant in all regressions except the Treatment A regression on α. They show that both students and practitioners initially became more

[32] Alternative specifications were tested and do not change the key results.
[33] The order of questions was randomized.
[34] It is not always straightforward to classify legal systems as one or the other. For example, the European Court is founded in a civil system but has been adopting elements of common law such as precedent. Scotland in the UK and Quebec in Canada have separate legal traditions with a strong element of civil law. Tetley (2000) provides a discussion focusing on mixed systems.
[35] In other regressions, we included the competition authorities dummies but excluded the nationality dummies; or included the nationality dummies but excluded the competition authorities dummies. We found that this makes no difference to the results: the coefficients on the included legal system dummy variables remain statistically insignificant, as in the regressions in Table 4.

belief conservative (i.e. reluctant to find against) with experience of the experimental task but the effect then tends to flatten off.

The coefficient on *Decision Type* is statistically significant, and supports Result 4 above: subjects were tougher if the decision is to refer the merger than they were if it is to prohibit the merger.

The coefficient on *Nsignals* indicates the extent to which subjects take into account the effect that the volume of evidence has on the sample variance. Except for the Treatment A coefficient in the regression φ, all other coefficients on *Nsignals* are statistically significant. In Treatment C, we find practitioners make a partial adjustment (in contrast with the full adjustment found in the bivariate analysis above). They lower the required proportion of adverse signals to find against the merger, but not by enough to maintain the same probability of Type 1 error. The Treatment B students decided similarly, but with a slightly lesser adjustment for volume of evidence and correspondingly greater change in the probability of Type 1 error. Treatment A students made no such adjustment. While the details differ from the earlier bivariate analysis, when taken together with those results, the general message is one where subjects – especially practitioners – take the volume of evidence into account, but not fully.

RESULT 10. *Practitioners (and to a lesser extent students) take the volume of evidence at least partially into account when making decisions*.

The coefficients on the standards of proof are generally significant and confirm Results 5 through 9 above. The coefficients with Treatment C practitioners tend to be at least as large or larger, and tend to have a larger range than those with Treatments A and B students. The largest coefficients are on 'beyond reasonable doubt' (Standard1) and ARCSE (Standard4), as hypothesized, and the coefficients on the two standards are not too dissimilar. The other patterns from the bivariate analysis are also broadly reproduced. The quantitative impact of moving between high, medium and low standards is at least as great as for the impact of decision type. If we focus on practitioners, we find that, relative to the low burden of proof baseline of 'the evidence raises a concern', the fraction of required adverse signals to decide against a merger goes up by around 0.3 and the implied α goes down by around 0.6 with

'beyond reasonable doubt' and ARCSE; φ increases by 0.15 and α decreases by 0.3-0.4 with the hypothesized medium burden of proof standards ('it is likely' and 'balance of evidence'); and φ goes up by just 0.08 and α goes down by just 0.2 with the other hypothesized low burden of proof standard ('realistic prospect').

Since the student age variance is small, it is not especially interesting that Age is not statistically significant in Treatments A and B. With practitioners, however, we find that φ becomes lower, and α becomes larger, with older subjects ($p < 0.05$, two tailed, in both cases).[36]

RESULT 11. *Practitioners become tougher with age*.

The remaining variables reveal no significant patterns. With the exception of the Treatment B φ regression, there is no evidence of gender effect. There is also no evidence of an effect of common versus civil law traditions or of European versus other competition agencies.[37] Result 3 on the insignificant effect of professional training is also confirmed.

5. Discussion

This paper provides confirmation of an important tenet of law and economics – namely, that standards of proof are an important influence on the way evidence is used to determine judgment. We find first that both lay students and specialist practitioners can be influenced by the required standard; and second, that relative error costs are also influential for interpreting a given standard. Although there is some previous experimental evidence on this in relation to lay subjects (see below), we provide empirical support for the general approach to optimal legal standards (e.g. Posner, 1973; Rubinfeld and Sappington, 1987) as applied to professional decision makers. Furthermore, our approach enables us better to understand how standards are actually interpreted: a striking result was that the ARCSE criterion (accurate, reliable, consistent and sufficient evidence) proposed by the European Court of Justice stands

[36] Also, and as in the bivariate analysis, in alternative regression specifications where we used years of work experience in alternative to or in combination with age, years of work experience was not found to be related to either φ or α.

[37] This result is robust to different regression specifications, for example if one excludes either nationality or authority variables, as noted in an earlier footnote; or if one excludes Norway as not belonging to the European Union; or if one has dummies for the UK instead of common law dummies for UK and Canada together.

out as placing almost as high a standard of proof as the 'beyond reasonable doubt' criterion from criminal law. It does not provide a 'middle level' standard of proof.

We had expected our approach to find differences between practitioners in different legal systems. At a broad level, there is a recent legal literature that has examined the apparent differences between civil law and common law jurisdictions (e.g. Posner,1999; Clermont and Sherwin, 2002).[38] However, we find no evidence for such effects amongst our practitioner treatment. This may be because the key early stages of merger appraisal are essentially inquisitorial in most countries, and cases rarely go to court. The literature on international differences in merger control decisions also focuses on differences in economic theories more than legal traditions (e.g. Fox, 2002). We have found that law and training do matter (compare our practitioner and student treatments). However, we were unable to distinguish identifiable differences between individual jurisdictions, possibly because our sample size was too small.

Garside et al. (2008) find that experience of UK Competition Commission chairmen in 'abuse of market power' inquiries is positively related to the probability of finding against the firm under investigation. They interpret this in relation to experimental evidence on bias and overconfidence; for example, police officers are more likely to claim a witness is deceitful as their experience increases.[39] Our methodology abstracts from that particular source of bias as we fix the probability that any specific piece of evidence is misleading. This suggests a different mechanism working through the individual's willingness to reject the hypothesis of 'no harm'. Although we do not find that specific work experience has a significant effect, we do find that the age of practitioners matters. This is consistent with Garside et al. (2008) in that older practitioners tended to be tougher.

This is not the first experimental paper on standards of proof (e.g. Kagehiro and Stanton, 1985; Dhami, 2008; Glöckner and Engel, 2008). However, the previous literature differs from our work in at least three important ways. First, in our experiment subjects' decisions were incentivized financially. Second, the existing literature makes a virtue out of using student subjects by investigating issues in jury trials, whereas we consider decision making by

[38] For example: the commercial standard of proof is expected to be higher in civil law countries (Clermont and Sherwin, 2002); and there are also different balances between 'inquisitorial' and 'adversarial' approaches (Posner, 1999).

[39] The evidence appears not to control separately for age.

a specialist agency where training, job selection and experience provide an important new dimension. Our evidence supports the view that professional decision making is different. This is further discussed below. Third, the earlier experimental literature presents subjects with a story on which to pass judgment, typically a synopsis of a real case. This inevitably introduces a strong social dimension into what types of evidence are more important, and the story-based experimental design is unable to control for this.[40] We have been able to vary one controlled element of the appraisal problem at a time in order to understand precisely how each element affects the decision.

