DOI:
10.1002/bdm.2073

Document Version
Peer reviewed version

Running Head: Reference points

Comparing what to what, on what scale?
The impact of item comparisons and reference points in communicating risk and uncertainty

Tim Rakow[1], Emily Blackshaw[1], Christina Pagel[2], David S. Spiegelhalter[3]

[1] King's College London, UK

[2] University College London, UK

[3] Cambridge University, UK

Corresponding Author

Dr Tim Rakow

Addison House, Floor 2

Department of Psychology,

Institute of Psychiatry, Psychology and Neuroscience,

King's College London,

London, SE1 1UL,

UK.

Email: tim.rakow@kcl.ac.uk

Tel: +44 (0)20 7848 6228

Main text: 9297 words (excluding Abstract, Tables, Figures and References)

Funding Acknowledgment: This work was funded by the National Institute for Health Research (NIHR) Health Services and Delivery Programme (project number 14/19/13). The views and opinions expressed in the paper are those of the authors, and do not necessarily reflect those of the Health Services and Delivery Research Programme, NIHR, UK National Health Service or the UK Department of Health.

Abstract

The mandated public availability of individual hospital's audit data for children's heart surgery in the UK creates a challenging scenario for communicating these complex and sensitive data to diverse audiences. Based on this scenario, we conducted three experiments with the aim of understanding how best to help lay people understand these data, and the practical goal of improving the public presentation of these data. The experiments compared different outcome measures for displaying the survival rate (percentage scale vs. the ratio of the predicted/observed rates) and presentation formats (individual hospital vs. all hospitals shown) for outcomes data presented relative to prediction intervals generated by a risk model that adjusts for case mix. Our data highlight how easily *inappropriate* comparisons can influence evaluations of complex data: for instance, both a survival ratio of 1 and the presence of other hospitals seemingly provided reference points that resulted in inappropriately harsh evaluations of some hospitals. By drawing on evaluability theory, we demonstrate how to enhance people's understanding of these complex data while also discouraging inappropriate comparisons, which has implications for communicating risk and uncertainty, and for choice architecture design in a range of contexts.

Keywords: risk communication; uncertainty; prediction intervals; reference-dependent evaluation; evaluability

Introduction

Judgments are rarely – if ever – made in isolation. Rather, estimates of quantities and the evaluations implied by people's choices are subject to a range of contextual factors (Parducci, 1968; Stewart, 2009). For instance, judgments often assimilate towards previously given 'anchor' values, an effect that has variously been attributed to: providing a start-point from which people adjust insufficiently (Epley & Gilovich, 2001); the selective priming of anchor-consistent evidence (Mussweiler & Strack, 1999); or changing the way we view the response scale (Frederick & Mochon, 2011). Sometimes, however, judgments can contrast away from a reference value; as when objects appear larger when placed alongside "small" objects than when placed alongside "large" ones (e.g., the "Ebbinghaus illusion") or when fruit juice tastes better after tasting diluted juice than after tasting undiluted juice (Zellner, Allen, Henley & Parker, 2006).

Thus, judgments are plastic. One reason for this is that evaluations do not depend solely upon the properties of the item under consideration, but are influenced by comparisons with other items. For example, Brown, Gardner, Oswald and Qian (2008) asked participants to rate their satisfaction with 11 different salaries ranging from £17K to £28K. The distribution of the intervening salary values was manipulated (between-subjects) to be either positively or negatively skewed, with the result that a salary of £22K was assessed more favourably when it ranked high in the salary set (negative skew set) than when it ranked low in the set (positive skew set). Workplace survey data reported by Brown et al. (2008) suggest that such effects have implications beyond the lab: with higher employee turnover and lower mean job satisfaction in companies with more positively skewed salary distributions.

Another source of comparison-based context dependence are the reference points that influence how an option's attribute dimensions are evaluated. For instance, framing outcomes (e.g., $-outcomes, health benefits) as potential losses in a choice or transaction rather than as potential gains seemingly increases the sensitivity to differences in values ('loss aversion'; Kahneman, Knetsch & Thaler, 1991) and makes preferences relatively more risk seeking (Kahneman & Tversky, 1979). Establishing a new aspiration value can have a similar effect, which can be understood as sub-aspirational outcomes being perceived as losses and supra-aspirational outcomes being treated as gains. For example, once a target or expectation is set for earnings, each dollar below that reference point weighs more heavily than each dollar above the reference point when earnings are evaluated, and therefore exerts greater influence on the behaviours that determine how much one earns (Camerer, 2000).

In this paper, we examine a domain in which this tendency to make relative judgments rather than absolute evaluations could be crucial: medical audit for surgical outcomes. Such audits are designed to monitor the clinical outcomes at an institution, but when these data become publically available they lend themselves to unintended, and potentially spurious, comparisons between institutions. We present

three experiments that speak to the role of context dependence in people's understanding and interpretation of such data, using the example of risk-adjusted survival rate data for heart surgery in children. Specifically, we examine how choices about the way these data are presented affect comprehension for, and evaluations of, these data – particularly when these 'design choices' are understood in terms of the comparisons that they encourage (or discourage) people to make. These experiments represent part of a larger project in which we aimed to find a better way to present complex data about medical outcomes to the general public (Pagel *et al.,* 2017); and our experiments were designed to inform and assess our progress towards that goal. In this piece of translational research, we use current theories of relative judgment and context-dependent preference as heuristic tools to assist in the design of our experimental manipulations and to guide how we should make sense of the data (rather than providing a test of those theories). Therefore, our experiments and findings have relevance beyond just the specifics of our project, and we believe that the lessons from this investigation can inform how to present information for a variety of medical decisions as well as informing the design of risk communications and choice menus in other domains.

Since 2000, the UK National Health Service (NHS) has collected data for every paediatric heart operation in the UK and Ireland. Using hospital data linked to independently verified life status information (from the UK Office of National Statistics), a 30-day survival status is assigned for each operation (i.e., whether the patient survived at least 30 days beyond surgery). These data are used by 14 individual centres to monitor their own performance; but are also collated by a national audit body that has published the survival data for each centre since 2013. The national audit body monitors the survival rates and intervenes if there is evidence that a hospital's survival rate is lower than it "should be".

A determination of what the survival rate "should be" for each hospital requires that each hospital's case mix is taken into account using risk adjustment (Pagel, Crowe, Brown & Utley, 2014) – i.e., adjusting for each patient's age, diagnosis, complexity of surgery, etc. Consequently, the audit body does not simply report "raw" survival rates for each unit. Rather, using a validated risk-adjustment model (Pagel et al., 2013) it reports raw survival rates *relative* to what is *predicted* for each hospital (given its case mix) and summarises the data as a survival ratio (actual survival rate / predicted survival rate). The audit body also places *control limits* around each hospital's survival ratio because we should expect some variation due to sampling variability or due to factors absent from the risk model that are unrelated to the hospital's standard of care. Specifically, the risk model is used to generate prediction intervals for each hospital. A central 95% prediction interval demarks a range of survival rates such that if a hospital's (underlying) survival rate is "as predicted" there is only a 5% chance that its survival rate (for the audit period) falls outside that range. An extended 99.8% prediction interval is also plotted which will exclude the observed survival rate only 0.2% of the time if sampling variability is the only source of deviation from the mean predicted survival rate of the risk model. When the survival rate falls outside the

prediction interval, this is described as having "survival higher/lower than predicted", or (more correctly) as indicating that there is "some evidence that the survival rate is higher/lower than predicted". When the survival rate falls beyond the *extended* prediction range, this is described as showing "survival *much* higher/lower than predicted" or "*strong* evidence that the survival rate is higher/lower than predicted". Figure 1 (right) displays the prediction intervals and observed survival rates for a hypothetical set of data. These intervals can be transformed to intervals for the survival ratios as shown in Figure 1 (left), which is the display used by the audit body. Thus, when a hospital is evaluated, it is evaluated against its own benchmarks – not against other hospitals. This point is critical to the focus of our studies, in which we examine the extent to which people's evaluation of these plots imply a comparison between hospitals (be that deliberate, or not).

It need hardly be stated that these data are complex and difficult to understand – even for the specialists whose work is informed by the data. However, these data are publically available and of interest to various audiences (e.g., families of children with congenital heart disease, journalists, service managers) – often with limited knowledge of medical statistics. It is therefore important that these data are presented in a way that can be understood and interpreted by a range of audiences. Consequently, in tandem with a recent update to the risk model (Brown *et al.,* 2017; Rogers *et al.,* 2017) we performed a series of experiments to examine the best way to present these data[1]. Drawing on our pilot work indicating that people had difficulty understanding the survival ratio measure, and suspecting that a survival ratio of 1.0 (which represents a "survival rate as predicted") might serve as a salient reference point for the evaluation of a hospital's data, our first experiment compared the current format of data presentation against a different graphic for presenting these data. This alternative graphic used the (raw) percentage survival rate as the outcome measure, which, compared to the survival ratio, should be both a more familiar type of measure (because percentages are used frequently) and a less complex measure (because it requires fewer steps of calculation than the survival ratio). We therefore predicted that this alternative percentage survival rate plot would be more straightforward to understand and interpret.

Experiment 1a – Comparing Reactions to Two Types of Prediction Interval Plot

This first experiment in our series was therefore designed to determine whether we could improve people's understanding of these complex data by presenting the survival-rate data on a percentage scale rather than as a survival ratio (as had been used in recent public communications of these data). Additionally, we wished to determine whether people's reactions to the data (e.g., their evaluation of each hospital) were affected by the choice of graphic used to present the data.

