Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. doi: 10.18637/jss.v000.i00

aggreCAT: an R Package for Mathematically Aggregating Expert Judgements

Elliot Gould*
University of Melbourne

Charles T. Gray
Newcastle University

Aaron Willcox
University of Melbourne

Rose O'Dea
University of New South Wales

Rebecca Groenewegen
University of Melbourne

David P. Wilkinson
University of Melbourne

Abstract

Structured elicitation protocols, such as the IDEA protocol, may be used to elicit expert judgements in the form of subjective probabilities from multiple experts. Judgements from individual experts about a particular phenomena must therefore be mathematically aggregated into a single prediction. The process of aggregation may be complicated when judgements are elicited with uncertainty bounds, and also when there are several rounds of elicitation. This paper presents the new R package **aggreCAT**, which provides 27 unique aggregation methods for combining individual judgements into a single, probabilistic measure. The aggregation methods were developed as a part of the Defense Advanced Research Projects Agency (DARPA) 'Systematizing Confidence in Open Research and Evidence' (SCORE) programme, which aims to generate confidence scores or estimates of 'claim credibility' for over 4000 research claims from the social and behavioural sciences. We provide several worked examples illustrating the underlying mechanics of the aggregation methods. We also describe a general workflow for using the software in practice to facilitate uptake of this software for appropriate use-cases.

Keywords: mathematical aggregation, expert judgement, DARPA SCORE, replicability, R.

1. Introduction

Expert judgement is frequently used to inform forecasting about uncertain future events across a range of disciplines, including ecology, conservation science, human geography, political science, and management (Sutherland, Dicks, Everard, and Geneletti 2018). Judgements from groups of experts tend to perform better than a single expert (Goossens, Cooke, Hale, and Rodic-Wiersma 2008), and it is best-practice to elicit judgements from diverse groups so that group members can bring "different perspectives, cross-examine each others' reasoning, and share information", however judgements or forecasts must then be distilled into a single forecast, ideally accompanied by estimates of uncertainty around those estimates (Hanea, Wilkinson, McBride, Lyon, van Ravenzwaaij, Singleton Thorn, Gray, Mandel, Willcox, Gould, and et al. 2021). Judgements from multiple experts may be combined into a single forecast using either behavioural approaches that force experts into forming consensus, or by using mathematical approaches (Goossens *et al.* 2008).

Although there are a variety of methods for mathematically aggregating expert judgements into single point-predictions, there are few open-source software implementations available to analysts or researchers. The R (R Core Team 2017) package **expert** provides three models of expert opinion to combine judgements elicited from groups of experts (CITE) , and **SHELF** implements only a single method (weighted linear pool) for aggregating expert judgements (CITE). Other R packages providing methods to mathematically aggregate expert judgements do so for non-point predictions, for example, **opera**, which generates time-series predictions (CITE). In this paper we present the **aggreCAT** package, which provides 27 different methods for mathematically aggregating judgements within groups of experts into a single forecast.

1.1. DARPA SCORE program and the repliCATS project

The **aggreCAT** package, and the mathematical aggregators therein, were developed by the repliCATS (Collaborative Assessment for Trustworthy Science) project as a part of the SCORE program (Systematizing Confidence in Open Research and Evidence), funded by DARPA (Defense Advanced Research Projects Agency) (Alipourfard, Arendt, Benjamin, Benkler, Bishop, Burstein, Bush, Caverlee, Chen, Clark, Dreber, Errington, Fidler, Fox, Frank, Fraser, Friedman, Gelman, Gentile, Gordon, Griffin, Gulden, Hahn, Hartman, Holzmeister, Hu, Johannesson, Kezar, Kline Struhl, Kuter, Kwasnica, Lee, Lerman, Liu, Loomas, Luis, Magnusson, Bishop, Miske, Mody, Morstatter, Nosek, Parsons, Pennock, Pi, Pujara, Rajtmajer, Ren, Salinas, Selvam, Shipman, Silverstein, Sprenger, Squicciarini, Stratman, Sun, Tikoo, Twardy, Tyner, Viganola, Wang, Wilkinson, and Wintle 2021). The SCORE program is the largest replication project in science to date, and aims to build automated tools that can rapidly and reliably assign "Confidence Scores" to research claims from empirical studies in the Social and Behavioural Sciences (SBS). Confidence Scores are quantitative measures of the likely reproducibility or replicability of a research claim or result, and may be used by consumers of scientific research as a proxy measure for their credibility in the absence of replication effort (Alipourfard *et al.* 2021).

Replications are time-consuming and costly (Isager, van Aert, Bahnik, Brandt, Desoto, Ginner-Sorolla, Krueger, Perugini, Ropovik, van't Veer, Vranka, and Lakens 2020), and studies have shown that replication outcomes can be reliably elicited from researchers (Gordon, Viganola, Bishop, Chen, Dreber, Goldfedder, Holzmeister, Johannesson, Liu, Twardy, Wang, and Pfeiffer 2020). Consequently, the DARPA SCORE program generated

Confidence Scores for > 4000 SBS claims using expert elicitation based on two very different strategies – prediction markets (Gordon *et al.* 2020) and the IDEA protocol (Hemming, Burgman, Hanea, McBride, and Wintle 2017), the latter of which is used by the repliCATS project (Fraser, Bush, Wintle, Mody, Smith, Hanea, Gould, Hemming, Hamilton, Rumpff, Wilkinson, Pearson, Singleton Thorn, Ashton, Willcox, Gray, Head, Ross, Groenewegen, Marcoci, Vercammen, Parker, Hoekstra, Nakagawa, Mandel, van Ravenzwaaij, McBride, Sinnot, Vesk, Burgman, and Fidler 2021). A proportion of these research claims were randomly selected for direct replication, against which the elicited and aggregated Confidence Scores are 'ground-truthed'. These aim of the DARPA SCORE project is to aid the development of artificial intelligence tools that can automatically assign Confidence Scores.

The repliCATS IDEA protocol

The repliCATS project adapted and deployed the IDEA protocol to elicit crowd-sourced judgements from diverse groups about the likely replicability of SBS research claims (Fraser *et al.* 2021). The IDEA ('Investigate', 'Discuss', 'Estimate' and 'Aggregate') protocol is a four-step structured elicitation protocol that draws on the 'wisdom of crowds' to elicit subjective judgements about the likelihood of uncertain events (Hemming *et al.* 2017, figure 1). To collect expert judgements about the replicability of SBS claims, we asked participants to estimate the "probability that direct replications of a study would find a statistically significant effect in the same direction as the original claim", eliciting estimates of uncertainty in the form of upper and lower bounds on those point-estimates. Judgements were elicited using the repliCATS platform (Pearson, Fraser, Bush, Mody, Widjaja, Head, Wilkinson, Sinnott, Wintle, Burgman, Fidler, and Vesk 2021), a multi-user cloud-based software platform that implements the IDEA protocol, between July 7th 2019 and November 30th 2020.

For a single claim under assessment, between 4 and 15 experts individually drew on background information to provide estimates of the probability, including 4 numeric data points and one character data point: an upper and lower bound, and best estimate of the event probability, as well as justifications for their estimates, and a value on the likert binary scale up to 7 rating the individuals' degree of comprehension of the claim (Round 1, *Investigate*). In the *Discuss* phase, three-point estimates from each group member are anonymously presented to the group, who then collectively discuss differences in opinion and provide potential evidence for these differences. Group members subsequently provide a second set of probabilistic judgements (Round 2, *Estimate*). Thus, for a single assessment, 2 sets of judgements are elicited from each expert (*pre-* and *post*-group discussion).

During the fourth step, *Aggregate*, judgements are mathematically aggregated into a single Confidence Score or forecast of replicability. The repliCATS project developed 27 different methods for mathematically aggregating judgements elicited from groups of experts into Confidence Scores (Hanea *et al.* 2021). We developed the **aggreCAT** package to implement these aggregation methods and deliver Confidence Scores for over 4000 SBS research claims as a part of the DARPA SCORE project.

2. Introducing the aggreCAT package

In this paper we aim to provide a detailed overview of the **aggreCAT** package so that re-

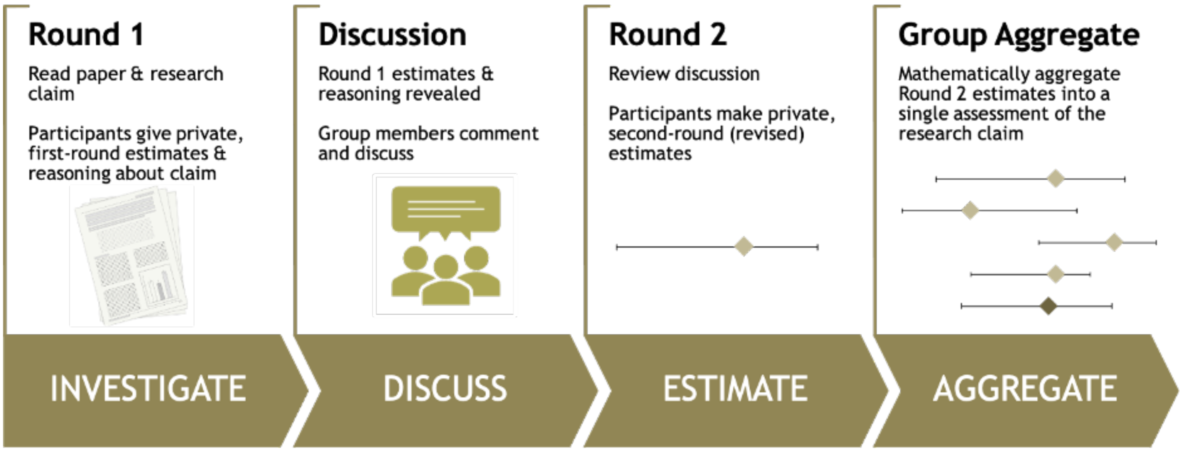

Figure 1: The IDEA protocol as deployed by the repliCATS project (reproduced with permission from Wintle et al. 2021).

searchers may apply the aggregation functions described in (Hanea *et al.* 2021) to their own expert elicitation datasets where mathematical aggregation is required. Note that judgements that have already been subjected to behavioural or consensus aggregation may not be subsequently mathematically aggregated, however individual elicited judgements may be aggregated mathematically as an alternative or complement to behavioural or consensus-based aggregation.

We begin by formulating the problem of mathematically aggregating expert judgements. Each method, and its data requirements is summarised in Table 5. Before outlining key aggregation methods, we briefly summarise package datasets, which were collected by the repliCATS project. By first describing the datasets before describing the aggregation methods in detail, we aim to provide a grounded understanding of the different outputs of expert elicitation using the repliCATS IDEA protocol, and the inputs available to the aggregation functions.

Next, we describe and illustrate the main types of aggregators, which may be categorised according to their data requirements, mathematical properties and computational implementation (Section 4). By selecting representative functions of each key aggregator type and applying them to a subset of focal claims, we demonstrate the internal mechanics of how these methods differently operationalise the data to generate forecasts or Confidence Scores. We do not give advice on the circumstances in which each method should be used, instead, choice of aggregation method should be informed by the mathematical properties of the method, the desired properties of an aggregation, and the purpose for which the aggregation is being used. For a detailed description of each method as well as a discussion of their relative merits, see (Hanea *et al.* 2021).

Finally, we provide a detailed workflow for aggregating expert judgments for multiple forecasts, using multiple aggregation functions, as implemented by the repliCATS project in the course of delivering $>$ 4000 Confidence Scores for the DARPA SCORE program. The **aggreCAT** package provides a set of supporting functions for evaluating or ground-truthing aggregated forecasts or Confidence Scores against a set of known-outcomes, as well as functions for visualising comparisons of different aggregation methods and the outcomes of performance evaluation. We describe and demonstrate this functionality in the presentation of the repliCATS workflow. The workflow is representative of the probable challenges faced by the researcher in the course of mathematically aggregating groups of forecasts, and should equip the reader to use **aggreCAT** for their own datasets; it exemplifies how to extend the **aggreCAT** package to any expert judgement dataset from any domain in which there are multiple judgements from multiple individuals that must be combined into a single forecast.

3. Mathematically Aggregating Expert Judgements

Mathematically, the aggregation methods can be divided into three main types:

- Un-weighted linear combination of best estimates, transformed best estimates or distributions,
- Weighted linear combinations of best estimates, transformed best estimates and of distributions, where weights are proxies of forecasting performance constructed from characteristics of participants and/or their judgements, and
- Bayesian methods that use participant judgements as data with which to update both uninformative and informative priors.