From a behavioral economics viewpoint, we found general evidence of belief conservatism, with both students and practitioners. For example, subjects required more than 50% on the 'balance of evidence' questions and correspondingly tended to have low α values. This suggests evidence for a general psychological mechanism at work in the direction of belief conservatism, albeit one that can be sensitive not only to the standard of proof but also to the cost of Type 1 error (i.e. incorrectly finding against the proposed merger). The latter was true also for students, and possibly reflected a perceived lower moral cost of making the wrong decision.[41] That subjects – especially practitioners – took the volume of evidence into account suggests the usefulness of using both our φ measure and (in a complementary way) our α measure. This is connected to the idea that subjects may form expectations as if they are undertaking a form of statistical inference.[42] Nevertheless, most of our qualitative results are robust to whether α, or the simple fraction φ of adverse signals, is employed.

The use of students as a subject pool is one of the classical criticisms of standard economic experiments. We found that professionalization (e.g., general training or selection into the role of competition policy decision makers) made subjects more discriminating in dealing with standards of proof, thus validating the usefulness of using practitioners for the

[40] This problem may be more severe for mergers than for a criminal trial because of the possible differences in economic theories to be applied in the former; e.g. the weight given to market share evidence compared with demand elasticity evidence.

[41] See Levitt and List (2007) for an example of moral cost framework.

[42] The idea that expectations are formed inferentially traces its roots back at least to Rappaport (1985), and is developed in Goldberg and Frydman (1996) and Frydman and Goldberg (2003), who allow agents to conduct hypothesis tests over models; in Foster and Peyton Young (2003), who consider hypothesis testing by agents on their opponents' repeated games strategies; and in Menzies and Zizzo (2007, 2009). Scott and Nowak (2005) describe certain environments where hypothesis testing is an optimal way to learn, and Bacchetta and van Wicoop (forthcoming) show that belief conservatism can be formally related to the size of the adjustment costs. For some experimental evidence on belief conservatism, see Edwards (1968) and Menzies and Zizzo (2007, 2008).

external validity of our experiment. Equally interesting, however, is the fact though that a number of broad qualitative findings - in relation to, say, standards of proof and average degree of belief conservatism or toughness - applied to students as well as practitioners, showing that having just a student sample would have been surprisingly (if only partially) useful even in the highly technical context of our experiment. This nuanced conclusion is consistent with previous experimental work that has compared students with other types of professional (Friedman and Sunder, 1994). Reassuringly, using an online setting did not appear to produce a loss of experimental control for any of our key findings, with most results in Treatment A mirroring those of Treatment B except where effects appear small (such as the potential effect of age on toughness in making decisions).[43]

6. Conclusions

The aim of this paper was to understand how legal decision making is influenced by alternative standards of proof in the face of limited and conflicting evidence. The context of merger regulation provides a flexible framework for investigating this and a number of related issues. Our central finding is that standards count. Understanding how they count then becomes an important question: for example, we found that the ARCSE criterion (accurate, reliable, consistent and sufficient evidence) proposed by the European Court of Justice (ECJ, 2005) was interpreted as placing almost as high a standard of proof as the 'beyond reasonable doubt' standard from criminal law.

Both lay and professional samples adapt their decisions in the light of the stated standard. Both also adopt a conservative reluctance to find harm in a proposed merger and they weigh the costs of Type 1 errors. We further find that professionalization does matter importantly in at least three ways. First, it sharpens distinctions between standards. Second, it is associated with a more sophisticated weighing of different volumes of evidence. Third, it is associated with a greater adjustment for the potential cost of errors.

We are unable to distinguish precisely how this professionalization effect works, but we do find that it is not particularly associated with years of work experience, gender, legal versus economics training, nationality or civil versus common law background. We also find

[43] Two examples of experiments which have compared an online environment with a laboratory environment are Charness et al. (2007) and Srzek and Baron (2007).

evidence that older practitioners tend to be tougher. These results still leave open explanations based on agency-provided training, competition agency culture and self-selection of individuals into careers in competition agencies.

Finally, the fact that standards matter to professionals means that more attention should be given to identifying appropriate standards of proof, as well as clarifying economic theories of harm. This is important both in establishing agency norms and when comparing apparent differences in decisions across jurisdictions.

References

Bacchetta, Philippe, and Eric van Wincoop. Forthcoming. "Infrequent Portfolio Decisions: A Solution to the Forward Discount Puzzle." *American Economic Review*.

Bailey, David. 2003. "Standard of Proof in EC Merger Proceedings: A Common Law Perspective." *Common Market Law Review* 40:845-888.

Baron, Jonathan, and Helena Srzek,. (2007) "The Value of Choice in Insurance Purchasing." *Journal of Economic Psychology* 28:529-544.

Charness, Gary, Ernan Haruvy, and Doron Sonsino. 2007. "Social Distance and Reciprocity: An Internet Experiment." *Journal of Economic Behavior and Organization* 63:88-103.

Clermont, Kevin M., and Emily Sherwin, 2002. "A Comparative View of Standards of Proof." *The American Journal of Comparative Law* 50:243-275.

Comanor, William S., and Lawrence J. White. "Market Power or Efficiency: A Review of Antitrust Standards" *Review of Industrial Organization* 7:105-116.

Competition Commission. 2003. "Merger references: Competition Commission Guidelines." http://www.competition-commission.org.uk/rep_pub/rules_and_guide/pdf/ CC2.pdf .

Dawson, Beth, and Robert G. Trapp. 2004. *Basic and Clinical Biostatistics*. 4th ed. New York: Lange Medical Books McGraw-Hill.

Dhami, Mandeep K., 2008, 'On Measuring Quantitative Interpretations of Reasonable Doubt' *Journal of Experimental Psychology: Applied* 414:353-63.

Department of Justice. Federal Trade Commission. 2006. "Commentary on the Horizontal Merger Guidelines" http://www.justice.gov/atr/public/guidelines/215247.pdf.

Edwards, Wayne. 1968. "Conservatism, in Human Information Processing." 17-52 in *Formal Representation of Human Judgment*, edited by Benjamin Kleinmuntz. New York: Wiley.

Emson, Raymond. 2008. *Evidence*, 4th ed. Basingstocke: Palgrave-Macmillan.

Foster, Dean P., and Young H. Peyton. "Learning, Hypothesis Testing, and Nash Equilibrium." *Games and Economic Behavior* 45:73-96.

Fox, Eleanor (2002), "United States and European Merger Policy: Fault Lines and Bridges for Mergers that Create Incentives for Exclusionary Practices." *10 George Mason Law Review* 471:474-475.

Friedman, Daniel, and Shyam Sunder. 1994. *Experimental Methods: A Primer for Economists*. Cambridge: Cambridge University Press.

Frydman, Roman, and Michael D. Goldberg. 2003. "Imperfect Knowledge Expectations, Uncertainty-Adjusted Uncovered Interest Rate Parity, and Exchange Rate Dynamics." 145-182 in *Knowledge, Information and Expectations in Modern Macroeconomics* edited by Philippe Aghion, Roman Frydman, Joseph Stiglitz and Michael Woodford. Princeton: Princeton University Press.

Department of Justice. Federal Trade Commission. 2008. *Hart-Scott-Rodino Annual Report Fiscal Year 2007*. http://www.ftc.gov/os/2008/11/hsrreportfy2007.pdf .

Garside, Ludivine, Paul Grout, and Anna Zalewska. 2008. "Does within-tenure experience make you tougher? Evidence from competition law." Unpublished manuscript. University of Bath, School of Management.

Glöckner, Andreas, and Christoph Engel. 2008. "Can we trust intuitive jurors? An experimental analysis" Working Paper 2008/36, Max Planck Institute for Research on Collective Goods. Bonn.