Method

[1] The explanatory website for these data, which incorporates the insights gained from these and other experiments in its design, can be seen at: http://childrensheartsurgery.info

Participants. Seventy-seven first-year undergraduate psychology students (69 female; all within the 18-29 year-old age-band) from a UK university volunteered as part of an optional course activity.

Design. The study employed an independent groups design, whereby participants were randomly assigned to one of two conditions. Participants were either shown plots with survival data displayed as ratios (observed/expected survival) as used in the annual audit report (ratio-plot condition, N = 39); or were shown plots displaying percentage survival rates which had been designed for use in an explanatory website (percentage-plot condition, N = 38). Examples of these plots are shown in Figure 1. Responses to questions about these plots were the dependent variables in the experiment (see Materials below). From the (broader) perspective of our project, this provided a test of whether the "new" (percentage) plot designed for the explanatory website improved upon the "old" (survival ratio) plot from the audit report. As such, this does not provide a "pure" test of a single design feature of these plots. Thus, in addition to changing the scale dimension (ratio vs. percentage) – which we regarded as the most important change – we also changed the orientation of the outcome scale (vertical to horizontal) to facilitate embedding the figure within a table on the explanatory website that we were developing, which showed information about each hospital in a separate row. (Subsequent to this experiment, we implemented this on our website.) Additionally, because pilot work had shown that some people mistook the prediction interval diagram (Figure 1, left) for a bar chart, we changed the shading of the figure to reduce the chance of this misinterpretation.

*****Figure 1 HERE*****

Materials. An online survey was constructed using *Qualtrics* software. The first part of the survey comprised demographic questions (age, gender and educational background). The next part presented information about survival rate data for paediatric heart surgery, which was adapted from the national audit report for these data (see *Supplementary Materials*). This information (approximately 850 words) was identical for the two conditions, with two exceptions. First, the graph shown at the end of the information differed according to the participant's assigned condition (ratio-plot vs. percentage-plot). Second, the sentence of information that accompanied this graph differed according to the graph that they saw, as follows:

> *Ratio-plot condition:* "Figure 1 shows on the Y-axis the survival ratio (actual survival/predicted survival) for all units, and the number of surgical 30-day episodes on the x-axis."
>
> *Percentage-plot condition:* "Figure 1 shows on the Y-axis the number of surgical 30-day episodes for all units, and the observed survival rate on the x-axis."

Thus, in keeping with the audit report that prompted our investigation, participants received definitions of the variables shown in the plot of the data, but no detailed explanation of the surgical outcome variable.

The main part of the survey comprised three categories of questions about these data: comprehension, interpretation and evaluation questions. All questions were accompanied by graphical displays of hypothetical (though plausible) survival rate data for 14 hospitals, with hospitals labelled with letters (A-N).

Question Set One consisted of a graphical display of hypothetical data, about which participants were asked comprehension questions (e.g., *What does the light grey shaded area mean?*). For Question Set Two, participants were asked for their interpretation of these data (e.g., *Please select which hospital(s) has a survival rate that is higher than predicted*). Question Set Three consisted of graphical displays of six hypothetical datasets (with order of dataset randomised). Questions asked for participant evaluations of the hospitals whose survival-rate data were displayed in each graph. There were four questions for each hospital, with ratings on a 1 ("strongly agree") to 7 ("strongly disagree") scale:

I am concerned about hospital X

I feel confident about hospital X

I would recommend using hospital X

I would discourage people from using hospital X

These responses were blocked by question for a given dataset (i.e., participants provided 14 separate "concerned" ratings for each of Hospitals A-N, before providing 14 separate "confident" ratings for each of these same 14 hospitals, and so on). To facilitate viewing the graph while also providing ratings, each set of 14 ratings was split into two subsets of 7 (A-G, H-N) which were displayed on successive pages of the on-line survey.

The datasets used in each of the question sets (*Table 1*) were carefully constructed (manipulated) to investigate a variety of possible interpretations of the data; allowing us to explore how key features of the graphical displays (e.g., control limits, sample size information) influenced participants' evaluations. All participants were shown all six data sets In Table 1, and answered the questions associated with them.

Participants completed the questions for a graph literacy scale (Galesic & Garcia-Retamero, 2010) at the end of the questionnaire. This scale has 13 questions that test people's facility with graphs and the conventions that usually govern the graphical display of data, and has demonstrated satisfactory levels of convergent validity and reliability.

*****Table 1 HERE*****

Procedure. Participants completed the study in individual testing rooms or in a quiet (sparsely populated) office area. Participants first read on-screen task instructions introducing the study, were given detailed information concerning their right to withdraw and anonymity, and were asked for their consent to participate. Participants completed the questionnaire; then participated in one of two

(shorter) studies (one reported as Experiment 1b; the other not reported here). Finally, participants completed an 'extended debriefing' exercise: an educational activity providing information about the study methods. Overall, the study session lasted approximately 50-60 minutes per participant.

Results and Discussion

Comprehension data. Table 2 summarises the responses to the comprehension questions, by condition. Accuracy for questions about the "labelling" of the regions of a prediction interval plot (Questions 1-5) were essentially at ceiling, with no significant differences between conditions. However, the question "What does the black dot mean?" (Question 6) provoked many errors, with significantly lower accuracy for the ratio-plot (41.0%) compared to percentage-plot (71.1%), $\chi^2(1, N=77) = 7.04$, $p = .008$. The most common answer to this question for participants in the ratio-plot condition was "observed survival rate" (N = 19, 48.7%), which was incorrect; plus two participants who responded with "other" gave "actual observed survival" and "actual survival rate" as free-text responses. Participants expressed more doubt than for other comprehension questions, with 3 participants in the ratio-plot condition and 4 participants in the percentage-plot condition selecting "I do not know".

The majority of participants in both conditions answered Questions 7 and 8 (about volume of procedures) correctly, with the mean accuracy for both questions (combined) being 88.4% for the ratio-plot condition, and 73.7% for the percentage-plot condition. When examined separately for each question, this difference in accuracy between the conditions did not reach significance, $\chi^2(1, N=77) = 2.24$, $p = .135$, and $\chi^2(1, N=77) = 3.34$, $p = .068$, for Questions 7 ("fewest") and 8 ("greatest") respectively. Almost all errors in the percentage-plot condition were reversing the hospitals that had performed the greatest/fewest number of procedures.

*****Table 2 HERE*****

An overall accuracy score was calculated for each participant by summing the number of correct responses to Questions 1 through 8 (possible range 0-8; observed range 4-8). The mean (SD) accuracy scores were 7.15 (0.74) for the ratio-plot condition and 7.03 (1.13) and percentage-plot condition. These means did not differ significantly, $t(75) = 0.59$, $p = .559$, $d = 0.13$.

Interpretation data. High levels of accuracy were observed for all interpretation questions (Table 3). For the purposes of these analyses, the correct answer for the question "please indicate hospital(s) with a survival rate higher than predicted" includes two possibilities: either only Hospital C (survival higher than predicted); or both Hospitals C (survival higher than predicted) and G (survival much higher than predicted). An overall interpretation score was calculated for each participant by summing the number of correct responses for the five interpretation questions (possible range 0-5, observed range 1-5). Mean interpretation scores did not differ significantly between the ratio-plot condition (M = 4.67, SD = 0.77) and percentage-plot condition (M = 4.63, SD = 0.97), $t(75) = 0.18$, $p = .861$, $d = 0.05$.

There was some consistency in the incorrect responses given for these questions. For Question 1, the majority of participants who answered incorrectly (N = 5) gave responses indicative of hospitals that were close the 1.00 ratio, or the centre of the predicted range. For Question 2, eight of the nine participants who answered incorrectly, included Hospitals C and G in their selection. Hospital G is in the survival 'much higher than predicted' range, whereas Hospital C is incorrect because it is in the survival 'higher than predicted' area. As expected, there was a significant positive correlation between the total scores for comprehension and interpretation, $r(75) = .33$, $p = .004$.

*****Table 3 HERE*****

Graph literacy. Participants' scores on the graph literacy scale were high, with a mean (SD) of 10.2 (1.5) out of 13. There was no significant difference in graph literacy score between conditions, $t(75) = 0.25$, $p = .804$, $d = .06$. There was not a strong correlation between graph literacy score and comprehension total score ($r(75) = .03$, $p = .778$), or graph literacy and interpretation total score ($r(75) = .18$, $p = .128$).[2]

Interim summary. Overall, participants displayed high levels of comprehension and appropriate interpretation of the data shown in the graphs; though we should note that our participants were academically successful and had some statistical training and so are not representative of the general population. Overall scores for comprehension and interpretation did not differ significantly between conditions, suggesting that for many purposes there is no clear advantage to using ratio-plots over percentage-plots, or *vice versa*. Importantly, however, participants in the percentage-plot condition were significantly better at identifying the 'meaning' of the black dot that signifies each hospitals level of performance. We regard this as a strong argument for using the percentage-plot display because it suggests that even people (such as our participants) who are used to interpreting quantitative data have difficulty in understanding the survival-ratio outcome measure.