However, the **aggreCAT** package user might wish to categorise the aggregation methods according to aspects of their computational implementation and data requirements, because these inform the function arguments as well as the type and form of the data that is parsed to the aggregation functions. These aspects include:

- Elicitation requirement, number of elicitation rounds: the majority of aggregation methods require data from only a single round of judgements, i.e. the final post-discussion estimates. However, some aggregation methods require data from both rounds of judgements, which may be elicited using the IDEA protocol or other similar structured elicitation protocol in which there are two rounds of judgements.
- Elicitation requirement, single point or three point elicitation: several aggregation methods use only a single data point elicited from individuals (their best estimate), however, most aggregation methods require a best estimate, and estimates of uncertainty in the form of upper and lower bounds.
- Number of claims / forecasts assessed by the individual: some weighted aggregation methods consist of weights that are calculated from properties of participant judgements across multiple forecasting questions, not just the target claim being aggregation. Secondly, for aggregation methods that calculate variance in estimates, variance cannot be calculated on a single data point. While 2 is the mathematical minimum, the user should give consideration to what minimum number of claims should be used to reliably calculate measures of variance.
- Supplementary data requirements: several aggregation methods require supplementary data collected either in addition to or as part of the repliCATS IDEA protocol, some of which will need additional qualitative coding before being parsed to the aggregation function.

The data and structured elicitation protocol requirements are described in Table 5. All aggregation methods requiring a single round of estimates can therefore be applied to expert judgments derived from any structured elicitation protocol that generates, lower, upper, and best estimates from each individual (i.e. not just the IDEA protocol), and does not enforce behavioural consensus.

Notation and Problem Formulation

Here we describe some preliminary mathematical notation used to represent each aggregation method. For the mathematical specification of each individual aggregation function, please consult (Hanea *et al.* 2021) or the **aggreCAT** package function documentation.

The total number of research claims, *claim*, or unique forecasts being assessed, C , is indexed by $c = 1, ..., C$. The total number of individuals / experts / participants is denoted by N, and is indexed by $i = 1, ..., N$. Each claim *outcome* (i.e. the outcome of a replication study) assumes binary values, where the value is 0 if the claim is false, and 1 if the claim is true. '`TRUE`' claims are claims where the replication study found a significant result in the same direction as the original research claim, and '`FALSE`' claims are those where the replication study *did not* find a significant result in the same direction as the original study. For each claim c, an individual i assesses the probability of a claim replicating by providing three probabilities: a lower bound $L_{i,c}$, an upper bound $U_{i,c}$, and a best estimate $B_{i,c}$, satisfying the inequalities: $0 \leq Li, c \leq Bi, c \leq Ui, c \leq 1$.

Every claim is assessed by multiple individuals, and their probabilities are aggregated using one of the aggregation methods to obtain a group or aggregate probability, denoted by $\hat{p}_c$. The aggregated probability calculated using a specific method, is given by $\hat{p}_c\,(MethodID)$. Each aggregation is assigned a unique $MethodID$ which is the abbreviation of the mathematical operation used in calculating the weights. Note that all Best, Lower and Upper estimates are taken to be `round 2` judgements from the repliCATS IDEA protocol Figure 1), unless appended by a "1", where they are `round 1` judgements, e.g. $B1_{i,c}$ denotes the `round 1` Best estimate from individual i for claim c.

Weighting Expert Forecasting Performance

Equal-weighting of judgements are less calibrated, accurate and informative than weighted aggregation methods where judgements from experts who performed well in similar judgement tasks are more heavily weighted (Hanea *et al.* 2021). Proxies for forecasting performance, such as breadth and variability of qualitative reasons used by experts to justify their judgements, can be used to form weights in the absence of measures of experts' prior performance (Hanea *et al.* 2021).

The aggregation methods other than the `AverageWAgg()` and Bayesian approaches in **aggreCAT** each employ weighting schemes that are informed by proxies for good forecasting performance whereby experts' estimates are weighted differently by measures of reasoning, engagement, openness to changing their mind in light of new facts, evidence or opinions presented in the discussion round, extremity of estimates, informativeness of estimates, asymmetry of estimate bounds, granularity of estimates, and by prior statistical knowledge as measured in a quiz.

Below, we define standardised notation for describing weighted linear combinations of individual judgements where un-normalised weights are denoted by w_method and normalised weights by $\tilde{w}\_method$ (Equation 1). Given that for all aggregation methods weights are normalised, and that the normalisation process is the same for each aggregation method, the equations for the aggregation methods are presented for un-normalised weights.

$$\hat{p}_c\left(MethodID\right) = \frac{1}{N}\sum_{i=1}^{N} \tilde{w}\_method_{i,c} B_{i,c} \tag{1}$$

By default, weights are calculated across all claims on a per-individual, per-claim basis, such that judgements for the same individual are weighted differently across all claims they have provided judgements for. There are some exceptions to this default: `GranWAgg`, `QuizWAgg`, `IndIntWAgg` `IndIntAsymWAgg`, `VarIndIntWAgg`, `KitchSinkWAgg`. Note that `IndIntWAgg,` and methods that include its weighting function `weight_nIndivInterval()` as a component, re-scale weights by a fixed measure across all claims. Hence, for aggregation methods that use information from multiple claims other than the target claim for which the Confidence Score is being computed, each individual claim c is indexed by $d = 1, ..., C$. Where the default weighting is used, this is coded into each function. However, where more complex and method-specific weighting methods are used, modularised functions have been created for ease of debugging. These function names are prefixed with `weight_`.

3.1. Package datasets

The **aggreCAT** package includes the core dataset `data_ratings` consisting of judgements elicited during a pilot experiment exploring the performance of IDEA groups in assessing replicability of a set of claims with "known outcomes." "Known-outcome" claims are SBS research claims that have been subject to replication studies in previous large-scale replication projects[1]. Data were collected using the repliCATS IDEA protocol at a two day workshop[2] in the Netherlands, on July 2019, at which 25 participants assessed the replicability of 25 unique SBS claims. In addition to the probabilistic estimates provided for each research claim assessed, participants were also asked to rate the claim's plausibility and comprehensibility, answer whether they were involved in any aspect of the original study, and to provide their reasoning in support of their quantitative estimates, which were used to form measures of reasoning breadth and engagement (Fraser *et al.* 2021).

`data_ratings` is a *tidy* '`data.frame`' wherein each *observation* (or row) corresponds to a single value in the set of `values` constituting a participant's complete assessment of a research claim. Each research claim is assigned a unique `paper_id`, and each participant has a unique (and anonymous) `user_name`. The variable `round` denotes the round in which each `value` was elicited (`round_1` or `round_2`). `question` denotes the type of question the `value` pertains to; `direct_replication` for probabilistic judgements about the replicability of the claim, `belief_binary` for participants' belief in the plausibility of the claim, `comprehension` for participants' comprehensibility ratings, and `involved_binary` for involvement in the original study. An additional column `element` maintains the tidy structure of the data, while capturing the multiple `values` that comprise a full assessment of the replicability (`direct_replication`) of a claim; `three_point_best`, `three_point_lower` and `three_point_upper` denote the best estimate and lower and upper bounds respectively. `binary_question` describes the `element` for both the plausibility rating (`belief_binary`) and in-

[1]Many labs 1, 2 and 3 Klein, Ratliff, Vianello, Adams Jr., Bahnic, Bernstein, Bocian, Brandt, Brooks, Brumbaugh, Cermalcilar, Chandler, Cheong, Davis, Devos, Eisner, Frankowska, Furrow, Galiani, Hasselman, Hicks, Hovermale, Hunt, Huntsinger, IJzerman, John, Joy-Gaba, Barry Kappes, Kreuger, Kurtz, Levitan, Mallet, Morris, Nelson, Nier, Packard, Pilati, Rutchick, Schmidt, Skorinko, Smith, Steiner, Storbeck, Van Swol, Thompson, van 't Veer, Vaughn, Vranka, Wichman, Woodzicka, and Nosek (2014), Klein, Vianello, Hasselman, Adams, Reginald B. Adams, Alper, Aveyard, Axt, Babalola, Štěpán Bahník, Batra, Berkics, Bernstein, Berry, Bialobrzeska, Binan, Bocian, Brandt, Busching, Rédei, Cai, Cambier, Cantarero, Carmichael, Ceric, Chandler, Chang, Chatard, Chen, Cheong, Cicero, Coen, Coleman, Collisson, Conway, Corker, Curran, Cushman, Dagona, Dalgar, Rosa, Davis, de Bruijn, Schutter, Devos, de Vries, Doğulu, Dozo, Dukes, Dunham, Durrheim, Ebersole, Edlund, Eller, English, Finck, Frankowska, Ángel Freyre, Friedman, Galliani, Gandi, Ghoshal, Giessner, Gill, Gnambs, Ángel Gómez, González, Graham, Grahe, Grahek, Green, Hai, Haigh, Haines, Hall, Heffernan, Hicks, Houdek, Huntsinger, Huynh, IJzerman, Inbar, Åse H. Innes-Ker, Jiménez-Leal, John, Joy-Gaba, Kamiloğlu, Kappes, Karabati, Karick, Keller, Kende, Kervyn, Knežević, Kovacs, Krueger, Kurapov, Kurtz, Lakens, Lazarević, Levitan, Neil A. Lewis, Lins, Lipsey, Losee, Maassen, Maitner, Malingumu, Mallett, Marotta, Međedović, Mena-Pacheco, Milfont, Morris, Murphy, Myachykov, Neave, Neijenhuijs, Nelson, Neto, Nichols, Ocampo, O'Donnell, Oikawa, Oikawa, Ong, Orosz, Osowiecka, Packard, Pérez-Sánchez, Petrović, Pilati, Pinter, Podesta, Pogge, Pollmann, Rutchick, Saavedra, Saeri, Salomon, Schmidt, Schönbrodt, Sekerdej, Sirlopú, Skorinko, Smith, Smith-Castro, Smolders, Sobkow, Sowden, Spachtholz, Srivastava, Steiner, Stouten, Street, Sundfelt, Szeto, Szumowska, Tang, Tanzer, Tear, Theriault, Thomae, Torres, Traczyk, Tybur, Ujhelyi, van Aert, van Assen, van der Hulst, van Lange, van 't Veer, Vásquez-Echeverría, Vaughn, Vázquez, Vega, Verniers, Verschoor, Voermans, Vranka, Welch, Wichman, Williams, Wood, Woodzicka, Wronska, Young, Zelenski, Zhijia, and Nosek (2018), Ebersole, Atherton, Belanger, Skulborstad, Allen, Banks, Baranski, Bernstein, Bonfiglio, Boucher, Brown, Budiman, Cairo, Capaldi, Chartier, Chung, Cicero, Coleman, Conway, Davis, Devos, Fletcher, German, Grahe, Hermann, Hicks, Honeycutt, Humphrey, Janus, Johnson, Joy-Gaba, Juzeler, Keres, Kinney, Kirshenbaum, Klein, Lucas, Lustgraaf, Martin, Menon, Metzger, Moloney, Morse, Prislin, Razza, Re, Rule, Sacco, Sauerberger, Shrider, Shultz, Siemsen, Sobocko, Weylin Sternglanz, Summerville, Tskhay, van Allen, Vaughn, Walker, Weinberg, Wilson, Wirth, Wortman, and Nosek (2016), the Social Sciences Replication Project Camerer, Dreber, Holzmeister, Ho, Huber, Johannesson, Kirchler, Nave, Nosek, Pfeiffer, Altmejd, Buttrick, Chan, Chen, Forsell, Gampa, Heikensten, Hummer, Taisuke, Isaksson, Manfredi, Rose, Wagenmakers, and Wu (2018) and the Reproducibility Project Psychology aac (2015).