Goldberg, Michael D., and Roman Frydman. 1996. "Imperfect Knowledge and Behavior in the Foreign Exchange Market." *Economic Journal* 106: 869-893.

Goodie, Adam S., and Edmund Fantino. 1996. "Learning to Commit or Avoid the Base-Rate Error." *Nature* 380: 247-249.

Hay, George A. and Gregory J. Werden. 1993. "Horizontal Mergers: Law, Policy, and Economics." *American Economic Review: Papers and Proceedings* 83: 173-177.

Kagehiro, Dorothy K, and W. Clark Stanton. 1985. "Legal vs. Quantified Definitions of Standards of Proof" *Law and Human Behavior* 9:159-178.

Keane, Adrian. 2008. *The Modern Law of Evidence*, 7th ed. Oxford: Oxford University Press.

Levitt, Steven D., and John A List. 2007. "What Do Laboratory Experiments Measuring Social Preferences Reveal about the Real World?" *Journal of Economic Perspectives* 21: 153-174.

Louveaux, Bernard and Paul Gilbert. 2005. 'The Standard of Proof under the Competition Act." *European Competition Law Review* 26:173-177.

Menzies, Gordon D., and Daniel J. Zizzo 2004. "Inferential Expectations" Discussion Paper no. 187 Oxford University, Oxford.

Menzies, Gordon D., and Daniel J. Zizzo. 2007. "Exchange Rate Markets and Conservative Inferential Expectations." Discussion Paper no. 2. Australian National University Centre for Applied Macroeconomic Analysis, Canberra.

Menzies, Gordon D., and Daniel J. Zizzo. 2008, "Do Only Economists Rely on Statistical Significance?" Social Science Research Network Discussion Paper. Norwich and Sydney.

Menzies, Gordon D., and Zizzo, Daniel J. 2009. "Inferential Expectations." *B.E. Journal of Macroeconomics* [*Advances*]. 9: Article 42.

Office of Fair Trading. 2003. *Mergers: Substantive assessment guidance*. London. Available at http://www.oft.gov.uk/shared_oft/business_leaflets/enterprise_act/oft516.pdf

Ordover, Janusz A., and Robert J. Reynolds. 2002. "Archimedean leveraging and the GE-Honeywell transaction" *Antitrust Law Journal*, 70:171-198.

Patterson, Donna E. and Carl Shapiro. 2001. "Trans-Atlantic Divergence in GE/Honeywell: Causes and Lessons" *Antitrust Magazine*, Fall.

Posner, A. Richard. 1973. "An Economic Approach to Legal Procedure and Judicial Administration" *The Journal of Legal Studies*, 2:399-458.

Posner, A. Richard. 1999. "An Economic Approach to the Law of Evidence." *Stanford Law Review* 51:1477-1546.

Rappoport, Peter. 1985. "Unfalsified Expectations: An Alternative Perspective on Modelling Expectations in Macroeconomics." research report 85-16. New York University, C.V. Starr Center for Applied Economics, New York.

Rubinfeld, L. Daniel. 1985. "Econometrics in the Courtroom" *Columbia Law Review*, 85:1048-1097.

Rubinfeld, L. Daniel. and David E. M. Sappington. 1987. "Efficient Awards and Standards of Proof in Judicial Proceedings" *The RAND Journal of Economics* 18: 308-315.

Schlossberg, S. Robert. 2008. *Mergers and Acquisitions: understanding the antitrust issues.* 2nd ed. Chicago: ABA Section of Antitrust Law.

Scott, Andrew (2006) 'Tweedledum and tweedledee?: regime dynamics in US and EC merger control.' 72-108 in *Handbook of research in trans-Atlantic antitrust,* edited by Philip Marsden. Cheltenham: Edward Elgar Publishing.

Scott, Clayton D., and Robert D. Nowak. 2005. "A Neyman-Pearson Approach to Statistical Learning." *IEEE Transactions on Information Theory* 51: 3806-3819.

Shavit, Tal, Doron Sonsino, and Uri Benzion. 2001. "A Comparative Study of Lotteries-Evaluation in Class and on the Web." *Journal of Economic Psychology* 22: 483-491.

Tetley, William 2000. "Mixed Jurisdictions: Common Law vs. Civil Law (Codified and Uncodified)." *Louisiana Law Review* 60:677-738.

Vesterdorf, Bo. 2005. "Standard of proof in merger cases: reflections in the light of recent case law of the Community Courts." *European Competition Journal* 1:3-33.

Wallsten, Thomas S., David V. Budescu, and Rami Zwick. 1993. "Comparing the Calibration and Coherence of Numerical and Verbal Probability Judgments." *Management Science* 39: 176-190.

Appendix A: Legal Analysis of Standards of Proof in the EU, USA and UK

Vesterdorf (2005) defines the standard of proof as 'the threshold that must be met before an adjudicator decides that a point is proven in law' (p.6). US law requires a 'preponderance of the evidence' as the civil standard of proof. There exists a potential higher standard of 'clear and convincing' evidence which is applicable in some cases where there is a very high cost of an adverse finding.[44] This is higher than the basic civil standard but not as high as the criminal standard of 'beyond reasonable doubt'. It does not appear to be used for merger cases. The civil standard in English law is similar, but phrased as 'balance of probabilities'. There is no intermediate standard, but the single civil standard is interpreted flexibly according to the circumstances of the specific case: 'the more serious the allegation or the more serious the consequences if the allegation is proved, the stronger must be the evidence before a court will find the allegation proved on the balance of probabilities' (Keane, 2008, p.108). Seriousness is associated with ex ante likelihood as much as potential harm.[45]

In the context of European law, Vesterdorf (2005) argues that 'the civil standard of balance of probabilities [also known as 'preponderance of the evidence'] is sufficiently flexible so that its intensity can vary depending on the interests at stake'. After arguing that 'the case law of the Community courts does not reveal, with any degree of clarity, whether the standard of proof incumbent on the Commission is one of the balance of probabilities, beyond reasonable doubt or any other intermediate standard' (p.25), the then President of the CFI concluded that for merger control 'something more than a pure balance of probabilities standard but most certainly something less than a criminal standard ought to apply' (p.31).[46]

An important test case at the highest European Court (ECJ, 2005) formed the following judgment in relation to evidential requirements, albeit with specific reference to a complex theory of harm: 'Not only must the Community Courts, inter alia, establish whether the

[44] The classic case determining this was Addington v State of Texas, 1979, 441 US 418 (USSC) – see Emson (2008) p.406.

[45] In *Re H*, Lord Nicholls concludes: 'When assessing the probabilities the court will have in mind as a factor, to whatever extent is appropriate in the particular case, that the more serious the allegation the less likely is that the event occurred and, hence, the stronger should be the evidence before the court concludes that the allegation is established on the balance of probability'. [1996] AC 563, at p. 586. Emson (2008) elaborates on the legal relationship between 'probability' and 'evidence': 'as serious impropriety is less likely to have occurred than something relatively trivial, a serious allegation requires highly cogent evidence to overcome that inherent unlikelihood and bring the probability of its occurrence over the 50% threshold' (p.404).