Evaluation data. Responses to the four evaluation items (*concerned, confident, recommend* and *discourage*) were combined into an overall evaluation score for each hospital within each data set, by reversing the scoring for the *confident* and *recommend* items and then averaging the four item scores together. Thus, '7' indicates a highly positive evaluation and '1' indicates a highly negative evaluation. Overall evaluation scores displayed high internal consistency: the correlation between each pair of responses for a given hospital, in a given data set, was typically fairly strong. Inspection of approximately 250 correlation coefficients per item (6 data sets x 14 hospitals x 3 correlations) revealed median correlations of $r \approx .7$ between "feel confident", "would recommend" or "would discourage"; and of $r \approx .6$ for correlations with "concerned". This slightly weaker correlation between "concerned" and the other items seems to be attributable to a small number of participants (perhaps just 2 or 3) using this scale in a

[2] Participants also indicated their satisfaction with the data via four statements. Responses varied greatly between participants, but were not significantly associated with condition, comprehension or interpretation – and so were not analysed further or reported here.

counterintuitive manner (e.g., agreeing with "I am concerned about" while also agreeing with "I feel confident about" and "I would recommend" but disagreeing with "I would discourage"). Nonetheless the data strongly suggest that the overall evaluation scores tap into a single dimension of evaluation (e.g., Cronbach's α often exceeded 0.9 and only rarely fell below 0.8; indicating very good scale reliability).[3]

Inspection of these evaluation scores separately by hospital, dataset and condition suggest that participants were highly sensitive to the location of the black dot with respect to its position on the outcome scale of the plot, and sensitive to the boundaries that separate the regions of the prediction-interval plot in a manner consistent with the labels attached to regions outside the predicted range. These hospital evaluations are reported separately for each dataset and condition in the *Supplementary Materials* (for reasons of space) and were subjected to a combined analysis as described next.

An items analysis was conducted, treating each mean evaluation for a given hospital in a given data set as a data point, with separate data points for the ratio-scale and percentage-scale plots.[4] This allowed detailed comparison between conditions; and showed that the pattern of means was very similar between conditions, with the item means correlating $r = .97$ between the two conditions. There was, however, a small but reliable tendency for more positive evaluations in the percentage-plot condition; with the mean (SD) evaluation score being 4.95 (1.02) in the percentage-plot condition and 4.73 (1.02) in the ratio-plot condition. These means differed significantly, $t(83) = 7.57$, $p < .001$, $d = 0.22$.

Next, we analysed the mapping between the survival outcomes for the hospitals that participants were shown, and the mean evaluation for each hospital that participants subsequently provided. Figure 2 shows mean evaluation (y-axis) as a function of each hospital's outcome (x-axis); with ratio-scale evaluations on the left and percentage-scale evaluations on the right. The upper pair of figures plots the data as a function of the outcome measure that participants saw when making their evaluations; while the lower pair of figures plot the data according to the outcome measure displayed in the other condition. Three regression lines are shown on each figure: Loess regression lines (solid red) fitted to the full data set; and two linear regression lines fitted separately for those hospital with action/predicted survival ratios above 1 (solid green line) and those with survival ratios below 1 (broken blue line). Note that a survival ratio above/below 1 also corresponds to a black dot shown above/below the mid-point of the predicted range; consequently, some information about survival ratio can (in principle) be inferred

[3] The responses to these four items showed sufficient internal consistency to suggest that future work could obtain acceptably reliable evaluations from a pair of items. The "recommend" item had the highest item-scale correlation for about 50% of the overall evaluation scores, suggesting it as the single best item. It could be paired with either one of the items that correlated more highly with the overall evaluation score (i.e., "confident" or "discourage". The concerned item had the lowest item-scale correlation for over 75% of the sets of responses (likely for the reasons discussed above), which suggests that it could be eliminated from future studies without harming internal consistency.

[4] There was missing data for five participants on two datasets (1a and 2) due to a coding error (which was rectified after one day of data collection).

from a percentage scale plot (cf. Figure 2d) whilst it is impossible to infer a hospital's percentage survival rate from a ratio scale plot (cf. Figure 2c).

******Figure 2 HERE******

Figure 2a shows that the ratio-scale plot yields evaluations that are very closely tied to the survival ratio, but, notably, the steepness of this relationship changes at a point close to a survival ratio of 1 (with the Loess curve switching from following one linear regression line to following the other). One could regard this as a 'penalty' for hospitals with survival ratios below 1; or as more discriminating evaluations when the survival ratio is below 1 in comparison to when it is above 1.

The mapping between hospital outcomes and evaluations follows a similar pattern for the percentage-scale plot, though evaluations are not so closely tied to the outcome measure plotted (percentage survival) as was the case for the ratio-scale plot. Note also that the 'penalty' for lower outcomes does not 'kick in' quite so early (Figure 2d) – with the break-point being around a survival-ratio of 0.995 rather than around 1.0. Seemingly, some tolerance or 'grace' is shown when evaluating points that sit a little below the middle of the predicted range on the percentage plot (relative to the ratio plot).

This tolerance associated with the percentage plot is also implied in Figure 3. This shows evaluations from both types of plot (ratio plot = diamond icons; percentage plot = open triangle icons) as a function of the position of the back dot relative to the predicted range (0 = lower 97.5% limit; 0.5 = mid-point; 1.0 = upper 97.5% limit). Black dots towards the end of the predicted range, or outside the predicted range yield similar evaluations for both kinds of plot, while black dots around the middle, or slightly below the middle of the predicted range yield more positive evaluations when presented via the percentage-plot. (Figure S2 in the *Supplementary Materials* shows further analyses of these data.)

******Figure 3 HERE******

Using multiple linear regressions for these data, we regressed *both* the survival ratio and percentage survival rate onto mean evaluation, doing this separately for each condition. This was done to further investigate whether participants' evaluations reflect more than just the numbers plotted on the outcome scale of the plot that they saw. In both regressions, the proportion of variance accounted for by the two predictors was sizeable, and this barely differed between conditions (ratio-plot R^2 = .929; percentage-plot R^2 = .922; both $p < .001$). Both predictors were significant in each regression, but the regression coefficients differed between the two analyses. In the ratio-plot condition, the coefficient ("weight") for survival ratio (β = .81, $p < .001$) was substantially higher than the coefficient for percentage survival (β = .18, $p < .001$); while, in the percentage-plot condition, the coefficients were

similar for the survival ratio (β = .54, p < .001) and percentage survival (β = .46, p < .001).[5] These findings are intriguing given that participants only saw *one* type of plot (though which plot they saw varied between conditions) and therefore only saw a scale displaying values for *one* of these outcome measures. This provides further evidence that participants extracted information from the plots above and beyond the values displayed on the outcome scale, and therefore incorporated information gleaned from the relative position of a hospital's outcome score within its prediction interval. For example, when viewing a percentage plot, one can estimate the survival ratio from where the black dot is relative to the bounds the prediction interval. We assume that participants did *not* undertake spontaneous and precise calculations of this nature; nonetheless, based on the mapping examined in these regressions it is *as if* they did so.

Summary of evaluation data. From these exploratory analyses we might characterise the evaluation data as follows. Hospitals with survival higher than predicted (but not significantly so) reap a progressive 'boost' to their evaluations as outcome rates increase, while hospitals with survival lower than predicted (but not significantly so) reap a progressive 'penalty' to their evaluations as outcome rates increase. However, consistent with the greater sensitivity to negative outcomes than to positive outcomes that is often reported (i.e., *loss aversion* whereby "losses loom larger than gains", Kahneman et al., 1991; Kahneman & Tversky, 1979), the progressive punishment for lower-than-predicted outcomes seems to be greater than the progressive reward for higher-than-predicted outcomes. Indeed, the plotted data bear a striking resemblance to the prospect theory value function which 'kinks' at a reference point as sketched out by Kahneman and Tversky. Importantly, on this reading of the data, the ratio-plot and percentage-plot seem to have a different reference point on the outcome scale that subjectively distinguishes 'good' from 'poor'. The rate-change in evaluation occurs close to the mid-point of the prediction interval (i.e., around survival ratio = 1.0) when outcomes are plotted as a survival ratio, while the point of rate change in evaluation is about a quarter of the distance into the prediction interval (from the lower limit) when outcomes are plotted as a percentage. Thus, potentially, the ratio-scale plot could promote undue concern about hospitals whose outcomes are within the lower half of the predicted range (where there is *no* evidence that the hospital is performing poorly).

Experiment 1b – Understanding the Reactions to Prediction Interval Plots

This second study (run alongside Experiment 1a) continued our comparison of ratio plots and percentage plots for presenting survival data and the corresponding prediction intervals. It employed methods that differed from those of Experiment 1a. Using a mixture of simple problem solving tasks and

[5] Because the two predictors in these regressions correlate highly (r = .82) we caution against treating these regression coefficients as precise estimates, because estimated regression coefficients are unstable when predictors are highly collinear. Nonetheless, using moderated regression, we did confirm that regression weights for the survival ratio differed between conditions (p < .001), as did those for percentage survival (p < .001).

a "think-aloud" question, we aimed to uncover some of the thinking that underpins people's interpretation and evaluation of these data.

Method

Participants. 49 participants who had participated in Experiment 1a took part: 24 participants from the ratio-plot condition and 25 participants from the percentage-plot condition.