[2]See Hanea *et al.* (2021) for details. The workshop was held at the annual meeting of the Society for the Improvement of Psychological Science (SIPS), <https://osf.io/ndzpt/>.

volvement (involved_binary) questions, whereas likert_binary is the element describing a participant's comprehension rating. Judgements are recorded in column value in the form of percentage probabilities ranging from (0,100). The binary_questions corresponding to comprehensibility and involvement consist of binary values (1 for the affirmative, and -1 for the negative). Finally, values corresponding to participants' comprehension ratings are on a likert_binary scale from 1 through 7. Note that additional columns with participant attributes can be included in the ratings dataset if required by the user, we include the group column in data-ratings, which describes the group number the participant was a part of. Below we show some example data for a single user for a single claim to illustrate this structure of the core data_ratings dataset.

```
R> library(tidyverse,quietly = TRUE)
R> library(aggreCAT)
R> aggreCAT::data_ratings %>%
+  dplyr::filter(paper_id == dplyr::first(paper_id),
+                user_name == dplyr::first(user_name)) %>%
+  print(., n = nrow(.))
```

```
# A tibble: 11 x 7
   round   paper_id user_name  question            element           value group
   <chr>   <chr>    <chr>      <chr>               <chr>             <dbl> <chr>
 1 round_1 100      7l8m7dmjdb direct_replication  three_point_lower    30 UOM1
 2 round_1 100      7l8m7dmjdb involved_binary     binary_question      -1 UOM1
 3 round_1 100      7l8m7dmjdb belief_binary       binary_question      -1 UOM1
 4 round_1 100      7l8m7dmjdb direct_replication  three_point_best     40 UOM1
 5 round_1 100      7l8m7dmjdb direct_replication  three_point_upper    45 UOM1
 6 round_1 100      7l8m7dmjdb comprehension       likert_binary         5 UOM1
 7 round_2 100      7l8m7dmjdb comprehension       likert_binary         4 UOM1
 8 round_2 100      7l8m7dmjdb belief_binary       binary_question      -1 UOM1
 9 round_2 100      7l8m7dmjdb direct_replication  three_point_lower    30 UOM1
10 round_2 100      7l8m7dmjdb direct_replication  three_point_upper    45 UOM1
11 round_2 100      7l8m7dmjdb direct_replication  three_point_best     39 UOM1
```

Not all data necessary for constructing weights on performance is contained in data_ratings. Additional data collected as part of the repliCATS IDEA protocol are contained within separate datasets to data_ratings. Participants provided justifications for giving particular judgemetns, and these are contained in data_justifications. On the repliCATS platform users were given the option to comment on others' justifications (data_comments), to vote on others' comments (data_comment_ratings) and on others' justifications (data_justification_ratings). Finally, **aggreCAT** contains three 'supplementary' datasets containing data collected externally to the repliCATS IDEA protocol: data_supp_quiz, data_supp_priors, and data_supp_reasons.

Quiz Score Data

Prior to the workshop, participants also completed an optional quiz on statistical concepts and meta-research that we expect participants to be aware of in order to reliably evaluate the replicability of research claims. Quiz responses are contained in data_supp_quiz and are used to construct performance weights for the aggregation method `QuizWAgg` where each participant receives a `quiz_score` if they completed the quiz, and `NA` if they did not attempt the quiz (see Hanea *et al.* 2021, for further details).

Reasoning Data

`ReasonWAgg` uses the number of unique reasons given by participants to support a Best Estimate for a given claim $B_{i,c}$ to construct performance weights, and is contained within `data_supp_reasons`. Qualitative statements made by individuals during claim evaluation were recorded on the repliCATS platform (Pearson *et al.* 2021) and coded as falling into one of 25 unique reasoning categories by the repliCATS Reasoning team (Wintle, Mody, Smith, Hanea, Wilkinson, Hemming, Bush, Fraser, Singleton Thorn, McBride, Gould, Head, Hamilton, Rumpff, Hoekstra, and Fidler 2021). Reasoning categories include plausibility of the claim, effect size, sample size, presence of a power analysis, transparency of reporting, and journal reporting (Hanea *et al.* 2021). Within `data_supp_reasons`, each of the reasoning categories that passed our inter-coder reliability threshold are distributed as columns in the dataset whose names are prefixed with `RW`, and for each claim `paper_id`, each participant `user_id` is assigned a logical `1` or `0` if they included that reasoning category in support of their Best estimate for that claim. See Section 4.4 for details on the `ReasonWAgg` aggregation method.

Bayesian Prior Data

The method `BayPRIORsAgg` uses Bayesian updating to update a prior probability of a claim replicating estimated from a predictive model (Gould, Willcox, Fraser, Singleton Thorn, and Wilkinson 2021) using an aggregate of the best estimates for all participants assessing a given claim c (Hanea *et al.* 2021). The prior data is contained in `data_supp_priors` with each claim in column `paper_id` being assigned a prior probability (on the logit scale) of the claim replicating in column `prior_means`.

Aggregation Wrapper Functions

Although there are 27 aggregation methods in total, we grouped methods based on their mathematical properties into eight 'wrapper' functions, denoted by the suffix `WAgg`, the abbreviation of *weighted aggregation*: `LinearWAgg()`, `AverageWAgg()`, `BayesianWAgg()`, `IntervalWAgg()`, `ShiftingWAgg()`, `ReasoningWAgg()`, `DistributionWAgg()`, and `ExtremisationWAgg()`. The specific aggregation *method* is applied according to the `type` argument, whose options are described in each aggregation wrapper functions' help page.

3.2. 'Tidy' Aggregation and Prescribed Inputs

The design philosophy of **aggreCAT** is principled on 'tidy' data (Wickham 2014). Each aggregation method expects a '`data.frame`' or '`tibble`' of judgements (`data_ratings`) as its input, and returns a '`tibble`' consisting of the variables `method`, `paper_id`, `cs` and `n_experts` (see Section 4.1 for illustration of outputs); where `method` is a character vector corresponding to the aggregation method name specified in the `type` argument. Each aggregation is applied as a summary function (Wickham and Grolemund 2017), and therefore returns a single row or observation with a single confidence score `cs` for each claim or `paper_id`. The number of expert judgements summarised in the aggregated confidence score is returned in the column `n_experts`. Because of the tidy nature of the aggregation outputs, multiple aggregations can be applied to the same data with the results of all aggregation methods row bound together in a single `tibble`.

Each aggregation function requires values derived from three-point elicitation (best-estimate, upper and lower bound), however, some methods require only the best-estimates for mathematical aggregation. For every aggregation function, the three-point elicitation values corresponding to the `direct_replication` question are required inputs. Of the `question` and `element`s other than the three-point elicitation `element`s belonging to the direct replication `question`, only the `comprehension` question with the `likert_binary` elements is required – this is an input into CompWAgg, which is used to weight participants judgements.

4. Focal Claim Aggregation

We now demonstrate how judgements elicited from a diverse group of individuals may be mathematically aggregated for a single forecasting problem, using the datasets provided by **aggreCAT**. We illustrate the internal mechanics of the weighting methods and the different data requirements of each of the different types of aggregators – namely; methods with non-weighted linear combinations of judgements, weighted linear combinations of judgements, re-scaled weighted linear combinations of judgements, methods that require supplementary data, and methods that require data elicited from the full IDEA protocol. Each group of methods differs in the type of judgements elicited (single- or three-point estimates), the number of elicitation rounds (one or two rounds), whether multiple forecasts / elicited judgements are used during confidence score computation for a target forecast / claim, and finally whether supplementary data is required for aggregation.

Here we demonstrate the application of aggregation methods for each group of methods using a set of ‘focal claims’ selected from the pilot study dataset supplied with the **aggreCAT** package. Below we subset the dataset `data_ratings` to include a sample of four claims with judgements from five randomly-sampled participants. From these focal claims, we select a target claim for which we will apply an exemplar aggregation method from each mathematical aggregator (Table 1).

```
R> set.seed(1234)
R> focal_claims <- data_ratings %>%
+  dplyr::filter(paper_id %in% c("24", "138", "186", "108"))
R> # select 5 users to highlight in focal claim demonstration
R> focal_users <- focal_claims %>%
+  dplyr::distinct(user_name) %>%
+  dplyr::slice_sample(n=5)
R> # filter out non-focal users from focal claims
R> focal_claims <- focal_claims %>%
+  dplyr::right_join(focal_users, by = "user_name")
```

```
R> focal_claims

# A tibble: 220 x 7
   round   paper_id user_name  question            element           value group
   <chr>   <chr>    <chr>      <chr>               <chr>             <dbl> <chr>
 1 round_1 108      lvr6dwbqag comprehension       likert_binary         7 UOM1
 2 round_1 108      lvr6dwbqag direct_replication three_point_upper    90 UOM1
 3 round_1 108      lvr6dwbqag direct_replication three_point_lower    40 UOM1
 4 round_1 108      lvr6dwbqag belief_binary       binary_question       1 UOM1
 5 round_1 108      lvr6dwbqag involved_binary     binary_question      -1 UOM1
 6 round_1 108      lvr6dwbqag direct_replication three_point_best     65 UOM1
 7 round_1 108      f35wd3nhdz direct_replication three_point_upper    60 UOM3
 8 round_1 108      f35wd3nhdz direct_replication three_point_lower    40 UOM3
 9 round_1 108      f35wd3nhdz direct_replication three_point_best     51 UOM3
10 round_1 108      f35wd3nhdz comprehension       likert_binary         6 UOM3
# ... with 210 more rows
```

Claim ID	User Name	Lower Bound	Best Estimate	Upper Bound
108	5nuvx01cj5	70	80	90
108	f35wd3nhdz	60	80	90
108	g31jx5dffs	70	85	90
108	kf7fxabzu0	50	60	70
108	lvr6dwbqag	40	65	90

Table 1: Focal Claim Data: Round 2 expert judgements for claim 108 derived from a subset of 5 claims and 5 participants from data ratings. Judgements are displayed as percentages.

4.1. Non-weighted linear combination of judgements

We first demonstrate the mechanics of mathematical aggregation and its implementation using the **aggreCAT** package with the simplest, unweighted aggregation method, `ArMean`. All other aggregation methods take this underlying computational blueprint, and expand on it according to the aggregation methods' requirements (See Box 1 for details). `ArMean` (Equation 2) takes the unweighted linear average (i.e. arithmetic mean) of the best estimates, $B_{i,c}$.

$$\hat{p}_c\,(ArMean) = \frac{1}{N}\sum_{i=1}^{N} B_{i,c} \tag{2}$$

Below we demonstrate the application of `ArMean` on a single claim `108` for a subset of participants who assessed this claim. We also illustrate this aggregation visually in Figure 2. `ArMean` is applied using the aggregation method `AverageWAgg()`, which is a wrapper function for several aggregation methods that calculate different types of averaged best-estimates (see `?AverageWAgg`). The function returns the Confidence Score for the claim in the form of a '`tibble`':

```
R> focal_claims %>%
+   dplyr::filter(paper_id == "108") %>%
+   AverageWAgg(type = "ArMean")

# A tibble: 1 x 4
  method paper_id    cs n_experts
  <chr>  <chr>    <dbl>     <int>
1 ArMean 108         74         5
```

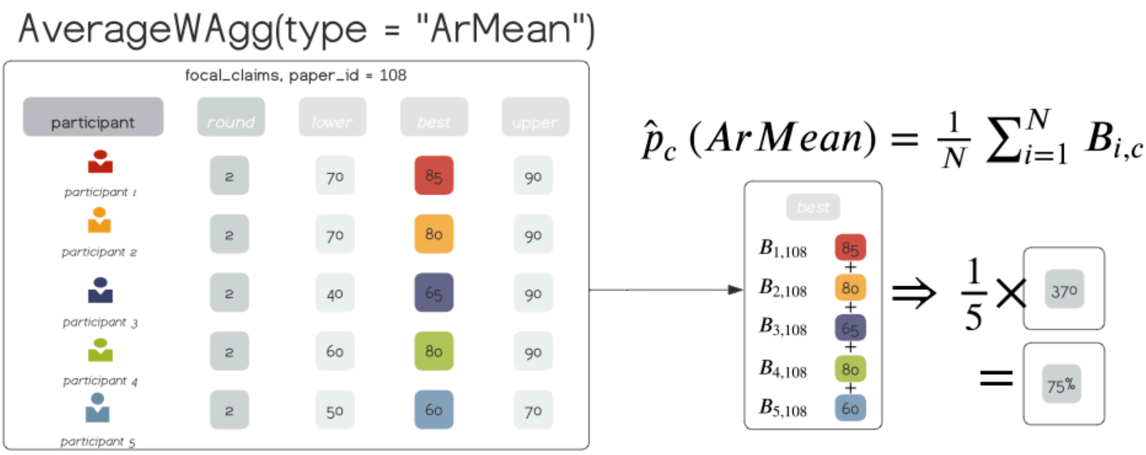

Figure 2: ArMean with `AverageWAgg()` uses the Estimates (shown in colour) from each participant to compute the mean. We illustrate this using a single claim `108` for a subset of 5 out of 25 participants from the `data_ratings` dataset. Note that the data representations in this figure are for explanatory purposes only, the data in the actual aggregation is tidy, with long form structure and format.

Box 1: Aggregation Workflow Blueprint

Argument Structure and Expected Form

Each aggregation *wrapper* function takes the following arguments: `expert_judgements`, `type`, `name`, `placeholder` and `percent_toggle`:

```
R> args(AverageWAgg)
function (expert_judgements, type = "ArMean", name = NULL, placeholder = FALSE,
    percent_toggle = FALSE, round_2_filter = TRUE)
NULL
```

The aggregation *method* to be applied by the aggregation *function*, is specified by the `type` argument, defaulting to `ArMean` in the above example. The resultant `tibble` of Confidence Scores includes the `name` of the aggregation method applied, defaulting to the `type` argument, but this can be overridden by the user if they supply a non-`NULL` value to `name`.