[46] Note that the civil law systems of continental Europe (e.g. France, Germany) do not usually have this nuanced attitude to standards of proof, and it is often argued that they have no difference between the criminal and civil standards, with both being at the high level of reasonable doubt (see Clermont and Sherwin, 2002). Although the European Court of Justice was founded in the civil system, it has evolved common law characteristics such as extensive use of precedent in competition cases. See also Bailey (2003).

evidence relied on is factually accurate, reliable and consistent but also whether that evidence contains all the information which must be taken into account in order to assess a complex situation and whether it is capable of substantiating the conclusions drawn from it. Such a review is all the more necessary in the case of a prospective analysis required when examining a planned merger with conglomerate effect' (#39).[47] This was widely considered as raising the standard rather than providing a 'middle level' standard of proof such as balance of probabilities. Scott (2006) argues that it takes the standard closer to that used in the USA where the insistence on proof of anti-competitive effects has been a feature of court decisions since the 1980s.[48]

A complementary take on the standard of proof can sometimes be found in either primary legislation ('hard law') or in agency guidelines ('soft law'). The DoJ/FTC (2006, p.1) horizontal merger guidelines set out that the US agencies will find against mergers which are 'likely to' create or enhance market power.[49] The EC guidelines also explain that the merger will be prohibited if it 'is likely to' impede effective competition (EC Non-Horizontal Guidelines, 2004) [50] The UK Competition Commission (2003, #1.19) guidelines use 'expectation' instead of 'likelihood' then draw these wordings together: "for the Commission to reach an adverse decision either the merger must have resulted in an SLC or the Commission must expect such a result. The Commission will usually have such an expectation if it considers that it is more likely than not that the SLC will result".

A distinctive feature of merger regulation in that it typically involves two stages (or phases). In the first stage, there is a relatively short review of the immediately available evidence and an explicit or implicit standard of proof against which to judge whether this evidence is sufficient either to allow the merger unchallenged or to refer it for further evidence gathering. For referred mergers, there is more evidence gathering and processing. A different standard of proof is applied at this later stage to decide whether to allow the merger or prohibit it.[51]

---

[47] The CFI later applied the same standard in the *Sony BMG* appeal.
[48] He cites *Continental T.V., Inc*. v. *GTE Sylvania Inc*., 433 U.S. 36 (1977) as being a turning point in this regard.
[49] In relation to the efficiency defence, the Guidelines reverse the presumption in favour of the merger and require 'clear and convincing evidence' to be provided by the merging parties.
[50] The ECMR Art.2.3 (also Art.8.3) prohibits a merger 'which would significantly impede effective competition', but it is a legal convention to phrase legislation as if the proof is certain once the requisite standard has been met.
[51] In mergers involving multiple markets, a challenge may take place against only part of the merger and the referral for more evidence may be selective with possible prohibition only in relation to certain markets. Such decisions are often called merger agreements subject to remedies (or undertakings). Single market mergers can

Apart from having different amounts of evidence at each stage, there are differences in error costs. While both stages might incur the same Type 2 decision error of allowing an anticompetitive merger, the Type 1 errors are quite different. In the first stage of merger control, the immediate cost of Type 1 error is the delay in completing an efficient merger plus the administrative costs of investigation. In the second stage, the cost of Type 1 error is that the potential efficiencies will never be achieved. A different wording for the standard of proof is generally applied in the first stage. This may reflect a judgment about the appropriate balance of error costs, or an adjustment to the evidentiary threshold to reflect the 'double chance' of the merger being allowed.[52]

There is little formal guidance on the proof required for a Second Request in the USA. The HSR Act[53] says that a Second Request should not be 'unduly burdensome'. The ABA book edited by Schlossberg (2008) says that there should be a Second Request '[i]f substantive antitrust concerns remain at the end of the initial waiting period, or if there has been insufficient time to conclude that the transaction does not raise anticompetitive concerns'. The EC is more explicit. The basic legislation (i.e. ECMR) states in Art.6.1(c) that the merger must be referred to Phase II if it 'raises serious doubts' that the merger will impede effective competition. The UK merger legislation was reformed in the Enterprise Act (2002) and there was an early test case at the new Competition Appeal Tribunal in which the standard required of the OFT to refer a merger was tested. As a result, the standard became that a merger must be referred if there is a 'reasonable belief that there is a realistic prospect' that the merger would significantly lessen competition.[54] In order to simplify wording of Phase I standards for the experiment, we truncated the UK wording to 'realistic prospect' for our probabilistic standard, and used 'evidence raises a concern' for the evidentiary standard.

---

also be remedied short of prohibition (e.g. partial divestiture of capacity; or behavioural remedies). For our purposes, it is sufficient to concentrate on whether the evidence should be interpreted as justifying any intervention in the merger and to consider the single market case as convenient shorthand.

[52] In Europe, merger prohibitions can be appealed to the Court of First Instance but the delay incurred during appeal means that the merger can rarely be resurrected even if successful. In the USA, the court is more immediately available if the firms fail to reach agreement with the agency, and this may restrict the discretion of the agency in interpreting the standard of proof.

[53] US Code Collection Title 15 Ch.1, sect.18a (which provides latest version of HSR Act) at http://www.ftc.gov/bc/docs/statute.pdf

[54] The subsequent OFT guidance (2004, #3.2) explain how it would interpret this: 'By the term 'realistic prospect', the OFT means not only a prospect that has more than a 50 per cent chance of occurring, but also a prospect that is not fanciful but has less than a 50 per cent chance of occurring…. where the information available to the OFT is full and extensive, the degree of likelihood that the OFT must require to believe that it may be the case that a merger may be expected to result in a substantial lessening of competition may be higher up the scale of probability (albeit less than 50 per cent) than compared to when there is less information available, particularly as regards central points in the analysis.'

Appendix B: Experimental Instructions

Introduction

This is an experiment on decision making in relation to merger control by a competition authority. You are assumed to be working for an agency which is required to make a decision on the basis of a 'competition test' (discussed below). **A particular merger should be *allowed unless* it fails this competition test.** We use the terminology that a merger should not '*harm*' competition; in practice, the formal law often expands this wording into phrases like 'significantly impedes effective competition' or 'substantially lessens competition'. 'Harm' to competition should be interpreted as the same as one of these phrases.

[*Treatment A only*: Please raise your hand if you have any questions of clarification at any point in the experiment.] [*Treatments B and C only*: You can view the experimental instructions online but you may also find it handy to download them as an Adobe Acrobat pdf file by clicking the link on the computer screen. To help preserve the scientific value of the experiment, we ask you to complete it in a single session if at all possible (it should last up to around 40 minutes), although you can log off and log back in if you have the need to. We also ask you to do the experiment on your own, without discussing it with your colleagues either during the experiment or afterwards (before the beginning of August 2008). Please email ccp.experiments@uea.ac.uk if you have any questions of clarification at any point in the experiment or if you have any login problems. Also email us if you want to know about the experiment results: we shall be happy to send (when ready) an electronic copy of the academic paper or report resulting from this research to all participants who have successfully completed the experiment and who email us a request to be debriefed.]

The experiment

A competition agency collects evidence relating to the competitive effects of each merger. We refer to this evidence as 'signals'. These should be taken to mean pieces of evidence such that each piece of evidence is of equal weight. For example, you can think of signals as being additional information about market definition, or about barriers to entry, or about how products can be substituted with those of other firms, etc. These signals are typically imprecise and do not always indicate the right answer. In order to provide some context, you may think of a signal as evidence that takes around one day for the investigating team to collect, decipher or evaluate. Each signal is an independently determined piece of evidence.