Materials. Paper based questionnaires were constructed, which included background information about prediction intervals (adapted from an explanatory website under development, see Footnote 1), for participants to read and refer to (see *Supplementary Materials*) and a graph as appropriate to their assigned condition (see Figure 4). Participants answered two comprehension questions, two inference questions, one question regarding the judgment of "acceptable" outcomes, and a sixth question asking participants to explain their previous answer (Table 4). For Questions 1-5, participants marked a dot on the graph to indicate their answer for hospitals where survival rates/ratios had been omitted (Figure 4).

Procedure. After completing the main study, the researcher showed participants two graphs (ratio plot, percentage plot) and participants indicated which one they had seen in the previous study. Contingent upon their answer, participants were assigned to the same condition (ratio-plot or percentage-plot) as Experiment 1a. Participants were told that the exercise should take approximately ten minutes, but reminded that there was no time limit. Participants read the background information and viewed the graph appropriate to their condition. Participants completed the exercise in individual rooms or a quiet sparsely-populated office space and then were thanked and debriefed.

*****Figure 4 HERE*****

*****Table 4 HERE*****

Results and Discussion

Comprehension and Inference. Table 4 illustrates that participants were highly accurate in their responses to Questions 1-4. All participants correctly answered the first comprehension question; and all but two participants (both in the ratio-plot condition) answered the second comprehension question correctly. When scoring Questions 3 and 4, a mark in the 'survival much lower than predicted' or 'survival lower than predicted' regions was treated as correct, and a mark elsewhere is treated as incorrect. Almost all participants answered Question 3 correctly, marking a dot in the 'survival much lower than expected' area (58% ratio-plot; 56% percentage-plot) or 'survival lower than expected' area (38% ratio-plot; 40% percentage-plot). One participant in each condition marked a dot in 'survival much higher than predicted' – which, arguably, *should* warrant contact from the audit body to determine whether there is good practice to be shared (though this is not current practice, or what was described in the information that participants read). For Question 4, almost all participants answered correctly, marking a dot in either 'survival lower than predicted' area (45.8% ratio-plot; 40% percentage-plot), or

'survival much lower than predicted' area (45.8% ratio-plot; 48% percentage-plot). None of the differences in accuracy between conditions were significant for these four questions, all $\chi^2(1, N=49) < 2.17$, all $p > .141$. Thus, in keeping with Experiment 1a, the responses to these comprehension and inference questions imply a good understanding of some of the key features of prediction-interval plots.

*****Figure 5 HERE*****

Judgement of minimum acceptable survival rates. Of the 44 (out of 49) participants providing a unique numeric response, the majority placed their threshold (for the "lowest acceptable survival rate") inside the survival-as-predicted area (65% ratio-plot; 52% percentage-plot). This implies a more stringent standard than the audit body applies, which the participants had been informed of and seemingly understood (given their answers to the two preceding questions). This aligns with our interpretation of Experiment 1a whereby survival rates are regarded as 'poor' when they are 'as predicted' but are below the predicted (point-estimate) survival rate – especially for the ratio-plot. The next most common area for the threshold was 'survival much lower than predicted' (21.7% ratio-plot; 19% percentage-plot), followed by 'survival lower than predicted' (13% ratio-plot; 28.6% percentage-plot). There was no difference between the conditions for the location of these thresholds (Mann-Whitney U = 211.5, p = .423). Figure 5 shows the exact thresholds by condition. The mean threshold in the ratio-plot condition of 0.989 was almost identical to the lower bound of the survival-as-expected region (0.99); as was the mean threshold of 95.7% in the percentage-plot condition (lower bound of 95.8%).

Content analysis of threshold explanations. Participants' responses were coded according to the presence/absence of eleven themes (a given answer could attract a 'present' code for more than one theme; see Table 5). Here, we highlight some key messages from this analysis: beginning with two of the more common themes, the prevalence of which also differed between conditions.

*****Table 5 HERE*****

Some participants indicated that they not only compared an observed survival rate to the predicted ranges for the identified hospital, but also to the predicted ranges for other hospitals (N = 15, Theme 5). One participant with a threshold in the 'survival much lower than predicted' range explained their choice: *"this is because, while it is out of the predicted range for the hospital itself, it is within the ranges for the other hospitals"*. Other examples of this type of explanation include: *"97.5%, because it would still lie within the 'survival as predicted' range, but not fall below a rate lower than 'average' of the other hospitals"*, and *"anything lower than the lowest range out of all the hospitals (hospital B) would to me be considered unacceptable".* There was a striking difference between conditions for this type of explanation: of the 15 participants who provided an explanation of this type, significantly more (N = 12) were from the percentage-plot condition, $\chi^2(1, N=44) = 7.86$, $p = .005$. This suggests that percentage-plots may encourage inter-institutional comparisons – which in this context – are *inappropriate* due to the variation in case mix between hospitals. This could be due to a number of factors, including

incidental features such as the horizontal orientation of the graph that we used. One possibility is that having a better understanding of percentage survival than the survival ratio measure – which is an important difference we identified in Experiment 1a – actually encourages people to 'do more' with the measure. Thus if percentages are straightforward to understand they may facilitate comparisons based on that measure. Further research is required to determine how to discourage inappropriate comparison between hospitals – but without harming people's understanding of the data.

Also, emphasising the importance of being within the 'survival as predicted' range (Theme 10) was significantly more common in the ratio-plot condition (N = 15) than in the percentage-plot condition (N = 7), $\chi^2(1, N=44) = 5.37$, $p = .02$. Thus, in tandem with percentage plots seeming to encourage inappropriate comparisons between hospitals, ratio plots may encourage appropriate consideration of survival rates *relative to* a hospital's own prediction interval; though this may partly be associated with the fact that more participants in the ratio-plot condition (N = 15) placed their threshold in the 'survival as predicted' range compared to the percentage-plot condition (N = 11). The prevalence of the other themes (Table 5) did not differ significantly between conditions.

Analyses were also conducted to examine the association between responses to Question 5 (i.e. the area in which a dot was marked to indicate lowest acceptable survival) and theme(s) present in the explanation thereof. One, perhaps unsurprising, finding was that threshold location was significantly associated with emphasising the importance of being within the 'as predicted' survival range (Theme 10); this theme being most prevalent when thresholds were within the survival-as-predicted area, $\chi^2(2, N=44) = 12.81$, $p = .002$. There was an association between comparing data against other hospitals (Theme 5) and location of lowest acceptable threshold, $\chi^2(2, N=44) = 6.40$, $p = .041$, which reflects a tendency to mark a lower threshold when explicit comparisons with other hospitals are made. There was also an association between taking the total number of procedures into account (Theme 4) and location of lowest acceptable threshold, $\chi^2(2, N=44) = 8.16$, $p = .017$. To provide an interpretation of this association, we note that none of the five participants who mentioned the number of procedures put their threshold in the survival-as-predicted range (whereas 26/39 of the other participants did so). Interestingly, a few participants seemed to be of the opinion that when hospital treated a larger number of cases, it should have a higher survival rate: *"a survival ratio below 0.97 for me is unacceptable, especially in hospital M where there [are] 1481 surgical episodes which is a lot",* and *"but when looking at hospital M (1481), a small decrease in % means a higher error in patient prediction"*.

Participants sometimes acknowledged that a hospital could fall outside of the predicted range by chance (N = 10, Theme 6). Some participants acknowledged chance, but chose to discount this information, *"if a score were to be below its prediction, it could be a sign that something is wrong within the hospital (ignoring the possibility that a low score could be due to chance)"*; *"although being outside of the predicted range could be a result of chance; it seems safer to trust the hospitals who have matched*

their predicted survival rate" and *"even if it is due to chance they should aim to improve survival rate."* Such reasoning represents a potentially important feature of people's reasoning about uncertainty: people may acknowledge chance or sampling variability as a *potential* cause yet be unwilling to accept this as a *satisfactory explanation* of the data.

Experiment 2 – Comparing Joint and Separate Evaluation of Prediction Interval Plots

In Experiments 1a and 1b, we observed some key differences in comprehension, interpretation and evaluation between the two plots that we examined. We judge that the main cause of these differences was changing the outcome scale from a survival ratio to a percentage. We favour this conclusion because the main difference in comprehension in Experiment 1a was for a question directly about that scale, and because the differences in interpretation and evaluation between conditions that we found map neatly onto the inherent differences between those two scales. Thus, while it is possible that changing the orientation, colour or shading of the plots makes some difference, we feel it less likely that these explain much of the difference (because, for example, comprehension and interpretation of features unrelated to the outcome scale differed little between conditions).

The findings of Experiments 1a and 1b highlight the importance of the comparisons that people can make on the basis of the information they have been provided. Seemingly, when presented with outcomes data as a survival ratio, our participants treated a ratio of 1.0 as a strict expectation and were harsher in their evaluation of outcomes falling below that level than when the data were presented as percentage survival rates. This occurred despite the fact that some variation either side of that ratio is expected if a hospital's survival rate is 'as predicted'. Moreover, participants had difficulty selecting an appropriate definition of the survival ratio (represented as a dot on the plot) – which calls into question just how successfully they can use prediction-interval data presented in that format to make informed decisions. On the other hand, while a percentage survival rate seems easier to understand, this may encourage inappropriate comparisons (indeed, perhaps even *because* this measure is more easily understood). Thus, in Experiment 1b, participants working with percentage survival rate plots were more inclined to explain their judgments on the basis of comparisons between hospitals and (accordingly) less inclined to mention comparing a hospital's outcomes against its prediction interval. However, comparing a hospital's percentage survival against its prediction interval (and not against other hospitals) is exactly what *should* be done (for the very reasons that call for risk-adjusted data and the use of prediction intervals).