Percentage values, counts, or other non probabilistic quantities are the default expected value type for ratings supplied to the `expert_judgements` argument of aggregation functions. By overriding the default value for the argument `percent_toggle` with `TRUE`, percentage values are converted to probabilities by dividing judgements over 100 within the aggregation functions.

When working with regularly updated data and developing a reproducible pipeline (Yenni, Christensen, Bledsoe, Supp, Diaz, White, and Ernest 2019) , it can be useful to put aggregation methods into 'placeholder' mode, whereby a placeholder value is returned by the aggregation function instead of computing a Confidence Score using the aggregation method. By setting `placeholder` to `TRUE`, the user can supply a placeholder Confidence Score, which defaults to 65%, the approximate average replication rate of SBS research claims (Camerer *et al.* 2018). Should the user wish to set an alternative value, they can create a modified version of `method_placeholder()` for themselves and store this within the global environment. This function will then be called by the aggregation method when the `placeholder` argument is set to `TRUE`.

Some functions expect additional arguments, especially those that rely on additional or supplementary data. See the *man* pages for details of additional arguments.

Mathematical Aggregation Computational Workflow Blueprint

Each aggregation function follows a general computational workflow 'blueprint' whereby the primary dataset `data_ratings`, parsed to the `expert_judgements` argument, is first pre-processed by `pre_process_judgements()`, weights are computed if applicable, subsequently the aggregation method is applied using `dplyr::summarise()`, and then finally the aggregated data is parsed to `postprocess_judgements()`.

The `preprocess_judgements()` function parses the primary dataset `data_ratings` through the argument `expert_judgements` to filter the required quantitative inputs for the aggregation method at hand. It uses two filtering arguments to control which round of judgements are used as inputs (`round_2_filter`), and whether the full set of three-point elicitation judgements should be used, or whether other additional elements must be returned (`three_point_filter`), including the `likert_binary` elements for participants' comprehensibility ratings, and the plausibility ratings under `binary_question` in column `element`. `three_point_filter` defaults to `TRUE` to provide only direct replication questions and associated values. Nearly all aggregation functions use only the round 2 judgements, so the `round_2_filter` defaults to `TRUE` (See Table 5 for required inputs of all aggregation methods). `preprocess_judgements()` further pre-processes the data to remove missing data, and returnq the data into an appropriate structure for calculating weights and applying the aggregation function with `dplyr::summarise()`.

```
R> data_ratings %>%
+   dplyr::group_by(paper_id) %>%
+   tidyr::nest()  %>%
+   dplyr::ungroup() %>%
+   dplyr::slice_sample(n = 1) %>%
+   tidyr::unnest(cols = c(data)) %>%
+   preprocess_judgements()

-- Pre-Processing Options --
```

```
i Round Filter: TRUE
i Three Point Filter: TRUE
i Percent Toggle: FALSE
# A tibble: 75 x 5
   round   paper_id user_name  element            value
   <chr>   <chr>    <chr>      <chr>              <dbl>
 1 round_2 118      7l8m7dmjdb three_point_best      50
 2 round_2 118      7l8m7dmjdb three_point_upper     60
 3 round_2 118      7l8m7dmjdb three_point_lower     40
 4 round_2 118      bvpgcjm09w three_point_best      45
 5 round_2 118      bvpgcjm09w three_point_upper     70
 6 round_2 118      bvpgcjm09w three_point_lower     30
 7 round_2 118      mnlww4qlq3 three_point_best      50
 8 round_2 118      mnlww4qlq3 three_point_lower     40
 9 round_2 118      mnlww4qlq3 three_point_upper     60
10 round_2 118      r7kn2x3mnj three_point_best      64
# ... with 65 more rows
```

The `preprocess_judgements()` function is exposed to the user to allow for data formatting in preparation for plotting, e.g. with **ggplot2** (Wickham 2016), or for developing bespoke aggregation functions / methods not supplied in `aggreCAT`.

For some aggregation methods, weights are necessary, and thus are computed prior to aggregation. Some aggregation methods compute weights using separate weighting functions (See Table 5), however, for aggregation methods with simpler weight computations, these are defined in-function, rather than being modularised.

After application of `preprocess_judgements()`, weights are constructed, and the aggregation method is applied, the function `postprocess_judgements()` then processes the variables into the final data frame that is returned by each aggregation function. The post processing function returns a '`tibble`' consisting of observations equal to the number of unique claims that were parsed to `postprocess_judgements()`, the `method`, `paper_id` , the Confidence Score `value`, as well as the total number of participants `n_experts` whose assessments were used in the aggregation.

4.2. Weighted linear combinations of judgements

We now demonstrate the construction of weights for forecasting performance, as well as the use of uncertainty bounds in addition to the Best Estimates (i.e. three-point estimates) in the aggregation computation. The aggregation method `IntWAgg` weights each participant's best estimate $B_{i,c}$ by the width of their uncertainty intervals, i.e. the difference between an individual's upper $U_{i,c}$ and lower bounds $L_{i,c}$. For a given claim c, a vector of weights for all individuals is calculated from their upper and lower estimates using the weighting function, `weight_interval()`, which calculates the interval width for each individual's estimate for the target claim. The weights are then normalised across the claim (by dividing each weight by the sum of all weights per claim). Normalised weights are then multiplied by the corresponding individual's best estimates $B_{i,c}$ andsummed together into a single Confidence Score (Figure 3).

4.3. Re-scaled weighted linear combinations of judgements

Individuals vary in the interval widths they give across different claims. `IndIntWAgg` is a variation on `IntWAgg` that accounts for cross-claim variation within individuals' assessments by rescaling the interval width weights for individual i for claim c relative to the widest interval provided by that individual across all claims C, (Equation 4). For the target claim, each individual's interval width is divided by the maximum interval width that same individual gave across all claims they have provided judgements for, using the weighting function `weight_nIndivInterval()` (Equation 3). The process of re-scaling is illustrated in Figure 3. Other aggregation methods that re-scale weights by using data from multiple claims other than the target claim under aggregation are `VarIndIntWAgg`, `IndIntAsymWAgg`, `KitchSinkWAgg` (applied with the wrapper function `IntervalWAgg()`) and `GranWAgg` (applied with the wrapper function `LinearWAgg()`), see Table 5.

$$w\_Interval_{i,c} = \frac{1}{U_{i,c} - L_{i,c}} \tag{3}$$

$$\hat{p}_c\left(IntWAgg\right) = \sum_{i=1}^{N} \tilde{w}\_Interval_{i,c} B_{i,c} \tag{4}$$

As for `AverageWAgg()`, when using the wrapper function `IntervalWAgg()` we supply the aggregation method names as a character vector to the `type` argument and the focal claim data frame to the argument `expert_judgements`, using `dplyr::bind_rows()` to bind the resultant Confidence Scores together:

```
R> dplyr::bind_rows(
+   aggreCAT::IntervalWAgg(expert_judgements = focal_claims %>%
+                            dplyr::filter(paper_id == "108"),
+                          type = "IndIntWAgg"),
+   aggreCAT::IntervalWAgg(expert_judgements = focal_claims %>%
+                            dplyr::filter(paper_id == "108"),
+                          type = "IntWAgg")
+ )

# A tibble: 2 x 4
  method     paper_id    cs n_experts
  <chr>      <chr>    <dbl>     <int>
1 IndIntWAgg 108       74           5
2 IntWAgg    108       74.8         5
```

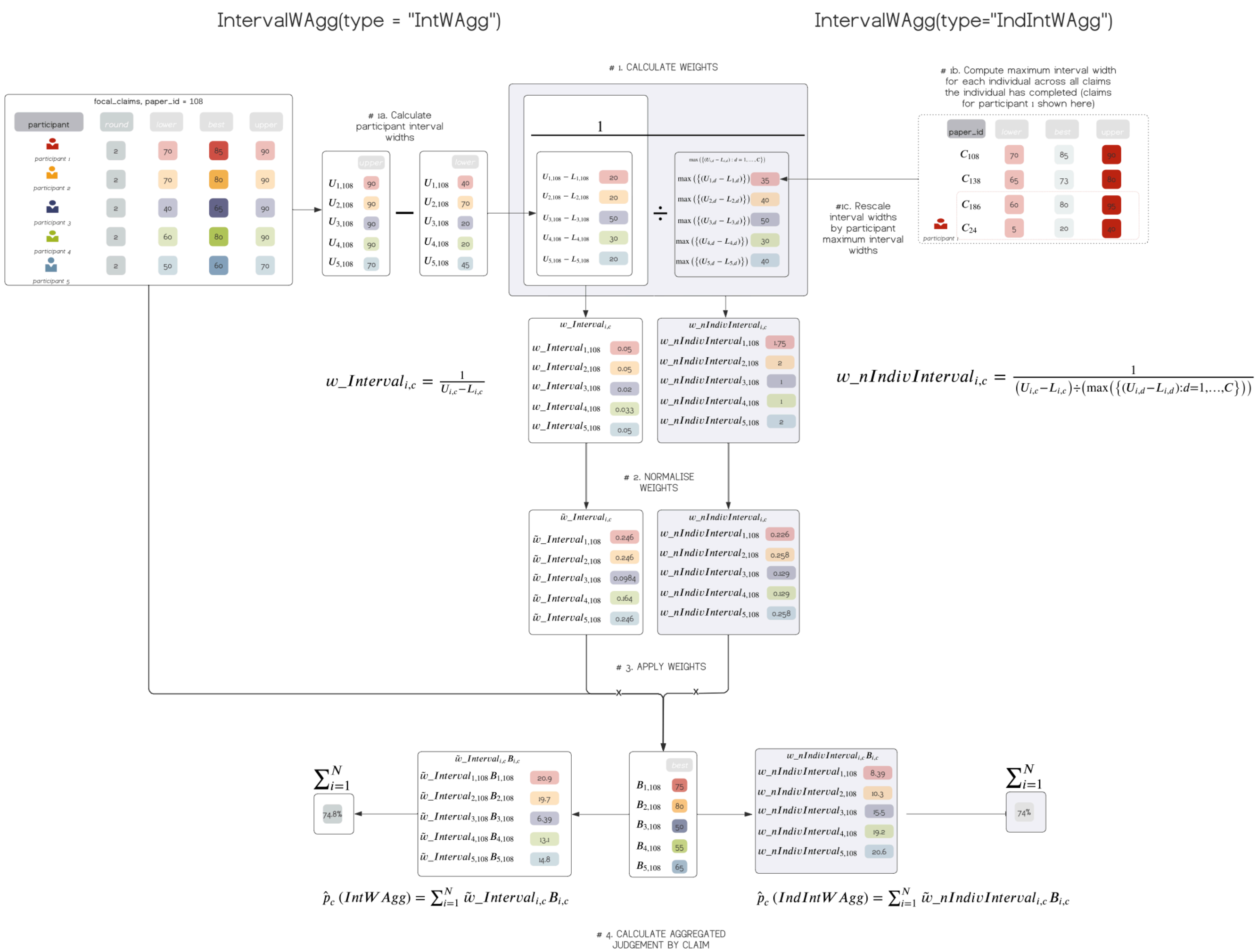

Figure 3: Example applications of mathematical aggregation methods a) `IntWAgg` and b) `IndIntWAgg` using the wrapper function a1) `IntWAgg` uses participants' upper and lower bounds to construct performance weights. b2) This weighting computation is modified in `IndIntWAgg` whereby the weights for each individual are re-scaled by the largest interval width across all claims for a given individual. We exemplify this rescaling process by illustrating the calculation of participant 1's maximum interval width across all claims they assessed in the demonstration dataset `focal_claims`. This is repeated for every individual who has assessed the target claim under aggregation.

4.4. Aggregation Methods Requiring Supplementary Data

In addition to the three-point elicitation dataset `data_ratings`, some aggregation methods require supplementary data inputs collected externally to the repliCATS IDEA protocol. Each aggregation wrapper function that requires supplementary data expects this data to be provided as a '`data.frame`' or '`tibble`' in addition to the main judgements that are provided to the `expert_judements` argument 5.

We illustrate the usage and internal mechanics of this type of aggregation with the method `ReasonWAgg`, which weights participants' best estimates $B_{i,c}$ by the breadth of reasoning provided to support the individuals' estimate (Equation 5). This method is premised on the expectation that multiple (unique) reasons justifying an individual's judgement may indicate their breadth of thinking, understanding and knowledge about both the claim and its context (Hanea *et al.* 2021) while also reflecting their level of engagement and general conscientiousness. These qualities are correlated with improved forecasting (Wintle *et al.* 2021). Thus, greater weighting of best estimates that are accompanied by a greater number of supporting reasons may yield more reliable Confidence Scores.

$$\hat{p}_c\left(ReasonWAgg\right) = \sum_{i=1}^{N} \tilde{w}\_reason_{i,c} B_{i,c} \tag{5}$$

`ReasonWAgg` is applied with the wrapper function `ReasoningWAgg()`, which uses the coded reasoning data `data_supp_reasons` (Section 3.1.2) to compute a vector of weights, $w\_reason_{i,c}$, the number of unique reasons provided by individual i in support of their estimate for claim c (Figure 4). Weights are then normalised across individuals, multiplied by the Best Estimates for that claim $B_{i,c}$ and weighted best estimates are then summed to generate the Confidence Score (Equation 5).