Each signal will suggest either that the merger harms competition or that it does not. We call a signal suggesting that the merger harms competition an '*adverse*' signal. As discussed above, signals will not be entirely reliable. Indeed, accurate decisions could always be made on the basis of just one signal if they were always 100% reliable. Specifically, you should assume that the reliability of the evidence is such that a signal conveys the truth *on average two times out of three* and is misleading on average one time out of three.

You need to weigh the evidence using a 'standard of proof' that we will give you. To give you an idea of what we mean by this, a standard of proof that we use in an example later on is on whether 'there is reasonable evidence' that the merger will harm competition.

There are two types of decisions you will be asked to make:

1. One relates to situations where you have the opportunity to request more evidence and time to consider it (you can think of this as buying more signals). Such a request is costly to the firms. Competition agencies call this a referral to a second stage of investigation. You should only '*refer*' the merger for collection of more evidence if

you consider that the required standard of proof is met to justify doing so. The alternative is that you '*clear*' the merger (i.e. allow the merger on the basis that the required standard of proof is not met).

2. The second type of decision is where you are not allowed to request any more evidence, but must either '*prohibit*' the merger or '*clear*' it. A prohibition means that the merger cannot proceed. (Practitioners should also note that the prohibition decision should be taken to include clearance subject to effective remedies.)

You will be asked to consider a set of 24 scenarios, one for each round. Each scenario corresponds to a different and unrelated merger case. In each round, you are given a number of signals, a standard of proof and a type of decision (i.e. refer or clear; prohibit or clear).

We do not ask you to make a decision directly on the merger in each scenario. Rather, what we ask you to do is **to consider what is the minimum number of adverse signals** (i.e., indicating that the merger would harm competition) **out of those available, that you would need in order for you *not* to clear the merger** (i.e., to either refer or prohibit)**.**

An incentive for you to answer carefully

For the scientific value of this study, it is very important that you consider each scenario seriously and make your decisions carefully. To help you focus your attention, at the end of the experiment the computer will take 2 rounds out of 24. It will then determine how, for each of these rounds, your rule for deciding against the merger fares in the light of the available evidence when a merger is randomly proposed by the computer. If, based on the given standard of proof and the other available information, your rule correctly decides in favour of or against the merger, you earn 1 point. If a merger which should be decided against is approved, or if a merger which should be approved is decided against, you earn 0 points.

[*Treatment A only:*
Thus, you can earn between 0 and 2 points. Each point is worth 5 pounds, and so, if you make good choices, you can earn up to 10 pounds. You also earn a participation fee of 3 pounds.]

[*Treatments B and C only:*
You also earn 1 additional point for participation. Thus, you can earn between 1 and 3 points. On 4th August 2008, we shall randomly determine four winners among all participants. Each winner will be awarded a gift voucher equivalent to 250 British pounds (i.e., approximately 320 Euros or 490 US dollars) from an Amazon.com, Inc. retail website. Participants who have scored 2 points will have double the chance of winning relative to participants who have scored 1 point; participants who have scored 3 points will have three times the chance of winning relative to participants who have scored 1 point.]

Example Scenario 1

Number of signals received on whether merger will harm competition:	**20**
You are asked to **refer the merger for further investigation** if and only if **there is reasonable evidence** that the merger will harm competition.	
Minimum number of adverse signals **out of 20** needed in order for you to refer the merger for further investigation?	

Suppose that you answer '0' to this question. This means that, no matter what the signals say, you believe that there is reasonable evidence that the merger will harm competition, and so that it should be referred for further investigation.

Suppose that you answer '5' to this question. Then you are requiring at least 5 signals out of 20 that the merger will harm competition before referring it for further investigation.

Suppose that you answer '15' to this question. Then you are requiring at least 15 adverse signals out of 20 before referring the merger for further investigation.

Suppose that you answer '20' to this question. Then you are requiring every signal to be adverse before deciding there is reasonable evidence to refer the merger for further investigation.

Example Scenario 2

Number of signals received on whether merger will harm competition:	**120**
You are asked to **prohibit the merger** if and only if it is **very likely** that the merger will harm competition.	
Minimum number of adverse signals **out of 120** needed in order for you to prohibit the merger?	

Suppose that you answer '36' to this question. Then you are requiring at least 36 signals out of 120 that the merger will harm competition before prohibiting it.

Suppose that you answer '61' to this question. Then you are requiring at least 61 adverse signals out of the 120 you receive before prohibiting the merger.

Suppose that you answer '84' to this question. Then you are requiring at least 84 adverse signals out of the 120 before prohibiting the merger.

Suppose that you answer '120' to this question. Then you are requiring every signal to be adverse before prohibiting the merger.

Before you start making decisions, there will be a short questionnaire, the only purpose of which is to make sure you have understood the instructions. [*Treatment A only:* **Raise your hand when you have completed the questionnaire.**]

Appendix C: Calculation of Alpha

In the experiment, we assume agents adopt a dichotomous belief structure based around a hypothesis test (inferential expectations). They are asked how many signals suggesting an adverse effect for competition would be required to decide against a merger.

The number of signals, n, is pre-determined (25 or 100) and respondents are told that the signals are unreliable one third of the time. That is, even if the merger is really harmless, one would still expect on average one third of the signals to indicate that it is not.

Let s_i be a Bernoulli signal. It is unity if it suggests the merger is harmful (an 'adverse' signal) and zero otherwise. Under the null that the merger is not harmful, the probability p of an adverse signal is one-third. If the true state is that the merger would be harmful, then we assume that the probability of an adverse signal is two-thirds. Since the number of parcels of evidence is 'large' (25 is borderline and 100 is ample) we can invoke the Central Limit Theorem to calibrate the distribution of adverse signals under the null.

$$\sum_{i=1}^{n} s_i \sim N\{np,\ np(1-p)\}$$

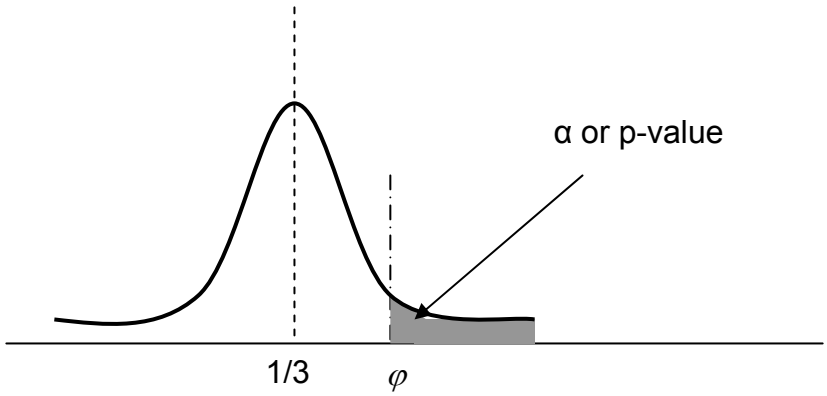
In the experiment, respondents declare how many adverse signals would be required to overturn the null H_0, in favour of the alternative H_1:[55]

$$H_0 : \textit{Merger is not harmful} \Rightarrow p = \frac{1}{3}$$

$$H_1 : \textit{Merger is harmful} \Rightarrow p = \frac{2}{3}$$

We back out our estimate of α from the distribution of Σs_i. We use this to calculate the probability of obtaining at least the declared number of signals, which is the p-value.[56]

Probability distribution of adverse signals under H_0



We compare this with a decision rule based only on the proportion to gauge if respondents take account of the volume of evidence. The two approaches differ because the

[55] Declaring a number of signals is equivalent to declaring a proportion for which the null would be rejected.
[56] We use the Bernoulli-Normal correction to allow for the difference between a discrete and continuous distribution. For example, if an agent rejects the null for 20 out of 25 signals, we work out the Normal Prob(Σs>19.5).

variance of the sample proportion, $p(1-p)/n$, is decreasing in n. In the following Table, we calculate the p-values corresponding to rejection of the null at various proportions.