In Experiment 2, we therefore sought to test whether the tendency for inappropriate comparisons between institutions could be reduced. To do so, we turned to evaluability theory (Hsee, 2000; Hsee, Loewenstein, Blount & Bazerman, 1999; Hsee & Zhang, 2010). Chris Hsee and colleagues have observed that when evaluating options, decision makers can seemingly rely on option attributes that they know to

be less important than other attributes. This occurs when the 'important' attribute is difficult to evaluate. For example, in a between-subjects comparison, participants stated higher willingness-to-pay (WTP) for a 5-ounce cup of ice-cream overfilled with 7 ounces of ice-cream compared to a 10-ounce cup underfilled with 8 ounces of ice-cream (Hsee et al., 1999). Unsurprisingly, when presented side-by-side (i.e., evaluation within-subjects) WTP was higher for the cup containing more ice-cream. An explanation of this (in terms of evaluability) is that everyone 'knows' that more ice-cream is better; but in the absence of a good feel for the per-ounce value one is forced to rely on a less important cue to price (e.g., overfilled cups are 'good', and under-filled cups are 'bad') if one cannot compare options side-by-side. Slovic and Peters (2006) reported a similar effect: between-subjects, participants rated a project with a 98% chance of saving 150 lives as better than a project that will save 150 lives. This also fits with evaluability theory. Everyone agrees that saving lives is important, but what value to place on X lives is notoriously difficult. It seems that this evaluation may be made easier by explicit reference to a high percentage: because everyone 'knows' that 98% is a good score even if they cannot say what one life or 150 lives are worth.

In Experiment 2, we therefore tested the effect upon evaluation of removing the possibility of the 'easy' comparison between hospitals, by the simple act of having participants evaluate one hospital at a time. Given the considerations discussed above, we were particularly interested to see what effect this had for those hospitals with survival rates that are below those of several other hospitals but which are still within the "outcomes as expected" range.

Method

Participants. 65 participants (54 female; all within the 16-54 age-band, 63% aged 16-24) were recruited from a UK university research volunteer list. Most were university students or employees; 63% had a university degree, though 8% had two years or fewer of post-16 education.

Materials and Design. An online survey was constructed using *Qualtrics* software, having a similar structure to Experiment 1a: demographics; audit report excerpt of approximately 850-words describing the audit procedures and data, and including an example of the type of graph (percentage plot) that participants would later see; comprehension questions; then evaluation questions. For this experiment, the comprehension questions (Table 6) were divided into two sets and interpolated by the provision of additional guidance (from our explanatory website) on prediction interval plots. The presentation order of these two sets of three questions was counterbalanced. This test of comprehension was included to test whether we could help people achieve a good understanding of the "basic" features of a prediction plot (appreciating that any set of complex data requires some explanation to the reader). Because these data are not relevant to the primary focus of this paper (i.e., the role of comparisons) we report them in the *Supplementary Materials*.

The stimuli for the evaluation questions were four of the six hypothetical datasets from Experiment 1a (1b, 2, 3 and 4; see Table 1) – always presented with percentage survival rate as the outcome measure – with order of dataset randomised. For each hospital, participants answered two of the four evaluation questions that had been used in Experiment 1a (*I feel confident about hospital X; I would recommend using hospital X*) with ratings on a 1-to-7 scale ("strongly disagree" to "strongly agree"). Each question was presented individually for a given hospital (i.e., only one question on the screen at a time) but the (sequential) presentation of these questions was blocked by question (order of questions was randomised) within dataset (order of hospitals within dataset was randomised). Thus, a participant progressed through the data sets, the order of which was randomised. For each data set, the participant answered a block of 'confident about' and a block of 'would recommend' questions, and those blocks could be in either order. And, within each block of questions, the participant was asked about each hospital, with the order of those hospitals randomised.

For the main manipulation in this experiment, participants were randomly assigned to one of two conditions. Participants saw plots showing data for all 14 hospitals in a dataset (combined condition, N = 32); or saw plots showing only the data for one hospital at a time (individual condition, N = 33). Thus when answering an evaluation question, participants in the individual condition only saw the data pertaining to the hospital identified in the current evaluation question, while participants in the combined condition saw plots similar to those in the percentage-plot conditions of Experiments 1a and 1b (e.g., Figure 1 right, or Figure 4 right).

Procedure. Participants completed the study in individual testing rooms following the same procedure as Experiment 1b; except that the debrief was a standard verbal debrief and the study session (which included a subsequent study, not reported here) lasted approximately 40-45 minutes per participant.

Results and Discussion

Evaluation scores. Responses to the two evaluation questions ("feel confident" and "recommend") were averaged together to create an overall evaluation score for each hospital within each dataset. These mean evaluations are summarised in Figure 6 (below) and are analysed in more detail (separately by dataset) in the *Supplementary Materials*. As per Experiment 1a, we used an items analysis to perform a 'global' comparison between the combined condition (joint evaluation) and the individual condition (separate evaluation). This involved finding the mean evaluation score for each hospital, separately for each data set and for each condition (14 x 4 x 2 = 112 data points in this analysis). This revealed that the mean evaluation score was higher in the individual condition (M = 5.24, SD = 0.87) compared to the combined condition (M = 5.02, SD = 0.97). This difference was small (d = 0.24) but statistically significant, $t(55) = 7.95$, $p < .001$. Figure 6 illustrates that the discrepancy is largely absent for hospitals with the very

highest survival rates, and may be more prominent for hospitals with lower survival rates. This reading of the data is corroborated by more detailed analyses reported in the *Supplementary Materials*.

******Figure 6 HERE******

Thus, presenting survival-rate data with accompanying prediction intervals for one hospital at a time results in more positive evaluations (relative to presenting survival-rate data and prediction intervals for several hospitals), and it seems particularly so for hospitals that have survival rates that fall towards the lower end of the distribution. Our favoured interpretation of this is that individual (separate) presentation of the data makes comparison to other hospitals difficult and forces participants to (appropriately) compare a hospital's survival rate to its own prediction interval. This results in more positive evaluations for those hospitals (that do not have the highest survival rates) which would otherwise suffer in a comparison against other hospitals.

However, one might ask whether it is appropriate that these hospitals, which do not have the highest adjusted survival rates, receive relatively positive evaluations. Perhaps these hospitals should not be evaluated positively? The rationale for using prediction intervals speaks to that question as follows. If a hospital's survival rate is within its prediction interval, we have no evidence that its survival rate is lower than predicted; and so this hospital should not receive an unduly negative evaluation solely on the basis of its survival rate. In contrast, if a hospital's survival rate falls below its predicted range there is some evidence that its survival rate is lower than predicted (given its case mix). This is a cause for concern, would trigger an investigation of a hospital's practice and performance, and therefore a negative evaluation is not inappropriate. Further inspection of the data suggests that participants' judgments in the individual condition may be broadly consistent with this nuanced view of how hospitals *should* be evaluated: the more "generous" evaluation of hospitals in the individual condition is largely restricted to hospitals having a survival *within* their predicted range (*Figure S3, Supplementary Materials*). A more detailed analysis in the *Supplementary Materials* is consistent with this interpretation: significant pairwise differences between conditions for a given hospital rarely occur when survival rates fall outside the predicted range, but are somewhat common when survival rates fall within the predicted range.

General Discussion

What are the lessons from our investigation? A first set of lessons relate to people's understanding and use of prediction interval plots, and may be restricted to that issue. A second set of lessons relate to the importance of comparisons in evaluating data, and have potentially wide-ranging implications for communicating risk and uncertainty, and for decision making in many domains.

First, we note that prediction interval plots and the accompanying survival rate data are not easy to read and interpret: errors in comprehension were seen in all three studies. However, with a careful choice of outcome measure (Experiment 1a) and some suitable explanation of these plots (Experiment 2), lay people can learn definitions and labels for the key features of these plots, and make broadly appropriate inferences from them.[6] Thus, with appropriate care in providing explanation, the advantages of using these kind of plots, such as avoiding creating inappropriate 'league tables' of hospitals (Spiegelhalter, 2005), can be made accessible to non-statisticians.

Different interested parties – including regulators, doctors, patients and their supporters, or journalists – might use these data in a variety of ways, and there is no 'correct' manner of presentation that can satisfy all their needs or requests. There is a practical demand for specific thresholds for action by authorities, which means that somewhat arbitrary boundaries need to be set, precisely analogous to the boundaries for determining statistical significance. While finer comparisons between hospitals are possible – for example there is certainly information in a hospital's position within its central interval – this is discouraged both because the dataset is not intended as a basis for patients choosing hospitals, and any attempt at ranking by journalists (or doctors) would be spurious. Therefore consideration of a hospital's performance relative to its own 'goalposts' appears a pragmatic compromise suitable for all parties.