The focal claim selected for aggregation using `ReasonWAgg` is `24`, the round two three-point estimates from the five focal participants for this claim are shown in Table 2. We first prepare the supplementary data for aggregation `data_supp_reasons`, subsetting only the participants contained in our `focal_claims` dataset. We also illustrate a subset of the supplementary data for our five focal participants for the focal claim `24` (see `?data_supp_reasons` for a description of variables):

```
R> data_supp_reasons_focal <- aggreCAT::data_supp_reasons %>%
+   dplyr::right_join(focal_users)
```

```
Joining with `by = join_by(user_name)`
```

```
R> data_supp_reasons_focal %>%
+   dplyr::filter( paper_id == 24) %>%
+   tidyr::pivot_longer(cols = c(-paper_id, -user_name)) %>%
+   dplyr::arrange(name) %>%
+   tidyr::separate(name,
+                   into = c("reason_num", "reason"),
+                   sep = "\\s", extra = "merge") %>%
+   dplyr::select(-reason) %>%
+   dplyr::group_by(paper_id, user_name) %>%
```

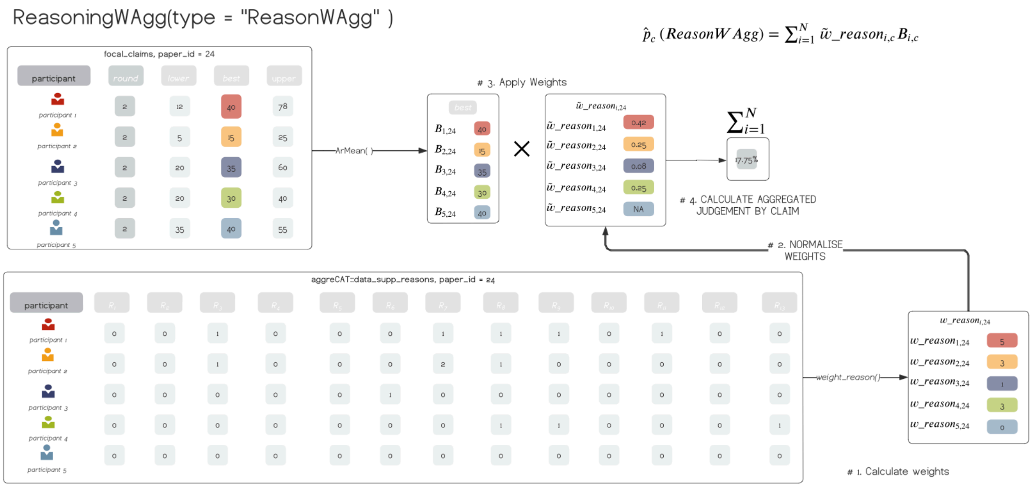

Figure 4: Illustration of the `ReasonWAgg` aggregation method for a subset of five participants who assessed claim `24`. `ReasonWAgg` is applied using the wrapper function `ReasoningWAgg()` and exemplifies aggregation methods that use supplementary data (`data supp ReasonWAgg`) collected externally to the IDEA protocol in the construction of weights and subsequent calculation of Confidence Scores. Weights are constructed by taking the sum of the number of unique reasons made in support of quantitative estimates for each participant, for the target claim.

```
+  tidyr::pivot_wider(names_from = reason_num) %>%
+  dplyr::arrange(user_name)

# A tibble: 5 x 13
# Groups:   paper_id, user_name [5]
  paper_id user_name  RW04  RW15  RW16  RW17  RW18  RW20  RW21  RW22  RW32  RW37
  <chr>    <chr>     <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
1 24       5nuvx01c~     0     1     0     0     0     1     0     2     0     0
2 24       f35wd3nh~     0     0     0     0     0     1     0     0     1     0
3 24       g31jx5df~     0     1     0     0     0     1     0     1     1     1
4 24       kf7fxabz~     0     0     0     0     0     0     0     0     0     0
5 24       lvr6dwbq~     0     0     0     0     0     0     1     0     0     0
# ... with 1 more variable: RW42 <dbl>
```

Claim ID	User Name	round	Lower Bound	Best Estimate	Upper Bound
24	5nuvx01cj5	1	5	15	25
24	5nuvx01cj5	2	5	11	17
24	f35wd3nhdz	1	20	30	40
24	f35wd3nhdz	2	10	15	20
24	g31jx5dffs	1	12	40	78
24	g31jx5dffs	2	5	20	40
24	kf7fxabzu0	1	35	40	55
24	kf7fxabzu0	2	10	30	50

24	lvr6dwbqag	1	20	35	60
24	lvr6dwbqag	2	20	35	50

Table 2: Focal Claim 24 judgements comprising best estimates, upper and lower bounds elicited from five participants. Judgements are displayed as percentages.

Confidence Scores estimating the replicability for claim `24` (Table 2) using the `ReasonWAgg` method are computed using `ReasoningWAgg()` and by providing the supplementary data to the `reasons` argument:

```
R> focal_claims %>%
+  dplyr::filter(paper_id == "24") %>%
+  aggreCAT::ReasoningWAgg(reasons = data_supp_reasons_focal,
+                          type = "ReasonWAgg")
```

Note that if there are zero participants with a Reasoning Score > 0 or all participants are missing a Reasoning Score, the log-odds transformed best estimate is returned instead (See `?AverageWAgg, type="LOArMean"`). The user can choose to flag this behaviour explicitly by setting the argument `flag_loarmean` to `TRUE`, which will generate new columns in the aggregation output '`data.frame`' named `method_applied` (with values `LOArMean` or `ReasonWAgg`), and `no_reason_score`, a logical variable describing whether or not there were no reasoning scores for that claim.

4.5. Bayesian Aggregation Methods

Both Bayesian methods `BayTriVar` and `BayPRIORsAgg` use the full three-point elicitation data, i.e., they use information contained in the uncertainty bound provided by individuals (upper $U_{i,c}$ and lower bounds $L_{i,c}$), in addition to Best Estimates, $B_{i,c}$. Like `IndIntWAgg` and other methods (Table 5), the Bayesian aggregation methods also construct weights from information encoded in participant assessments of claims other than the target claim under aggregation. In fact, the Bayesian methods require more than a single claim's worth of data to work properly execute due mathematical specification of the models (See `?BayesianWAgg` and below for details).

The two Bayesian methods use the elicited probabilities as data to update prior probabilities. `BayTriVar` incorporates three sources of uncertainty in best estimates: variability in best estimates across all claims, variability in estimates across all individuals, and claim-participant variability (which is derived from an individuals' upper and lower bounds). This Bayesian model, implemented using **R2JAGS** (Su and Yajima 2020), takes the log odds transformed individual best estimates, and uses a normal likelihood function to derive a posterior distribution for the probability of replication. The estimated confidence score is the mean of this posterior distribution.

`BayPRIORsAgg` is a modified version of `BayTriVar` where, instead of using default priors, priors are generated from a predictive model that estimates the probability of a claim replicating based on characteristics of the claim and publication (Gould *et al.* 2021). Priors are parsed as supplementary data to the wrapper function `BayesianWAgg()` using the argument `priors` (Section 3.1.3) with each claim having its own unique prior.

We illustrate aggregation of participant judgements using the method BayTriVar to generate a Confidence Score for the claim 108. Note that BayesianWAgg() expects best estimates in the form of probabilities, so to convert elicited values in the form of percentages within the data parsed to expert_judgements to probabilities, the logical value TRUE is supplied to the argument percent_toggle (Box 1):

```
R> focal_claims %>%
+   BayesianWAgg(type = "BayTriVar",
+                percent_toggle = TRUE) %>%
+   dplyr::filter(paper_id == "108")
```

```
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
Graph information:
   Observed stochastic nodes: 20
   Unobserved stochastic nodes: 4
   Total graph size: 230

Initializing model
```

```
# A tibble: 1 x 4
  method    paper_id    cs n_experts
  <chr>     <chr>    <dbl>     <int>
1 BayTriVar 108      0.699         5
```

The Confidence Score calculated for a given claim depends on data for other claims and participants included in the expert_judgements argument other than the target claim, because, by definition, BayesianWAgg() calculates the Confidence Score for a target claim using data from participants' assessments of other claims, and from all other claims in the 'data.frame' parsed to the expert_judgements argument. Because information about other claims than the target claim is used to calculate the Confidence Score for the target claim, what is included in the data supplied to the argument expert_judgements in BayesianWAgg() will alter the Confidence Score. Above, we calculated the Confidence Score for claim 108 but including information from three additional claims included in the focal_claims 'data.frame': 108, 138, 186, 24. However, if we were to supply assessments for only two claims to BayesianWAgg(), then we would observe a different result for focal claim 108:

```
R> focal_claims %>%
+   dplyr::filter(paper_id %in% c("108", "138")) %>%
+   aggreCAT::BayesianWAgg(type = "BayTriVar", percent_toggle = TRUE) %>%
+   dplyr::filter(paper_id == "108")
```

```
Compiling model graph
   Resolving undeclared variables
   Allocating nodes
Graph information:
   Observed stochastic nodes: 10
   Unobserved stochastic nodes: 2
   Total graph size: 116
```

```
Initializing model

# A tibble: 1 x 4
  method    paper_id    cs n_experts
  <chr>     <chr>    <dbl>     <int>
1 BayTriVar 108      0.739         5
```

The Confidence Score shifts from 0.7 to 0.74. Note that `BayesianWAgg()` cannot calculate confidence scores when judgements for only a single claim is provided to `expert_judgements()`, because by definition the underlying Bayesian model calculates variance across multiple claims and multiple participants:

```
R> focal_claims %>%
+   dplyr::filter(paper_id == "108") %>%
+   aggreCAT::BayesianWAgg(type = "BayTriVar",
+                          percent_toggle = TRUE)

Error in `aggreCAT::BayesianWAgg()`:
! Model requires n > 1 ids to successfully execute.
```

Although we have set $n = 2$ as the minimum number of claims for which variance is computed, it is up to the user to determine their own justifiable minimum for reliable variance calculations.

Finally, all of the previous methods illustrated in this section have been used with data generated using the IDEA elicitation protocol, however this elicitation method is not strictly necessary for the of these methods. Methods that *do* require the full IDEA protocol for their correct mathematical implementation, such as `ShiftingWAgg()`, which use two rounds of three-point judgements in which the second round judgements are revised after discussion, are listed in Table 5.