α for various proportions (φ)
where agents reject H_0

n	σ_p	φ=0.40	φ=0.50	φ=0.60	φ=0.70
100	0.05	0.07	0.00	0.00	0.00
25	0.09	0.23	0.03	0.00	0.00

Thus the implied α when an agent rejects H_0 on 40% of the signals is 0.07 if n = 100, but it is a less cautious 0.23 if n = 25.

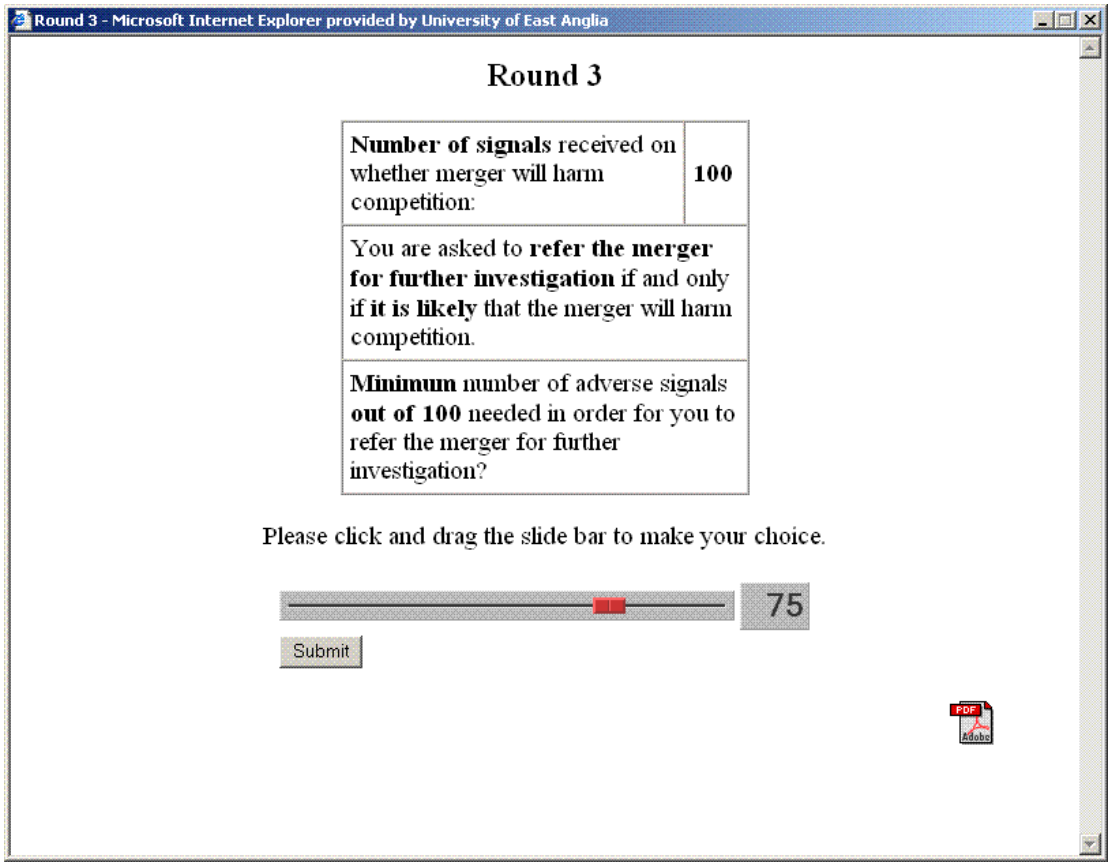
A subject's attitude to the volume of evidence can be ascertained in two ways. If the rejection proportion drops as she moves from n=25 to n=100, then it would appear that the agent is cognizant that a higher sample size allows more precise inference. If, however, the proportion does not drop by enough to preserve the alpha values, then it indicates that the agent has not taken enough account of the additional (four-fold) sampling power.

Agents who use very high proportions for their tests incur an increased probability of a type II error. As the following table shows, when agents start to nominate proportions in the vicinity of 0.6, this probability becomes non-trivial. The probability of a type II error (reject the null p=1/3 when in fact the alternative p=2/3 is true) is denoted β. The so called power, 1-β, contains essentially the same information.

β for various proportions (φ)
where agents reject H_0

n	σ_p	φ=0.40	φ=0.50	φ=0.60	φ=0.70
100	0.05	0	0	0.08	0.76
25	0.09	0	0.04	0.24	0.64