Second, our data give support to the adage that 'all judgments are relative'. In Experiment 1a we saw that when patient outcomes were summarised as a survival ratio, evaluations seemed to be highly dependent upon the comparison of the each hospital's survival ratio against 1.0. This yields a pattern of evaluations akin to loss aversion (Kahneman et al., 1991) which has the result of 'penalising' hospitals with survival rates 'as predicted' but which are descriptively – but not reliably – below the point prediction for their survival rate. This chimes with other research that finds adverse effects on the interpretation of data in graphics that emphasise a central tendency measure. For example, Rakow, Wright, Spiegelhalter and Bull (2015) found that displaying a population mean on a funnel plot encouraged inappropriately fine discrimination between data points lying either side of that reference line; and Broad, Leiserowitz, Weinkle and Steketee (2007) reported that people became unduly fixated on the 'most likely' path of a hurricane when this was shown on a meteorological forecast map and consequently (potentially to their detriment) discounted other highly likely possibilities.

This tendency for the mid-point of the prediction interval to act as a reference point was reduced by the use of a more familiar and more readily understood outcome measure: the percentage survival rate (see *Figure 1*). This, however, seems to reduce the (appropriate) tendency to compare a hospital's survival rate to its own prediction range, and increase the (inappropriate) tendency to compare hospitals

[6] After receiving written guidance on how to "read" a prediction plot, participants in Experiment 2 achieved almost 90% accuracy on a series of 6 questions designed to test comprehension for the information presented in these plots.

on their raw survival rates (Experiment 1b). However, we showed that this problem could be ameliorated by making the comparison between hospitals difficult by presenting hospitals individually with their prediction intervals (Experiment 2). Thus, when the only comparison that is facilitated by the data presentation is the (appropriate) comparison between a hospital and its own prediction range, evaluations tend not to be so harsh unless this is warranted by *clear* evidence that the hospital is performing worse than predicted.

The impact of our combined-versus-individual manipulation (Experiment 2) illustrates an important application of evaluability theory (Hsee et al., 1999; Hsee & Zhang, 2010). Note, however, that it represents a non-typical application of this theory. In the examples from the literature that we discussed, which are typical of those presented in support of evaluability theory (e.g., see Hsee, 2000), joint evaluation makes it obvious which option is best on the most important attribute. Thus, joint evaluation makes it easier to evaluate and use the most important attribute. For the scenario that we explored, joint evaluation seems to make it easier to evaluate the raw survival rate. Importantly, however, this is *not* the most important attribute to evaluate (because it takes no account of case mix). This insight from Experiment 2 led us to adopt separate presentation of the prediction intervals and survival rate data for each hospital as the default view in our explanatory website of these data, and to augment this with an explanation of why comparing hospitals on their raw survival rates is *not* appropriate. We deemed this particularly important because we had decided to use percentage survival rate to report outcomes because several lines of evidence (e.g., Experiment 1a) indicated that people had considerable difficulties with understanding what the survival ratio represented and how it should be interpreted (even though that measure does take some account of case mix).

The bottom line is that people will often seek to get a grip on data by making relative judgments – even when an absolute judgment is called for (Parducci, 1968) – partly because they tend to seize upon whatever information they find easiest to use (Hsee & Zhang, 2010). Therefore an important part of helping people to use data well is to have regard for which comparisons are valid and which are spurious, and to consider whether the data can be presented in such a way that only appropriate comparisons are encouraged.

References

Broad, K., Leiserowitz, A., Weinkle, J., & Steketee, M. (2007). Misinterpretations of the "cone of uncertainty" in Florida during the 2004 hurricane season. *Bulletin of the American Meteorological Society, 88,* 651-667. doi:http://dx.doi.org/10.1175/BAMS-88-5-651

Brown, G. D. A., Gardner, J., Oswald, A. J., & Qian, J. (2008). Does wage rank affect employees' well-being? *Industrial Relations, 47,* 355-389.

Brown, K.L., Rogers, L., Barron, D.J., Tsang, V., Anderson, D., Tibby, S. … et al. (2017). Incorporating comorbidity within risk adjustment for UK pediatric cardiac surgery. *The Annals of Thoracic Surgery, 104(1),* 220–222.

Camerer, C. F. (2000). Prospect theory in the wild: Evidence from the field. In D. Kahneman & A. Tversky (Eds.) *Choices, values, and frames* (pp. 288-300). Cambridge: Cambridge University Press.

Epley, N., & Gilovich, T. (2001). Putting adjustment back in the anchoring and adjustment heuristic. *Psychological Science, 12,* 391-396.

Frederick, S. W., & Mochon, D. (2011). A scale distortion theory of anchoring. *Journal of Experimental Psychology: General, 141,* 124-133.

Galesic, M., & Garcia-Retamero, R. (2010). Graph literacy: A cross-cultural comparison. *Medical Decision Making, 31,* 444-457.

Hsee, C. K. (2000). Attribute evaluability: Its implications for joint-separate evaluation reversals and beyond. In D. Kahneman & A. Tversky (eds.) *Choices, values, and frames* (pp. 543-577). Cambridge: Cambridge University Press.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. (1999). Preference reversals between joint and separate evaluations of options: A review and theoretical analysis. *Psychological Bulletin, 125,* 576-590.

Hsee, C. K., & Zhang, J. (2010). General evaluability theory. *Perspectives on Psychological Science, 5,* 343-355.

Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decisions under risk. *Econometrica, 47*, 263-291.

Kahneman, D., Knetsch, J. L., Thaler, R. H. (1991). Anomalies: The endowment effect, loss aversion, and status quo bias. *The Journal of Economic Perspectives, 5*, 193-206.

Mussweiler, T., & Strack, F. (1999). Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology, 35,* 136-164.

Pagel, C., Crowe, S., Brown, K., & Utley, M. (2014). The benefits and risks of risk-adjustment in paediatric cardiac surgery. *Heart, 100,* 528-529. doi:10.1136/heartjnl-2013-304848

Pagel, C., Jesper, E., Thomas, J., Blackshaw, E., Rakow, T., Pearson, N., & Spiegelhalter, D. (2017). Understanding children's heart surgery data: A cross-disciplinary approach to codevelop a website. *The Annals of Thoracic Surgery, 104(1),* 342–352.

Pagel, C., Utley, M., Crowe, S., Witter, T., Anderson, D., Samson R., … et al. (2013). Real time monitoring of risk-adjusted paediatric cardiac surgery outcomes using variable life-adjusted display: Implementation in three UK centres. *Heart, 99,* 1445-1450. doi:10.1136/heartjnl-2013- 303671.

Parducci, A. (1968). The relativism of absolute judgments. *Scientific American, 219,* 84-90.

Rakow, T., Wright, R. J., Spiegelhalter, D. J., Bull, C. (2015). The pros and cons of funnel plots as a risk communication aid for individual decisions about medical treatment. *British Journal of Psychology, 106,* 327-348. DOI: 10.1111/bjop.12081

Rogers, L., Brown, K.L., Franklin, R.C., Ambler, G., Anderson, D., Barron, D.J., … et al. (2017). Improving risk adjustment for mortality after pediatric cardiac surgery: The UK PRAiS2 Model. *The Annals of Thoracic Surgery, 104(1),* 211-219.

Slovic, P., & Peters, E. (2006). Risk perception and affect. *Current Directions in Psychological Science, 15,* 322-325.

Spiegelhalter, D. J. (2005). Funnel plots for comparing institutional performance. *Statistics in Medicine, 24,* 1185-1202. DOI: 10.1002/sim.1970

Stewart, N. (2009). Decision by sampling: The role of the decision environment in risky choice. *The Quarterly Journal of Experimental Psychology, 62*, 1041-1062.

Zellner, D. A., Allen, D., Henley, M., & Parker, S. (2006). Hedonic contrast and condensation: Good stimuli make mediocre stimuli less good and less different. *Psychonomic Bulletin & Review, 13,* 235-239.

Table 1. Experiment 1a: Variations of the datasets used to create the graphical displays presented to participants for the evaluation questions. Datasets are a variation of a baseline dataset, whereby the 14 cardiac centres are approximately normally distributed within the survival-as-predicted range).[a]

Dataset	Dataset manipulations / distinguishing features of the dataset
Baseline	All survival rates within the 'survival-as-predicted' range, (approximately normal) distribution consistent with sampling variability being the only source of variation from the mean prediction.
1a	All points within the 'survival as predicted' range, but (only) one plot below the 1.00 survival ratio
1b	One plot below the 1.00 ratio but with 'survival as predicted' and all other points above the 1.00 ratio (ranging from 'survival as predicted' to 'survival much higher than predicted'.
2	One hospital is in the survival-higher-than-predicted region, and the others spread across the 'survival as predicted' range.
3	The hospitals show greater variation in their survival rates than in the other datasets, with one hospital in each of the four defined regions beyond the 'survival as predicted' range, and the remaining hospitals distributed across the 'survival as predicted' range.
4	Two hospitals with equal values on the survival rate scale, but one is within the survival-as-predicted interval and one is in the survival-lower-than-predicted region; all but one hospital is within the survival-as-predicted interval.

[a] Figures showing each data set can be seen in the *Supplementary Materials*.

Table 2. Experiment 1a: Frequency (percentage) correct for each comprehension question by condition.