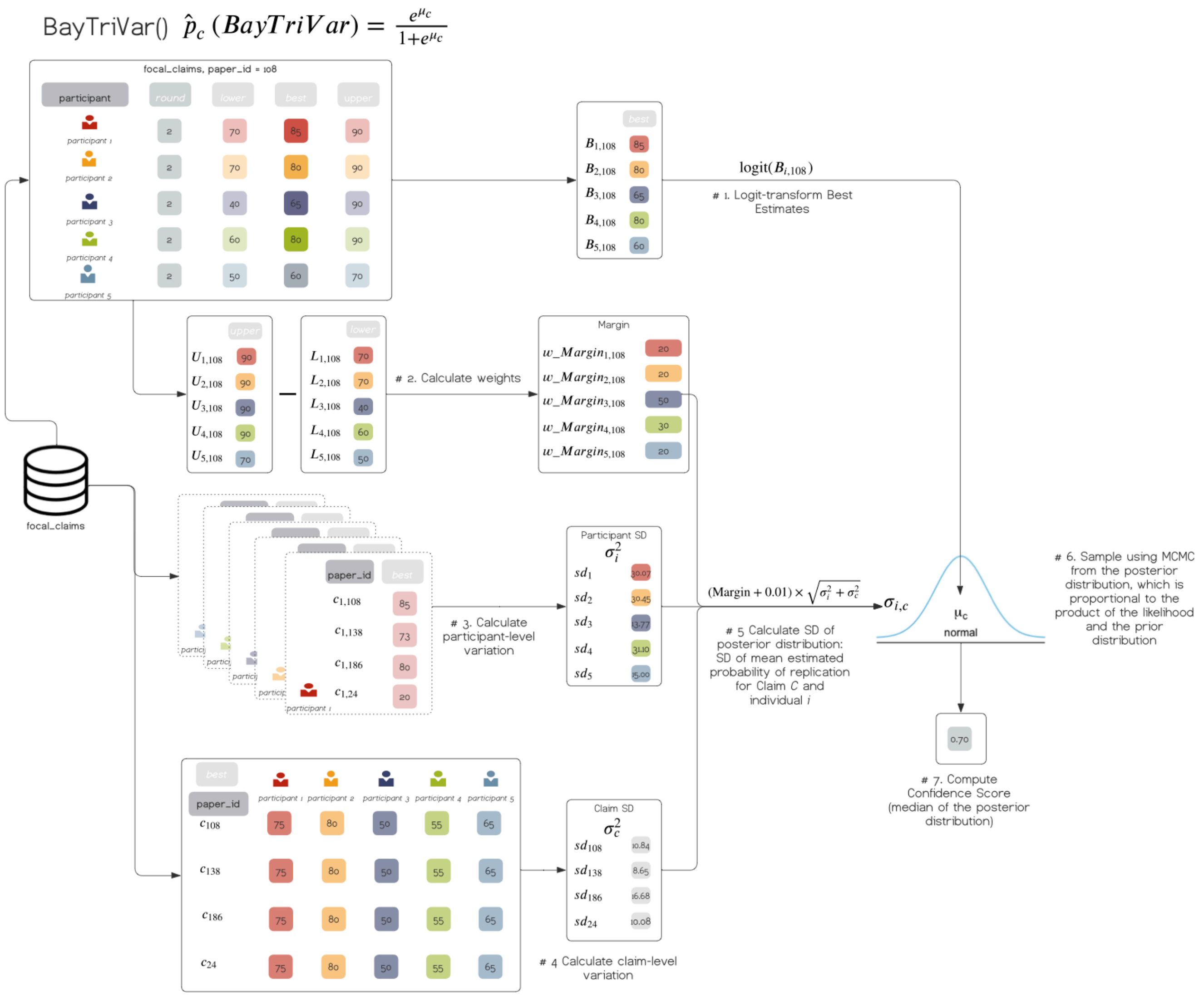

Figure 5: Illustration of BayTriVar applied with `BayesianWAgg()`for a single claim, `paper_id = 108` from the `focal_claims` data object. Note that the claims `138`, `186` and `24` contained in `focal_claims` are used in the calculation of pariticipant-level SD and claim-level SD, thus the Confidence Score returned by BayTriVar is sensitive to the other claims provided to argument `expert_judgements`.

5. An illustrative workflow for use in real study contexts

Throughout the SCORE program, 752 participants assessed more than 4000 unique claims using the repliCATS IDEA protocol, between 7th July 2019 and 25 November 2021. This required batch aggregation over multiple claims, and to generate Confidence Scores for multiple claims. We also applied multiple aggregation methods to the same claim so that we could compare and evaluate the different aggregation methods. We expect that these are not uncommon use-cases,consequently in this section we demonstrate a general workflow for using the **aggreCAT** package to aggregate expert judgements using pilot data from DARPA SCORE program generated by the repliCATS project.

5.1. Generating multiple forecasts

During expert-elicitation the analyst or researcher may be tasked with generating multiple forecasts for different problems or questions, and therefore it is useful to batch the aggregation. Since the **aggreCAT** package is designed using the principles of *tidy* data analysis (Wickham, Averick, Bryan, Chang, McGowan, François, Grolemund, Hayes, Henry, Hester, Kuhn, Pedersen, Miller, Bache, Müller, Ooms, Robinson, Seidel, Spinu, Takahashi, Vaughan, Wilke, Woo, and Yutani 2019), each aggregation function accepts a '`data.frame`' of raw three-point forecasts for one or more claims, C, parsed to the argument `expert_judgements`. The data pre-processing and aggregation methods are applied using a combination of calls to **tidyverse** functions, including `summarise` and `mutate`. From the user's perspective, this means that data processing and application of the aggergation methods is handled internally by the **aggreCAT** package, rather than by the user. The user is therefore free to focus their attention on the interpretation and analysis of the forecasts. Here we demonstrate the application of the `ArMean` aggregation method to four focal claims simultaneously:

```
AverageWAgg(focal_claims, type = "ArMean")
```

```
# A tibble: 4 x 4
  method paper_id    cs n_experts
  <chr>  <chr>    <dbl>     <int>
1 ArMean 108       74           5
2 ArMean 138       68.6         5
3 ArMean 186       57.6         5
4 ArMean 24        22.2         5
```

5.2. Comparing and Evaluating Aggregation Methods

In real study contexts, such as that of the repliCATS project in the DARPA SCORE program, it is of interest to compute Confidence Scores using multiple aggregation methods so that their performance might be evaluated and compared. Since different methods offer different mathematical properties, and therefore might be more or less appropriate depending on the purpose of the aggregation and forecasting, a researcher or analyst might want to check how the different assumptions embedded in different aggregation methods influence the final Confidence Scores for a forecast – i.e. how robust are the results to different methods and therefore to different assumptions?

From a computational perspective, multiple aggregation methods must first be applied to the forecast prior to comparison and evaluation. This can be achieved by applying each different aggregation method to `focal_claims`, and binding the results together with **dplyr**'s `row_-bind()`. However, more elegant and succinct solutions can be implemented using **purrr**'s `map_dfr()` function (Henry and Wickham 2020, see Listing 1 and Listing 2).

```
R> confidenceSCOREs <-
+   dplyr::bind_rows(
+     AverageWAgg(focal_claims,
+                 "ArMean",
+                 percent_toggle = TRUE),
+     IntervalWAgg(focal_claims,
+                  "IndIntWAgg",
+                  percent_toggle = TRUE),
+     IntervalWAgg(focal_claims,
+                  "IntWAgg",
+                  percent_toggle = TRUE),
+     ShiftingWAgg(focal_claims,
+                  "ShiftWAgg",
+                  percent_toggle = TRUE),
+     BayesianWAgg(focal_claims,
+                  "BayTriVar",
+                  percent_toggle = TRUE),
+     ReasoningWAgg(focal_claims,
+                   reasons = aggreCAT::data_supp_reasons,
+                   percent_toggle = TRUE)
+   )

Compiling model graph
   Resolving undeclared variables
   Allocating nodes
Graph information:
   Observed stochastic nodes: 20
   Unobserved stochastic nodes: 4
   Total graph size: 230

Initializing model

R> confidenceSCOREs

# A tibble: 24 x 4
   method     paper_id    cs n_experts
   <chr>      <chr>    <dbl>     <int>
 1 ArMean     108      0.74          5
 2 ArMean     138      0.686         5
 3 ArMean     186      0.576         5
 4 ArMean     24       0.222         5
 5 IndIntWAgg 108      0.740         5
 6 IndIntWAgg 138      0.685         5
```

```
 7 IndIntWAgg 186      0.561         5
 8 IndIntWAgg 24       0.19          5
 9 IntWAgg    108      0.748         5
10 IntWAgg    138      0.694         5
# ... with 14 more rows
```

After generating Confidence Scores using various aggregation methods, we then evaluate the forecasts. We evaluated the repliCATS pilot study forecasts against the outcomes of previous, high-powered replication studies (Hanea *et al.* 2021), which are contained in the `data\_outcomes` dataset published with **aggreCAT**. In this dataset, each claim `paper_id` is assigned an `outcome` of `0` if the claim did not replicate and `1` if the claim was successfully replicated:

```
R> aggreCAT::data_outcomes %>%
+  head()
```

```
# A tibble: 6 x 2
  paper_id outcome
  <chr>      <dbl>
1 100            1
2 102            0
3 103            0
4 104            1
5 106            0
6 108            1
```

The function `confidence_score_evaluation()` evaluates a set of aggregated forecasts or Confidence Scores against a set of known or observed outcomes, returning the Area Under the ROC Curve (AUC), the Brier score, and classification accuracy of each method (Table 3):

Method	AUC	Brier Score	Classification Accuracy
ArMean	1.00	0.10	100%
BayTriVar	1.00	0.12	100%
IndIntWAgg	1.00	0.10	100%
IntWAgg	1.00	0.09	100%
ReasonWAgg	1.00	0.09	100%
ShiftWAgg	1.00	0.13	75%

Table 3: AUC and Classification Accuracy for forecasts from the aggregation methods 'ShiftWAgg', 'ArMean', 'IntWAgg', 'IndIntWAgg', 'ReasonWAgg' and 'BayTriVar' for a subset of the repliCATS pilot study claims (focal_claims) and known outcomes.

5.3. Visualising Judgements, Confidence Scores and Forecast Performance

We include two functions for visualising comparison and evaluation of Confidence Scores across multiple aggregation methods for a suite of forecasts from multiple participants, `confidence_scores_ridgeplot()` and `confidencescore_heatmap()`. `confidence_scores_ridgeplot()` generates ridgeline plots using **ggridges** (Wilke 2021), and displays the distribution of predicted outcomes across a suite of forecasts for each aggregation method, grouped into separate 'mountain ranges' according to the mathematical properties of the aggregation method (Figure 6).

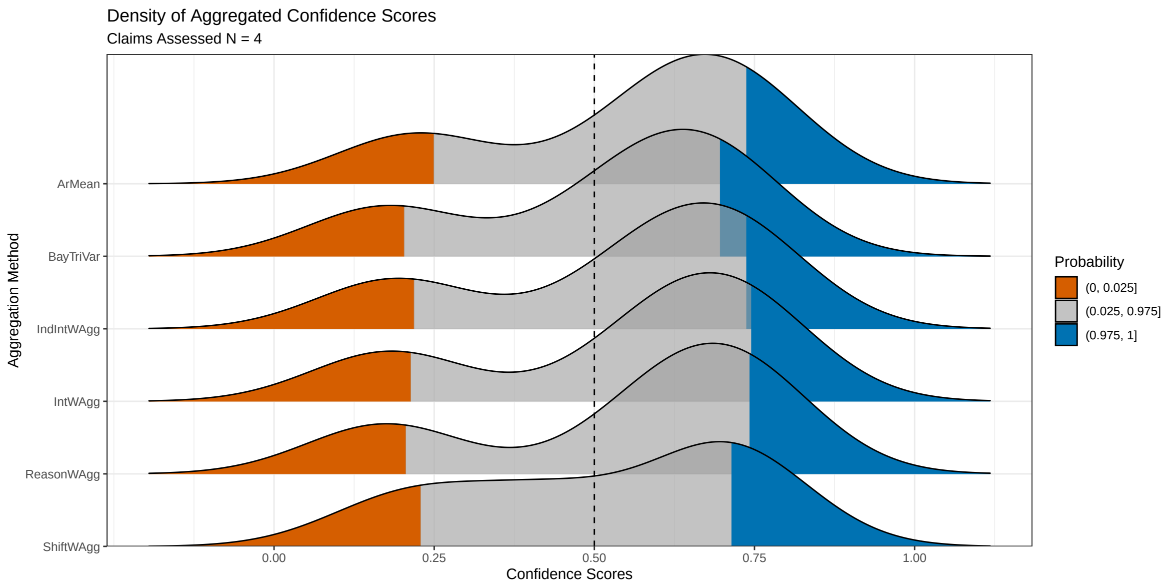

Figure 6: Ridgeline plots illustrating the distribution of aggregated Confidence Scores for the tibble `confidenceSCOREs`, grouped according to mathematical properties of each method.

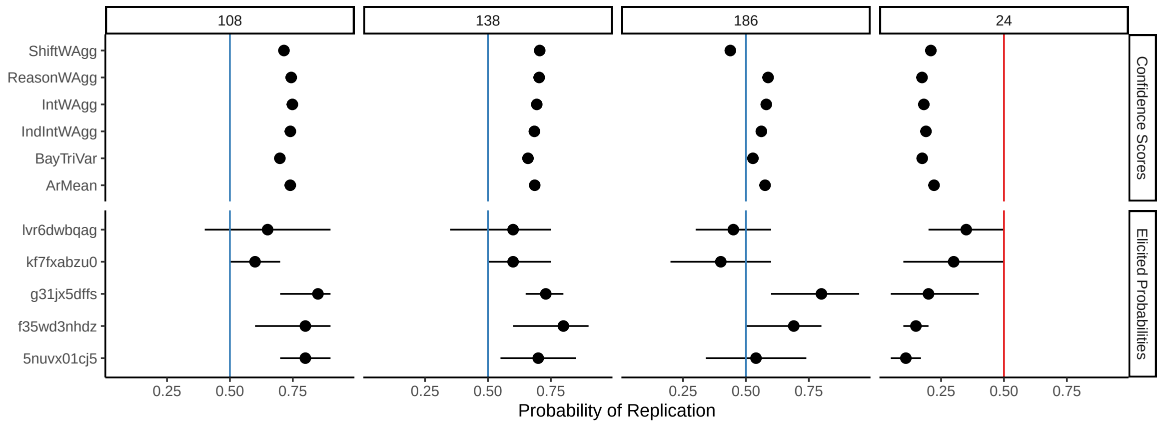

Figure 7: Confidence Scores for the aggregation methods `ArMean`, `BayTriVar`, `IntWAgg`, `IndIntWAgg`, `ReasonWAgg` and `ShiftWAgg` for four claims. Participants' three-point best estimates are displayed as black points, and their upper and lowr bounds displayed as black error bars. Confidence Scores are displayed as points within the upper row of plots. Lines are displayed vertically at the 0.5 probability mark, and their colour denotes the observed outcome under previous large-scale replication projects.