Figure 1. Sample Experimental Screenshots



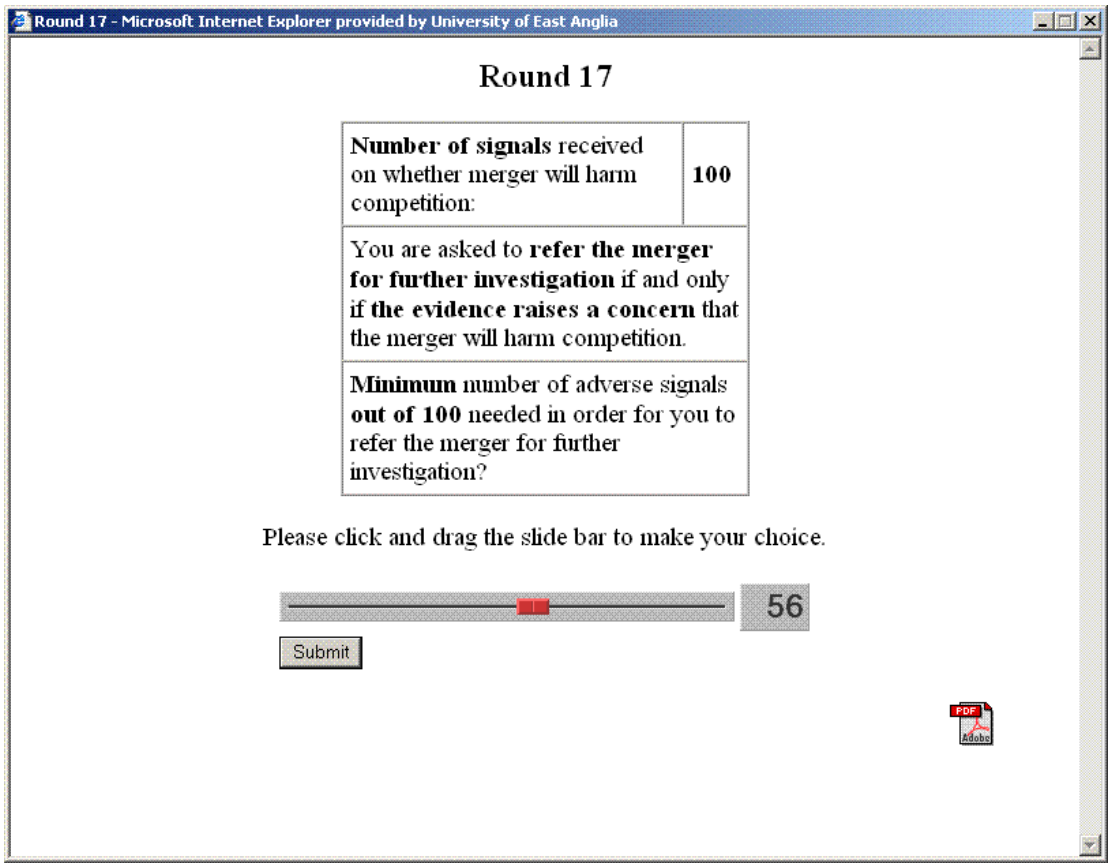
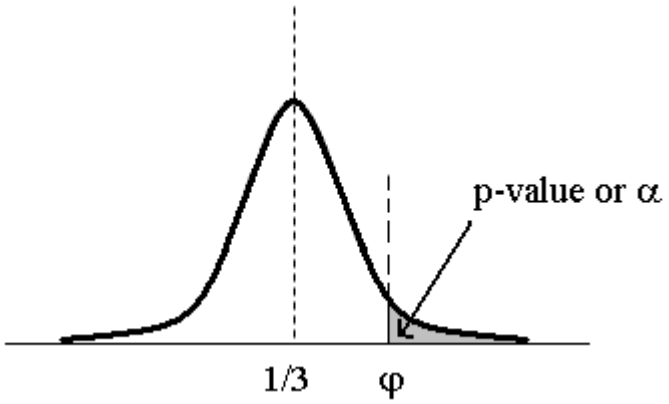
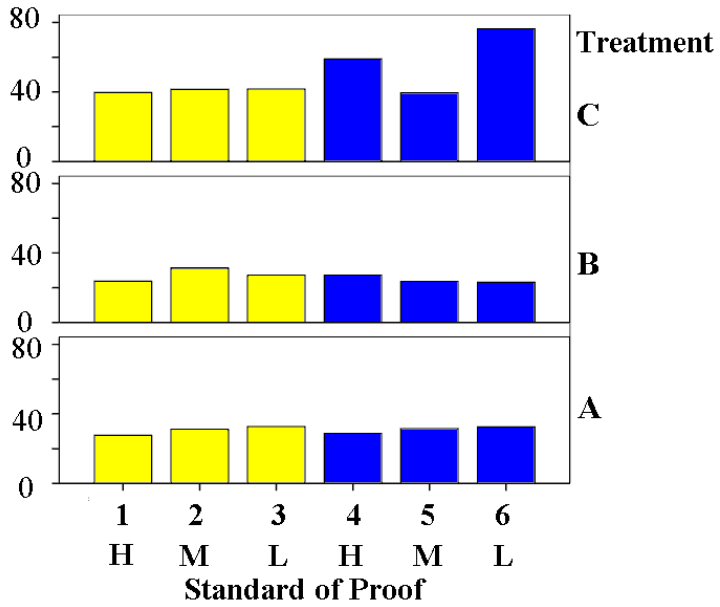


Figure 3. Probability Distribution of Adverse Signals



Notes: the graph assumes a null hypothesis of 'innocence' of the merger. That being the case, and if signals are incorrect 1/3 of the time, then the null hypothesis distribution will be centred around an average of 1/3 of the signals being adverse. This will hold as long as the proportion of adverse signals is not higher than φ. The test size α is given by the probability that the observed number of signals is greater than or equal to φ under the null hypothesis.

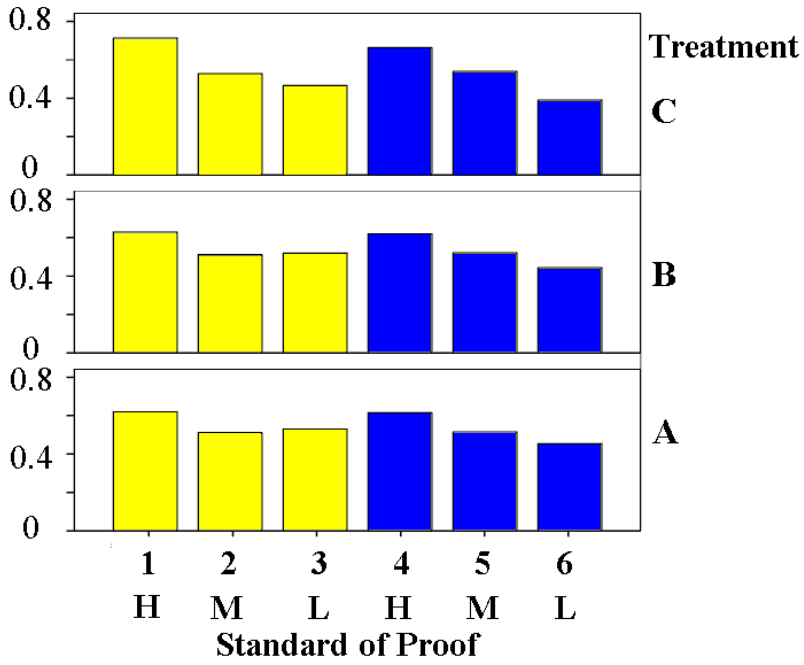
Figure 2. Average Time Taken to Respond to Each Question



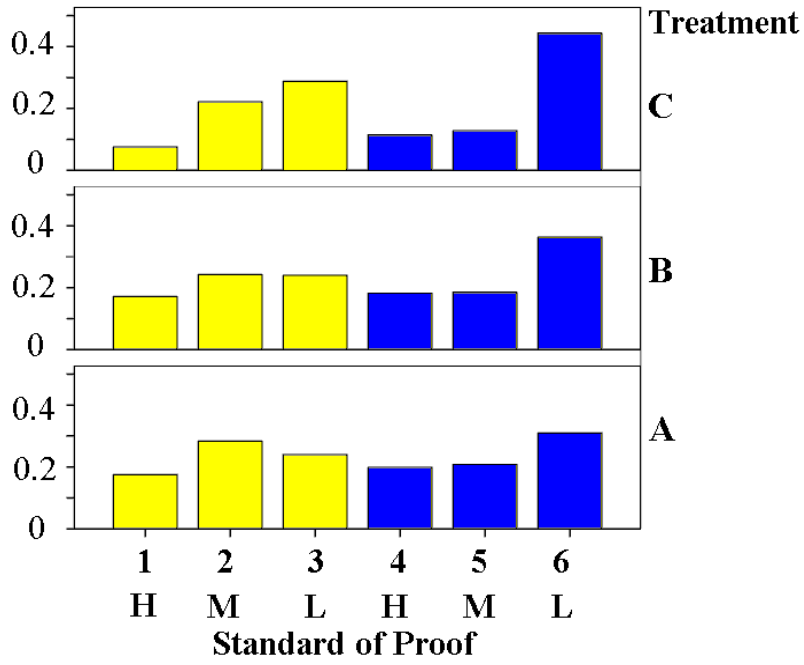
Notes: Treatment A involved students in an experimental lab; Treatment B involved students accessing the experiment by web interface; and Treatment C involved agency practitioners accessing by web interface. Standards of proof are as listed in Table 1: light and dark columns refer to probabilistic and evidence based standards, respectively; H, M and L respectively refer to a high, medium and low hypothesized standard of proof.