	Question stem (1-6): What does the ___ (area) mean?	Correct answer	Condition		Combined
			Ratio-plot (N = 39)	Percentage-plot (N = 38)	
1	*'White/yellow'*	'Survival as predicted'	39 (100%)	38 (100%)	77 (100%)
2	*'Light blue/orange area to the right of the yellow bar'*	'Survival higher than predicted'	39 (100%)	35 (92.1%)	74 (96.1%)
3	*'Dark blue/white area to the left of the yellow bar'*	'Survival much higher than predicted'	39 (100%)	37 (97.4%)	76 (98.7%)
4	*'Light grey area /orange to the left of the yellow bar'*	'Survival lower than predicted'	38 (97.4%)	36 (94.7%)	74 (96.1%)
5	*'Dark grey area/white to the right of the orange bar'*	'Survival much lower than predicted'	39 (100%)	38 (100%)	77 (100%)
6	*'Black dot'*	Ratio-plot: 'actual versus predicted survival ratio' Percentage-plot: 'observed survival rate'	16 (41.0%)	27 (71.0%)	43 (55.8%)
7	*Which hospital has performed fewest surgical procedures?*	Hospital A	34 (87.1%)	28 (73.7%)	62 (80.5%)
8	*Which hospital has performed greatest surgical procedures?*	Hospital N	35 (89.7%)	28 (73.7%)	63 (81.8%)

Table 3. Experiment 1a: Frequency (percentage) of correct responses to interpretation questions by condition

Question (Which hospitals have a survival rate that is …)	Condition		
	Ratio-plot	Percentage-plot	Combined
1. As predicted	36 (92.3%)	36 (94.7%)	72 (93.5%)
2. Much higher than predicted	34 (87.2%)	35 (92.1%)	69 (89.6%)
3. Higher than predicted	37 (94.9%)	36 (94.7%)	73 (94.8%)
4. Lower than predicted	36 (92.3%)	36 (94.7%)	72 (93.5%)
5. Much lower than predicted	39 (100.0%)	35 (92.1%)	74 (96.1%)

Table 4. Questions for Experiment 1b; with frequency (percentage) of correct answers by condition for Questions 1-4

	Category	Question	Condition	
			Ratio-plot (N=24)	Percentage-plot (N=25)
1	Comprehension	Use an X to mark a dot for hospital D, which would indicate that the hospital's survival rate is as predicted.	24 (100%)	25 (100%)
2	Comprehension	Use an X to mark a dot for hospital F, which would indicate that the hospital's survival rate is much higher than predicted.	22 (91.7%)	25 (100%)
3	Inference	Use an X to mark a dot for hospital I, which would indicate that the hospital's survival rate would prompt NICOR [the audit body] to contact the hospital for more information about its results.	23 (95.8%)	24 (96.0%)
4	Inference	Use an X to mark a dot for hospital K, which would indicate that the hospital's survival rate would prompt NICOR [the audit body] to contact the hospital for more information about its results.	22 (91.7%)	22 (88.0%)
5	Threshold judgment	Use an X to mark a dot for hospital M, which would indicate that the hospital's survival rate is the lowest survival rate that you would personally consider to be acceptable.	--	--
6	Qualitative	Please explain your answer [to question 5].	--	--

Table 5. Themes identified from participants' free-text explanations for their threshold values for the "minimum acceptable survival rate" (Experiment 1b)

Theme[a]	Example of theme present	Frequency (%) present	Theme associated with
1. Poor quality of care by hospital	*'The fact that the survival rate is outside the orange box would suggest that these might be problems with the surgical team.'*	9 (18.8%)	--
2. Personal significance of type of surgery (paediatric)	*'I think any hospital survival rate that is in any way below what is predicted is unacceptable. This is due to personal repercussions for families involved.'*	5 (10.4%)	--
3. Specific numerical value	*'When the survival rate is above the 1/3 of the predicted range, it is okay for me to accept it.'*	13 (27.1%)	--
4. Takes into account total number of procedures	*'A survival ratio below 0.97 for me is unacceptable, especially in hospital M where there 1481 surgical episodes which is a lot.'*	5 (10.4%)	Threshold level
5. Comparison to data of other hospitals	*'Anything lower than the lowest range out of all the hospitals (hospital B) would to me be considered unacceptable.'*	15 (31.1%)	Threshold level & Condition
6. Discusses role of chance	*'Although being outside of the predicted range could be a result of chance; it seems safer to trust the hospitals who have matched their predicted survival rate.'*	10 (20.8%)	--
7. Would like more information	*'Personally I would want to see survival rates no lower than anticipated, unless presented with certain facts about why this has not been the case.'*	1 (2.1%)	--
8. Considers type or complexity of surgery (paediatric)	*'A 96% survival rate is still a relatively high percentage, especially for something as complex as congenital heart disease.'*	1 (2.1%)	--
9. Discusses the prediction process or model	*'I would hope the hospital to at least be able to match the result that they have predicted or else the whole prediction process would prove meaningless.'*	3 (6.3%)	--
10. Emphasises importance of being within the 'as predicted' range	*'It is still within the predicted survival rate but its observed rate is at the lowest.'*	22 (45.8%)	Threshold level & Condition
11. Refers to any aspect of prediction or expectation shown	*'This is because, while it is out of the predicted range for the hospital itself, it is within the ranges for the other hospitals.'*	40 (83.3%)	--

[a] Initially, ten themes (Themes 1-10) were identified and coded; and coding was then checked (blind) by a second independent coder. Agreement was between 94% and 100% for all themes except Theme 5 (88%) and Theme 10 (52%). Therefore, discrepancies were resolved for Theme 5, by agreement between the two coders. The disagreement on Theme 10 led to the creation of Theme 11, so the distinction between 10 and 11 is that: for Theme 10, participants must refer to the importance of being within the "survival as predicted" range; whereas for Theme 11, participants refer to *any* aspect of prediction or expectation.

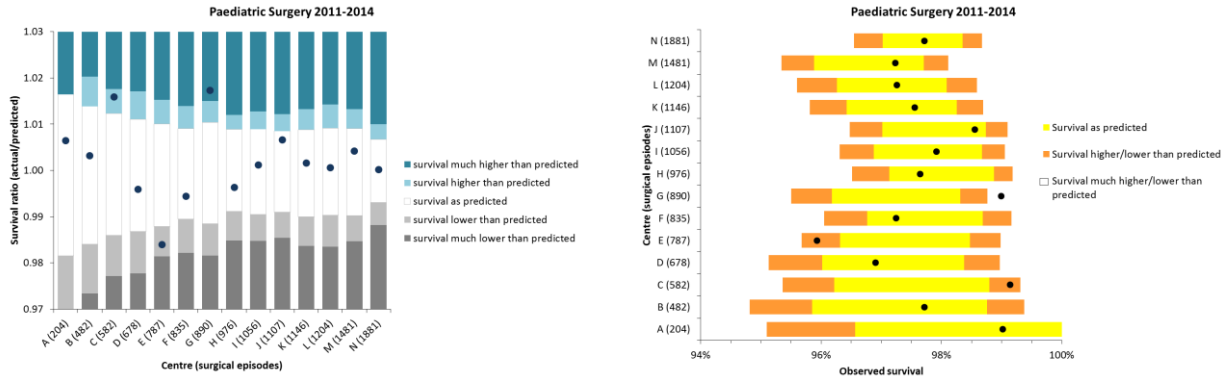

Figure 1. Stimuli for comprehension and interpretation questions in Experiment 1a: the ratio-scale plot (left) is the plot used in the annual audit report of children's heart surgery outcomes; the percentage-sale plot (right) was newly devised for an explanatory website for these data. Both plots display 30-survival rates in relation to predicted survival rates that adjust for case mix, and show 95% and 99.8% prediction intervals (indicated by differential shading).