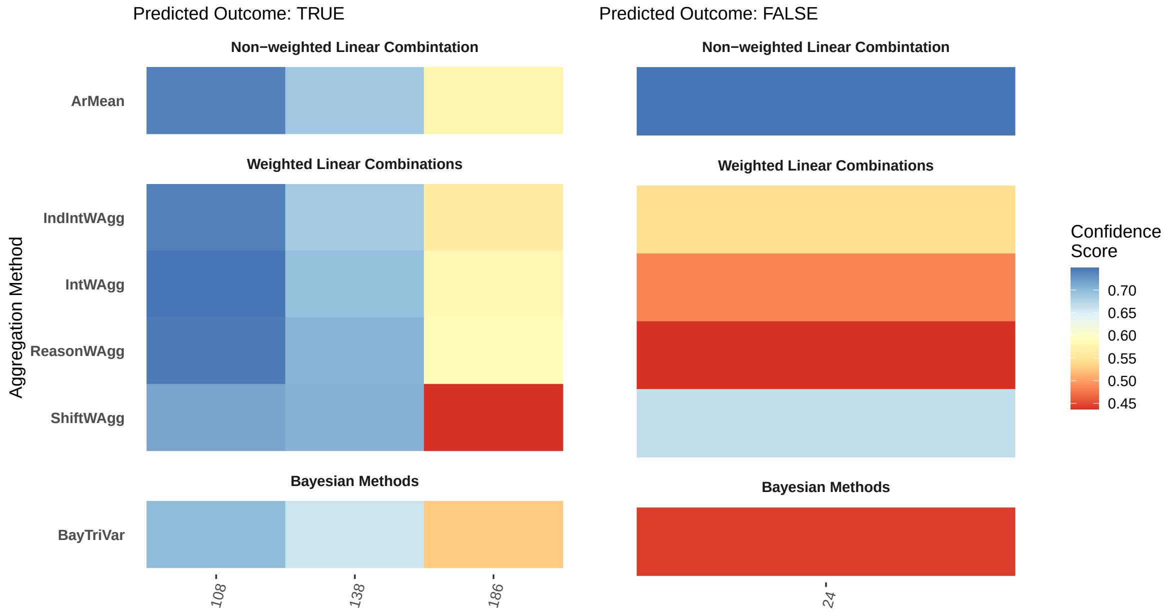

Figure 8: Blocked heatmap visualisation of confidence scores is useful for visually comparing aggregation methods and evaluating them against a set of known outcomes. In this example, Confidence Scores generated by six aggregation methods for the repliCATS pilot study are visualised for 25 claims. Claims where known outcomes succesfully replicated `outcome == TRUE` are presented in heatmap blocks on the left, and claims that failed to replicate are presented in heatmap blocks on the right. Confidence Scores generated by different aggregation methods are positioned along the y-axis, with vertical groupings according to the methods' mathematical properties. Colour and intensity of cells indicates the direction and degree of deviation respectively of the Confidence Scores from the known outcomes.

While `confidencescore_heatmap()` is useful for comparison of aggregation methods, `confidencescore_heatmap()` is useful for visual comparative *evaluation* of aggregation methods. `confidencescore_heatmap()` generates heatmaps of forecasted Confidence Scores for each aggregation method included in the dataset provided to the argument `confidence_-scores` organised with unique aggregation methods on the y-axis, and separate forecasts or `paper_ids` along the y-axis (Figure 8). The heatmap is blocked vertically according to the mathematical characteristics of each aggregation method, and horizontally into two groups, according to the binary outcomes in `data_outcomes`.

Horizontal grouping facilitates quick and simple evaluation of the aggregation methods. Perfectly accurate aggregation methods show dark blue squares in the left heatmap blocks, where the outcomes were `1` or `TRUE`, and dark red squares on the right heatmap blocks, where the actual outcomes were `0` or `FALSE`. Deviation from this expectation indicates which aggregation methods for which claim/forecast, for which outcome type were inaccurate, and to what degree.

For example, in Figure 8, for the dataset `confidenceSCOREs` the successful replication of most claims was accurately forecasted by most methods, except for several claims. Some methods performed better than others for some claims (e.g. `ReasonWAgg` for claims `109` and `138`.

Finally, creating bespoke user-defined plots is relatively easy – because **aggreCAT** functions return tidy '`data.frame`'s or '`tibble`'s, we can easily manipulate the raw judgements, aggregated Confidence Scores and outcome data to plot them with **ggplot2** (Wickham 2016) or other visualisation package. Below we plot the aggregated Confidence Scores along with the three-point judgements (subset using `preprocess_judgements()` on `focal_claims`, transforming judgements in percentages to probabilities by setting `percent_toggle` to `TRUE`, Figure 7, Listing 4).

5.4. Extending aggreCAT to other datasets

The aggregation methods supplied by the **aggreCAT** package can easily be applied to other forecasting problems. The only requirements are that the data inputs adhere to the required format (Box 1), and that the expert judgements are elicited using the appropriate method, as required by each aggregation method (see Table 5).

Judgement data provided to the `expert_judgements`, `data_justifications` or any supplementary data inputs argument must contain the requisite column names, and be of the correct data type, as described in each method's documentation (see `?data_ratings`, for example). At minimum the user must supply to `expert_judgements`: the `round` under which each judgement is elicited, a unique ID for each different forecasting problem `paper_id`, a unique `user_name` for each individual, and the `element` of the three point elicitation that the recorded response or `value` in that row corresponds to. The data is stored in long or tidy format such that each row or observation in the '`data.frame`' references only a single `element` of a participants' set of three point elicitation values. When applying aggregation methods requiring supplementary data to the elicitation data, the analyst should also adhere to the requirements stipulated for the relevant supplementary dataset described in the documentation.

Although several aggregation methods *require* judgements that are elicited using the IDEA protocol (See Table 5 for exceptions), most aggregation methods require only a single round of elicitation that generates a set of three points; a best estimate, and upper and lower bounds about those estimates, even if the IDEA protocol was used to elicit judgements. Hence, the aggregation functions contained in the **aggreCAT** package are unsuitable for use with judgements elicited with methods that aggregate behaviourally (e.g. using consensus) and therefore result in a single forecast value. Where the analyst elicits judgements for only a single round, the analyst should record the round in the judgements data as the character string `"round_1"`, and set the `round_2_filter` argument to `FALSE` in the aggregation wrapper function call.

Should the analyst wish to create their own aggregation functions, pre- and post-processing functions may be leveraged inside the functions (`preprocess_judgements()` and `postprocess_judgements()`, respectively), as we have illustrated in data preparation for Figure 7 (Listing 4). These processing functions modularise key components of the aggregation's computational implementation – namely the data wrangling that occurs before and after the actual mathematical aggregation.

Preparing your own Elicitation Data

We demonstrate how to prepare data for applying the **aggreCAT** aggregation methods with data collected using the IDEA protocol for an environmental conservation problem (Arlidge, Alfaro-Shigueto, Ibanez-Erquiaga, Mangel, Squires, and Milner-Gulland 2020) . Participants were asked "How many green turtles in winter per month would be saved using a total gillnet ban, with gear switching to lobster potting or hand line fishing required?". We take the required data for the `expert_judgements` argument from Table S51 of Arlidge et al. (2020), make the data long instead of wide, and then add the required additional columns `paper_id` and `question`:

```
R> green_turtles <-
+  dplyr::tribble(~user_name, ~round, ~three_point_lower,
+          ~three_point_upper, ~three_point_best,
+          "L01", 1,    10.00,  16.43,  10.00,
+          "L01", 2,    10.00,  16.43,  10.00,
+          "L02", 1,    500.00, 522.50, 500.00,
+          "L02", 2,    293.75, 406.25, 350.00,
+          "L03", 1,    400.00, 512.50, 400.00,
+          "L03", 2,    300.00, 356.25, 300.00,
+          "L04", 1,    32.29,  65.10,  41.67,
+          "L04", 2,    32.29,  65.10,  41.67,
+          "L05", 1,    6.67,   7.74,   6.67,
+          "L05", 2,    6.67,   7.74,   6.67) %>%
+  dplyr::group_by(user_name) %>% # pivot longer
+  tidyr::pivot_longer(cols = tidyr::contains("three_point"),
+               names_to = "element", values_to = "value") %>%
+  dplyr::mutate(paper_id = 1,
+         round = ifelse(round ==1, "round_1", "round_2"),
+         question = "direct_replication")
```

We can then apply multiple aggregation methods, using the same approach implemented for aggregation of the `focal_claims` dataset (Listing 3), with aggregated Confidence Scores shown in Table 4. Note that because the judgements are absolute values rather than probabilities, we set the `percent_toggle` argument for each aggregation wrapper function to FALSE (Listing 3).

Method	Question ID	Confidence Score	N (experts)
ArMean	1	141.67	5
IndIntWAgg	1	141.67	5
IntWAgg	1	15.26	5
ShiftWAgg	1	328.85	5

Table 4: Example aggregation of non-percentage / non-probabilistic estimates with several aggregation methods using Green Turtle dataset (Arlidge et al. 2020).

6. Summary and Discussion

The **aggreCAT** package provides a diverse suite of methods for mathematically aggregating judgements elicited from groups of experts using structured elicitation procedures, such as the IDEA protocol. The **aggreCAT** package was developed by the repliCATS project as a part of the DARPA SCORE program to implement the 27 aggregation methods described in Hanea et al. (2021).

There are very few open-source tools available to the researcher wishing to mathematically aggregate judgements. The **aggreCAT** package is therefore unique in both the diversity of aggregation methods it contains, as well as in its computational approach to implementing the aggregation methods. There is no other R or other software package with so many aggregation methods, and methods that use proxies of forecasting accuracy using weights.

The **aggreCAT** package is production-ready for application to data elicited during either a single workshop, or for contexts where data collection may be ongoing and continuous analysis is used for automating aggregation. Unlike other aggregation packages, the **aggreCAT** package is designed to work within the *tidyverse*. The package is premised on the principles of *tidy* data analysis whereby the user supplies '`data.frame`'s of elicited judgements, and the aggregation methods return '`data.frame`'s of aggregated forecasts. The benefits of this approach are three-fold. Firstly, the work of data-wrangling and application of the aggregation methods is handled internally by the aggregation methods, so that the researcher can focus on analysis and interpretation of the aggregation outputs. This is critical in data-deficient contexts where rapid assessments are needed, which is a common use-case for the use of expert derived forecasts. Secondly, the **aggreCAT** package is easily paired with other tidyverse tools, such as **purrr**, **dplyr**, and **ggplot2**, as exemplified through the repliCATS workflow described in Section 5.

Thirdly, application of the **aggreCAT** package aggregation methods and performance evaluation tools is scalable, which is evidenced by the application of the **aggreCAT** package to forecast the replicability of over 4000 research claims by the repliCATS project. The scalability and placeholder functionality allow the **aggreCAT** package to be built into production-ready pipelines for more complicated analyses where there are multiple forecasts being elicited and aggregated, where there are numerous participants, and where multiple aggregation methods are applied.

Finally, through the provision of built-in performance metrics, the analyst is able to 'ground-truth' and evaluate the forecasts against known-outcomes, or alternative forecasting methods (e.g. Arlidge *et al.* 2020).

The **aggreCAT** package is easily extensible and production-ready. Each aggregation function follows a consistent modular blueprint, wherein data-wrangling of the inputs and outputs of aggregation is largely handled by pre- and post-processing functions (`preprocess_judgements()` and `postprocess_judgements()`, respectively). This design expedites debugging by making it easier to pinpoint the exact source of errors, while also permitting the user to easily create their own custom aggregation methods.

Although the package currently requires data inputs to conform to nomenclature specific to the repliCATS project, future releases of the **aggreCAT** package will relax the data-input requirements so they are more domain-agnostic. We believe this to be a minimal barrier for adoption and application of the **aggreCAT** package. Ecologists should be no stranger to these naming conventions for data requirements, with packages like **vegan** also imposing strict nomenclature (Oksanen, Blanchet, Friendly, Kindt, Legendre, McGlinn, Minchin, O'Hara, Simpson, Solymos, Stevens, Szoecs, and Wagner 2020). We have illustrated how to extend and apply the package to data from domains beyond forecasting the replicability of research claims through our minimal example using forecasts generated using the IDEA protocol for a fisheries and conservation problem.