Figure 4. Average φ and α Values by Standard of Proof

(a) *Average φ values*



(a) *Average α values*



Notes: average fractions φ of required adverse signals to decide against the merger, and implied average α values (i.e. assuming statistical tests are conducted to determine φ), are provided. Treatment A involved students in an experimental lab; Treatment B involved students accessing the experiment by web interface; and Treatment C involved agency practitioners accessing by web interface. Standards of proof are as listed in Table 1: light and dark columns refer to probabilistic and evidence based standards, respectively; H, M and L respectively refer to a high, medium and low hypothesized burden of proof.

Table 1. Experimental Standards of Proofs

Index	Standard of Proof	Type of Standard	Hypothesized Standard of Proof
1	it is proved beyond reasonable doubt	Probabilistic	High
2	it is likely	Probabilistic	Medium
3	there is a realistic prospect	Probabilistic	Low
4	accurate, reliable, consistent and sufficient evidence (ARCSE)	Evidence based	High
5	the balance of evidence is	Evidence based	Medium
6	the evidence raises a concern	Evidence based	Low

Table 2. Average φ and α Values

(a) *Average φ values*

Treatment	Overall	Signals = 25	Signals = 100	Referral	Prohibition
A	0.549	0.552	0.547	0.528	0.571
B	0.545	0.551	0.539	0.534	0.556
C	0.561	0.571	0.551	0.525	0.598
Total	0.55	0.556	0.544	0.53	0.569

	Standard of Proof					
Treatment	1	2	3	4	5	6
A	0.62	0.512	0.53	0.615	0.516	0.454
B	0.63	0.512	0.52	0.621	0.521	0.443
C	0.714	0.527	0.466	0.663	0.539	0.39
Total	0.648	0.515	0.509	0.63	0.524	0.433

(a) *Average α values*

Treatment	Overall	Signals = 25	Signals = 100	Referral	Prohibition
A	0.229	0.237	0.222	0.249	0.21
B	0.226	0.237	0.214	0.24	0.212
C	0.199	0.201	0.198	0.239	0.16
Total	0.22	0.228	0.212	0.241	0.199

	Standard of Proof					
Treatment	1	2	3	4	5	6
A	0.175	0.283	0.241	0.199	0.208	0.31
B	0.171	0.243	0.24	0.182	0.185	0.363
C	0.076	0.222	0.288	0.114	0.128	0.443
Total	0.149	0.246	0.252	0.169	0.176	0.371

Notes: 1.Average fractions φ of required adverse signals to decide against the merger, and implied average α values (assuming statistical tests are conducted to determine φ), are provided both overall and classified by number of available signals (25 or 100), type of decision (merger referral or prohibition) and standard of proof (as listed in Table 1).

2. The relatively similar α values for standards 4 and 5 reflect the fact that averages are used. A few agents have very high αs (>0.5), and this pulls up these averages. The αs associated with high proportions (>0.5) are actually close to zero. The tables also average over n=25 and n=100, again making a simple mapping between proportions and αs problematic.

Table 3. Professional Background of Practicioners

Treatment		Economist	Lawyer	Neither
φ	Mean	0.56	0.571	0.547
	n	36	19	12
α	Mean	0.209	0.177	0.206
	n	36	19	12

Notes: average φ and α values by professional background are provided, together with corresponding sample sizes *n*.

Table 4. Tobit Random Effects Regressions on φ and α

(a) *Regressions on φ*

	Treatment A			Treatment B			Treatment C		
	β	*t*	*P*	β	*t*	*P*	β	*t*	*P*
Round	0.00993	3.13	0.002	0.00921	4.92	0	0.00562	2.39	0.017
Round Squared	-0.00031	-2.52	0.012	-0.00026	-3.53	0	-0.00018	-1.96	0.05
Decision Type	0.044	4.2	0	0.021	3.35	0.001	0.07	8.98	0
Nsignals	-0.009	-0.82	0.411	-0.015	-2.34	0.019	-0.021	-2.72	0.007
Standard1	0.176	9.65	0	0.19	17.64	0	0.326	24.36	0
Standard2	0.054	2.98	0.003	0.076	7.1	0	0.146	10.96	0
Standard3	0.078	4.29	0	0.082	7.62	0	0.08	6.03	0
Standard4	0.176	9.64	0	0.184	17.09	0	0.273	20.48	0
Standard5	0.06	3.29	0.001	0.083	7.76	0	0.15	11.26	0
Age	-0.002	-0.47	0.638	-0.002	-1.28	0.2	-0.004	-2.07	0.039
Gender	-0.026	-0.6	0.546	0.054	2.31	0.021	0.031	0.89	0.372
CmL National	0.074	0.8	0.424	-0.008	-0.18	0.858	-0.103	-1.36	0.175
ECN National	-0.066	-0.73	0.468	0.062	1.31	0.192	-0.232	-1.57	0.117
CmL Authority							0.029	0.38	0.704
ECN Authority							0.223	1.54	0.123
Economist							0.029	0.45	0.651
Lawyer							-0.001	-0.02	0.981
Constant	0.448	3.38	0.001	0.359	6.91	0	0.495	6.22	0
Log likelihood	129.098			479.04			569.554		
Sample size *n*	1368			3672			1608		

(b) *Regressions on α*

	Treatment A			Treatment B			Treatment C		
	β	*t*	*P*	β	*t*	*P*	β	*t*	*P*
Round	-0.00708	-0.96	0.335	-0.02263	-5.4	0	-0.01865	-3.1	0.002
Round Squared	0.00014	0.49	0.627	0.00066	4.06	0	0.00053	2.25	0.024
Decision Type	-0.074	-3.02	0.003	-0.039	-2.77	0.006	-0.138	-6.8	0
Nsignals	-0.125	-5.01	0	-0.126	-8.8	0	-0.09	-4.46	0
Standard1	-0.318	-7.39	0	-0.351	-14.16	0	-0.677	-18.04	0
Standard2	-0.095	-2.32	0.02	-0.184	-7.84	0	-0.332	-10.23	0
Standard3	-0.159	-3.85	0	-0.186	-7.93	0	-0.205	-6.45	0
Standard4	-0.293	-6.82	0	-0.332	-13.49	0	-0.584	-16.32	0
Standard5	-0.182	-4.42	0	-0.234	-10.01	0	-0.409	-12.64	0
Age	0.007	0.74	0.459	0.004	1.26	0.209	0.011	2.49	0.013
Gender	0.094	0.99	0.321	-0.088	-1.72	0.085	-0.054	-0.64	0.521
CmL National	-0.185	-0.89	0.372	0.057	0.61	0.543	0.26	1.41	0.159
ECN National	0.107	0.53	0.597	-0.186	-1.83	0.068	0.439	1.22	0.221
CmL Authority							-0.076	-0.41	0.682
ECN Authority							-0.442	-1.27	0.205
Economist							-0.159	-1.01	0.312
Lawyer							0.007	0.08	0.935
Constant	0.214	0.73	0.466	0.577	5.1	0	0.224	1.15	0.248
Log likelihood	-834.973			-2118.255			-755.55		
Sample size *n*	1368			3672			1608		

Notes: the regressions control for subject level non independence of observations at the level of subjects. The dependent variables are either the fraction φ of required adverse signals to decide against the merger, or the implied test size α value. *P* values are two tailed.