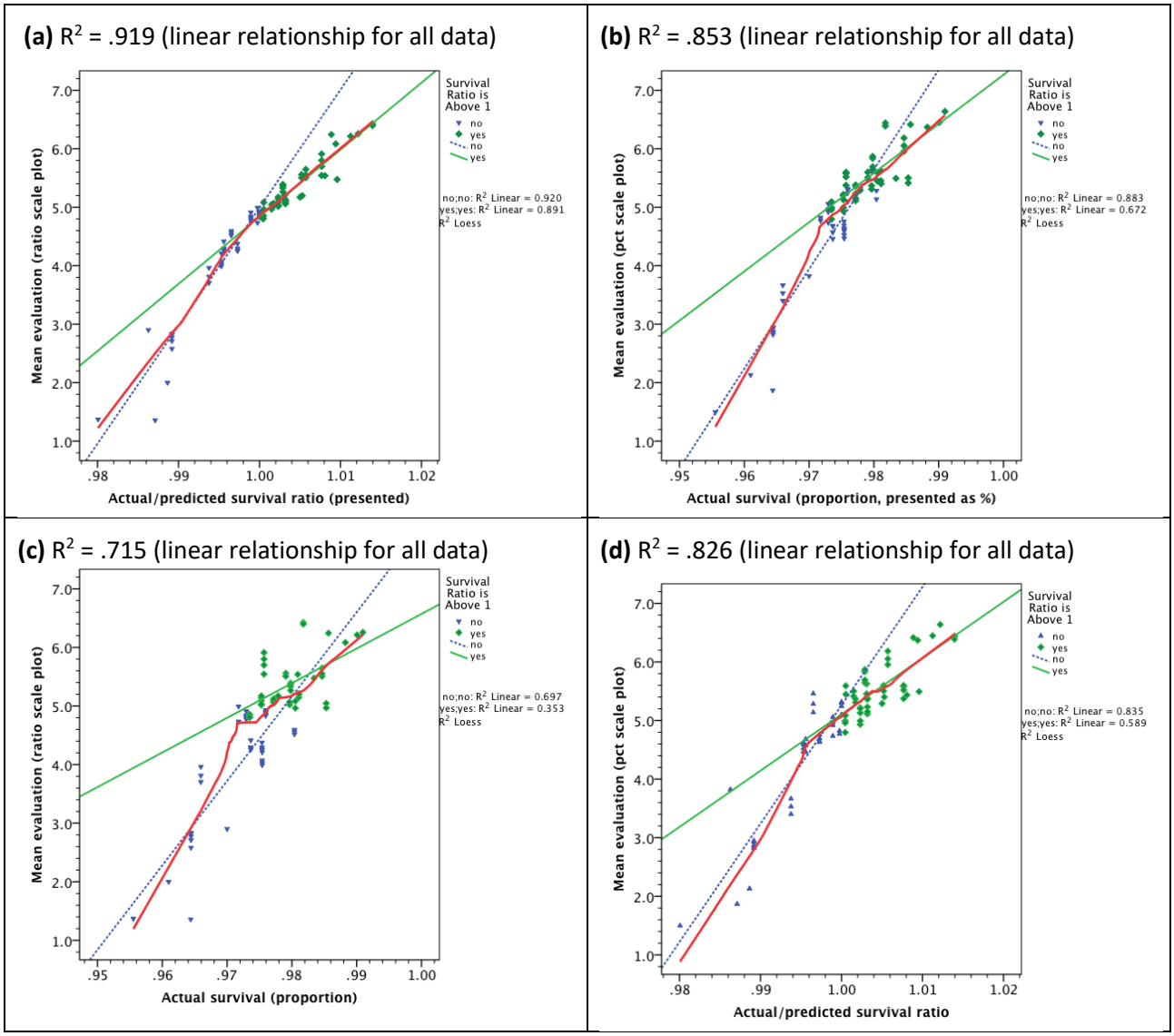

Figure 2. Experiment 1a: Mean evaluation by participants per hospital (vertical axis) plotted as a function of actual outcome in that hospital (survival ratio, a & d; or survival proportion, b & c) for the ratio plot (a & c) and percentage plot (b & d). The non-linear nature of each relationship is illustrated by a Loess regression line (solid red) and separate linear regressions for hospitals with survival ratios above 1 (solid green line) and those with survival ratios below 1 (dotted blue line).

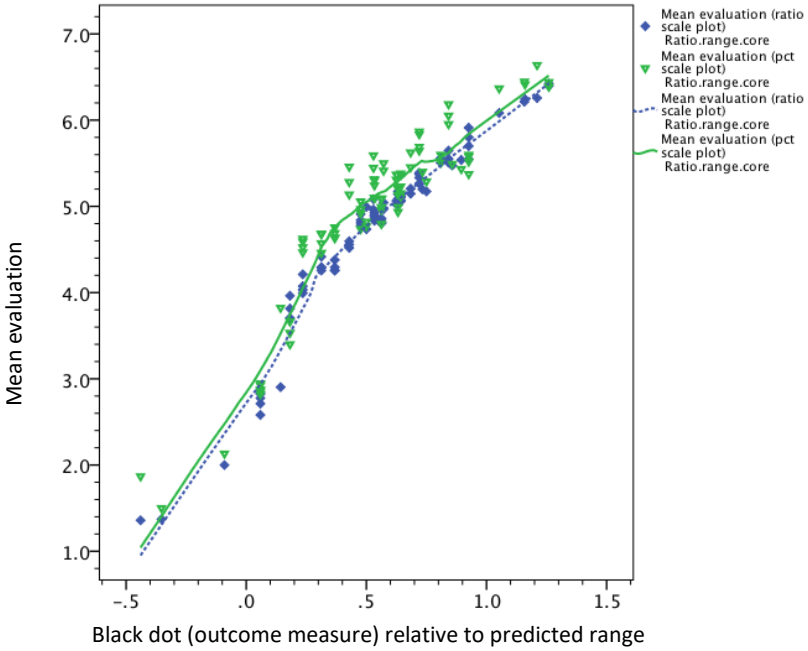

Figure 3. Experiment 1a: Mean evaluation by participants per hospital (vertical axis) plotted as a function of the position of actual outcomes relative to the prediction interval [prediction interval scaled from 0 to 1; hence x < 0.0 denotes "survival worse than predicted" and x > 1.0 denotes "survival better than predicted"]. Non-linear Loess regression lines are plotted separately for each condition (ratio plot = diamond icons; percentage plot = open triangle icons).

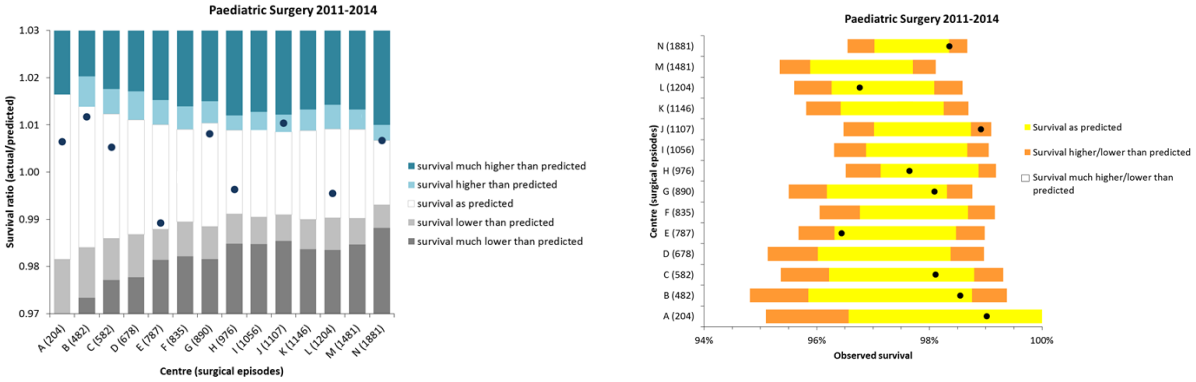

Figure 4. Stimuli for Experiment 1b: the ratio-scale plot (left) or percentage-sale plot (right) was shown to participants. Both plots display 30-survival rates in relation to predicted survival rates that adjust for case mix (though missing for some hospitals), and show 95% and 99.8% prediction intervals (indicated by differential shading).

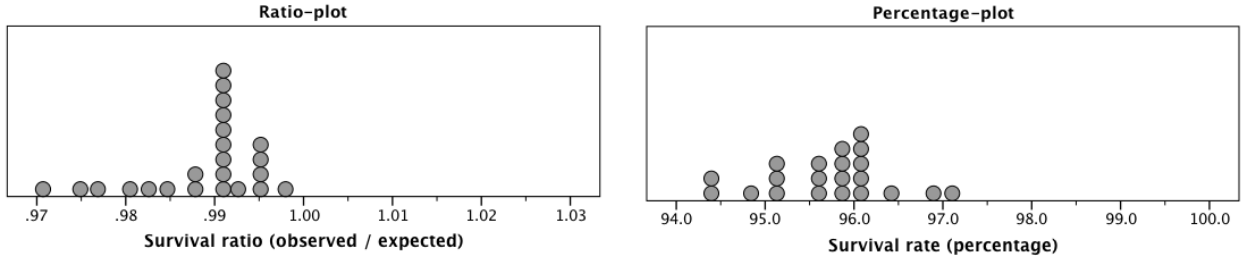

Figure 5. Experiment 1b: Distribution of minimum acceptable survival rates stated by participants: ratio-plot condition (left) and percentage-plot condition (right).

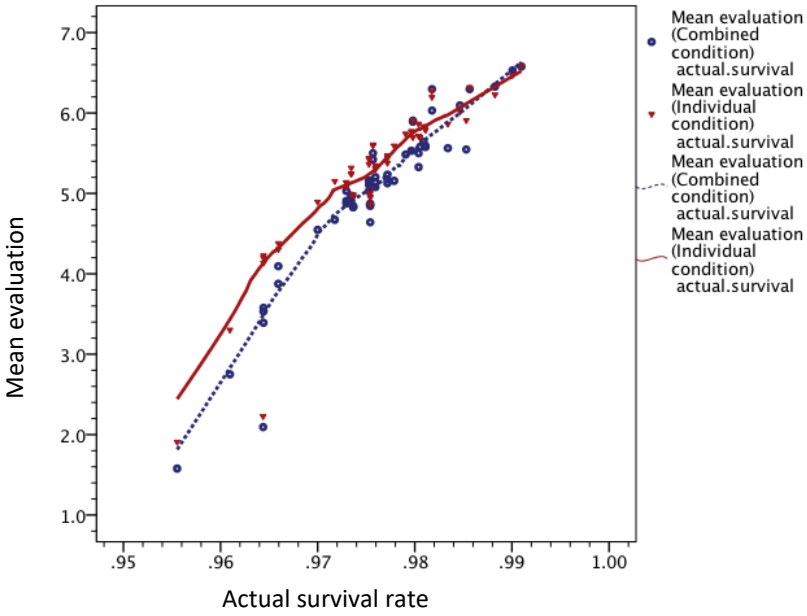

Figure 6. Mean evaluation by participants per hospital (vertical axis) as a function of the actual survival rate for those hospitals, plotted separately according to whether those evaluations were made with data for all hospitals showing (combined condition, blue circle icons) or only for the hospital being evaluated (individual condition, red triangle icons). Non-linear Loess regression lines are plotted separately for each condition (broken blue line = combined condition; solid red = individual condition).