The package will be actively maintained into the future, beyond the life of the DARPA SCORE program. Bug reports and feature-requests can easily be lodged on the **aggreCAT** GitHub repository using reproducible examples created with **reprex** (Bryan, Hester, Robinson, and Wickham 2021) on the repliCATS pilot study datasets shipped with the **aggreCAT** package.

We have described the computational implementation of the aggregation methods and supporting tools within the **aggreCAT** package, providing usage examples and workflows for both simple and more complex research contexts. Consequently, this paper should fully equip the analyst for applying the aggregation functions contained within the **aggreCAT** package to their own data. Where the analyst is uncertain as to *which* aggregation method is best for their particular research goals, the reader should consult Hanea et al. (2021) for a discussion on the mathematical principles and hypotheses underlying the design of the aggregation methods, as well as a comparative performance evaluation of each of the methods. In conclusion, the **aggreCAT** package will aid researchers and decision analysts in rapidly and easily analysing the results of IDEA protocol and other structured elicitation procedures where mathematical aggregation of human forecasts is required.

Table 5: Summary of aggregation methods and functions, including data requirements and sources.

Method	Description	Data Requirements	Weighting Function	Elicitation Rounds	Elicitation Method	Data Sources
AverageWAgg(): Averaged best estimates						
ArMean	Arithmetic mean of the best estimates		NA - Estimates are equally weighted	1	Single-point	$B_{i,c}$
Median	Median of the best estimates		NA - Estimates are equally weighted	1	Single-point	$B_{i,c}$
GeoMean	Geometric mean of the best estimates		NA - Estimates are equally weighted	1	Single-point	$B_{i,c}$
LOArMean	Arithmetic mean of the log odds transformed best estimates		NA - Estimates are equally weighted prior to transformation	1	Single-point	$B_{i,c}$
ProbitArMean	Arithmetic mean of the probit transformed best estimates		NA - Estimates are equally weighted prior to transformation	1	Single-point	$B_{i,c}$
LinearWAgg() Linearly-weighted best estimates						
DistLimitWAgg	Weighted by the distance of the best estimate from the closest certainty limit. Best-estimates closest to certainty limits are more strongly weighted		Calculated internally	1	Single-point	$B_{i,c}$
GranWAgg	Weighted by the granularity of best estimates		Calculated internally	1	Single-point	$B_{i,c}$
Judgement	Weighted by user-supplied weights at the judgement level	‘data.frame/tibble‘ with three columns (‘paper_id‘, ‘user_name‘, ‘weight‘)	Calculated internally	1	Single-point	$B_{i,c}$
Participant	Weighted by user-supplied weights at the participant level	‘data.frame‘/‘tibble‘ with two columns (‘user_name‘, ‘weight‘)	Calculated internally	1	Single-point	$B_{i,c}$
OutWAgg	Outliers are down-weighted.		‘weight_outlier()‘	1	Single-point	$B_{i,c}$
IntervalWAgg() Linearly-weighted best estimates, with weights influenced by interval widths						
IntWAgg	Weighted by interval width		‘weight_-interval()‘	1	Three-point	$B_{i,c}$, $U_{i,c}$, $L_{i,c}$
IndIntWAgg	Weighted by the re-scaled interval width (interval width relative to largest interval width provided by individual i.		‘weight_-nIndivInterval()‘	1	Three-point	$B_{i,c}$, $U_{i,c}$, $L_{i,c}$, $U_{i,d}$, $L_{i,d}$
AsymWAgg	Weighted by asymetry of intervals		‘weight_asym()‘, ‘weight_-nIndivInterval()‘	1	Three-point	$B_{i,c}$, $U_{i,c}$, $L_{i,c}$

Table 5: Summary of aggregation methods and functions, including data requirements and sources. *(continued)*

Method	Description	Data Requirements	Weighting Function	Elicitation Rounds	Elicitation Method	Data Sources
IndIntAsymWAgg	Weighted by individuals' interval widths and their asymetry		‘weight_asym()‘, ‘weight_nIndivInterval()‘	1	Three-point	$B_{i,c}$, $U_{i,c}$, $L_{i,c}$, $U_{i,d}$, $L_{i,d}$
VarIndIntWAgg	Weighted by the variation in individuals' interval widths across estimates		‘weight_varIndivInterval()‘	1	Three-point	$B_{i,c}$, $U_{i,c}$, $L_{i,c}$, $U_{i,d}$, $L_{i,d}$
KitchSinkWAgg	Weighted by everything but the kitchen sink - rewards narrow and assymetric intervals as well as the variability of individuals' interval widths across estimates.		‘weight_asym()‘, ‘weight_nIndivInterval()‘, ‘weight_varIndivInterval()‘	1	Three-point	$B_{i,c}$, $U_{i,c}$, $L_{i,c}$, $U_{i,d}$, $L_{i,d}$
ShiftingWAgg()	**Weighted by judgemetns that shift most after discussion**					
ShiftWAgg	Accounts for shifts in individuals' best-estimates, upper and lower bounds between rounds		Calculated internally	2	IDEA protocol or other structured protocol that generates multiple rounds of judgements using three-point elicitation	$B1_{i,c}$, $U1_{i,c}$, $L1_{i,c}$, $B1_{i,c}$, $U1_{i,c}$, $L1_{i,c}$
BestShiftWAgg	Weights constructed from shifts in best-estimates		Calculated internally	2	IDEA protocol or other structured protocol that generates multiple rounds of judgements of single point-estimates	$B_{i,c}$
IntShiftWAgg	Weights constructed from shifts in interval widths		Calculated internally	2	IDEA protocol or other structured protocol that generates multiple rounds of judgements using three-point elicitation	$B_{i,c}$, $U_{i,c}$, $L_{i,c}$
DistShiftWAgg	Weights constructed from degree of extrimisation shift between rounds		Calculated internally	2	IDEA protocol or other structured protocol that generates multiple rounds of judgements of single point-estimates	$B_{i,c}$

Table 5: Summary of aggregation methods and functions, including data requirements and sources. *(continued)*

Method	Description	Data Requirements	Weighting Function	Elicitation Rounds	Elicitation Method	Data Sources
DistIntShiftWAgg	Weights constructed by degree of interval narrowing and shift towards certainty bounds between rounds		Calculated internally	2	IDEA protocol or other structured protocol that generates multiple rounds of judgements using three-point elicitation	$B_{i,c}$, $U_{i,c}$, $L_{i,c}$
ReasoningWAgg() Linearly-weighted best estimates, with weights constructed from supplementary reasoning data						
ReasonWAgg	Weighted by the breadth of reasoning (number of supplied reasons) provided to support the individuals' estimate	'data_supp_-ReasonWAgg'	'weight_reason()'	1	IDEA protocol or other structured protocol to elicit reasoning, but only sinle round (round 2) of data used in aggregation calculation.	$B_{i,c}$, $w\_reason_{i,c}$
ReasonWAgg2	Weighted by the breadth of reasoning provided to support the individuals' estimate, rescaled by breadth of reasoning across all claims	'data_supp_-ReasonWAgg'	'weight_-reason2()'	1	IDEA protocol or other structured protocol to elicit reasoning, but only sinle round (round 2) of data used in aggregation calculation.	$B_{i,c}$, $w\_reason_{i,c}$, $U_{i,d}$, $L_{i,d}$, $w\_reason_{i,c}$
ExtremisationWAgg() Takes the average of best-estimates and transforms it using the cumulative distribution function of a beta distribution						
BetaArMean	Beta-transformed arithmetic mean of the best-estimates		NA - Estimates are equally weighted	1	Single-point	$B_{i,c}$
BetaArMean2	Beta-transformed arithmetic mean of the best-estimates, but only to confidence scores outside a specified middle range.		NA - Estimates are equally weighted	1	Single-point	$B_{i,c}$
DistributionWAgg() Calculates the arithmetic mean of distributions created from expert judgements.						
DistribArMean	Applies a non-parametric distribution evenly across upper, lower and best-estimates		NA - Estimates are equally weighted	1	Three-point	$B_{i,c}$, $U_{i,c}$, $L_{i,c}$
TriDistribArMean	Applies a triangular distribution to the upper, lower and best-estimates		NA - Estimates are equally weighted	1	Three-point	$B_{i,c}$, $U_{i,c}$, $L_{i,c}$
BayesianWAgg() Bayesian aggregation methods with either uninformative or informative prior distributions						
BayTriVar	Bayesian tripple variability method		NA - Estimates are equally weighted	1	Three-point	$B_{i,c}$, $U_{i,c}$, $L_{i,c}$

Table 5: Summary of aggregation methods and functions, including data requirements and sources. *(continued)*

Method	Description	Data Requirements	Weighting Function	Elicitation Rounds	Elicitation Method	Data Sources
BayPRIORsAgg	As per ‘BayTriVar‘ but with priors derived from external predictive models, updated with individuals’ best-estimates	‘data_supp_-BayPRIORsAgg‘	NA - Estimates are equally weighted	1	Three-point	$B_{i,c}$, $U_{i,c}$, $L_{i,c}$

Listings

Listing 1 Multiple aggregation methods can be applied by binding rows rather than using the purrr package, if preferred.

```
purrr::map2_dfr(.x = list(AverageWAgg,
                          IntervalWAgg,
                          IntervalWAgg,
                          ShiftingWAgg,
                          BayesianWAgg),
                .y = list("ArMean",
                          "IndIntWAgg",
                          "IntWAgg",
                          "ShiftWAgg",
                          "BayTriVar"),
                .f = ~ .x(focal_claims,
                          type = .y,
                          percent_toggle = TRUE)
)
```

Listing 2 If we wish to batch aggregate claims using a combination of aggregation methods that do and do not require supplementary data, we must aggregate them separately, since the methods that require supplementary data have an additional argument for the supplementary data that must be parsed to the wrapper function call. We can chain the aggregation of the methods that do not require supplementary data, and the methods that do require supplementary data together very neatly using dplyr's bind_rows function (Wickham *et al.* 2021) and the magrittr pipe `%>%` (Bache and Wickham 2020). Below we implement this approach while applying the aggregation methods `ArMean`, `IntWAgg`, `IndIntWAgg`, `ShiftWAgg` and `BayTriVar` to the repliCATS pilot program dataset `data_ratings`.

```
confidenceSCOREs <-
  list(
    AverageWAgg,
    IntervalWAgg,
    IntervalWAgg,
    ShiftingWAgg,
    BayesianWAgg
  ) %>%
  purrr::map2_dfr(
    .y = list("ArMean",
              "IndIntWAgg",
              "IntWAgg",
              "ShiftWAgg",
              "BayTriVar"),
    .f = ~ .x(aggreCAT::data_ratings, type = .y, percent_toggle = TRUE)
  ) %>%
  dplyr::bind_rows(
    ReasoningWAgg(aggreCAT::data_ratings,
                  reasons = aggreCAT::data_supp_reasons,
                  percent_toggle = TRUE)
  )
```

Listing 3 Bring your own data: non-probablistic values

```
turtle_CS <-
  list(
  AverageWAgg,
  IntervalWAgg,
  IntervalWAgg,
  ShiftingWAgg
) %>%
  purrr::map2_dfr(.y = list("ArMean",
                            "IndIntWAgg",
                            "IntWAgg",
                            "ShiftWAgg"),
                  .f = ~ .x(green_turtles, type = .y,
                            percent_toggle = FALSE)
  )
```

Listing 4 Visualising Confidence Scores

```
plot_cs <-
  confidenceSCOREs %>%
  dplyr::left_join(aggreCAT::data_outcomes) %>%
  dplyr::mutate(data_type = "Confidence Scores") %>%
  dplyr::rename(x_vals = cs,
         y_vals = method) %>%
  dplyr::select(y_vals, paper_id, data_type, outcome, x_vals)

plot_judgements <-
  aggreCAT::preprocess_judgements(focal_claims,
                                  percent_toggle = TRUE) %>%
  tidyr::pivot_wider(names_from = element,
                     values_from = value) %>%
  dplyr::left_join(aggreCAT::data_outcomes) %>%
  dplyr::rename(x_vals = three_point_best,
         y_vals = user_name) %>%
  dplyr::select(paper_id,
         y_vals,
         x_vals,
         tidyr::contains("three_point"),
         outcome) %>%
  dplyr::mutate(data_type = "Elicited Probabilities")

p <- plot_judgements %>%
  dplyr::bind_rows(., {dplyr::semi_join(plot_cs, plot_judgements,
                           by = "paper_id")}) %>%
  ggplot2::ggplot(ggplot2::aes(x = x_vals, y = y_vals)) +
  ggplot2::geom_pointrange(ggplot2::aes(xmin = three_point_lower,
                       xmax = three_point_upper)) +
  ggplot2::facet_grid(data_type ~ paper_id, scales = "free_y") +
  ggplot2::theme_classic() +
  ggplot2::theme(legend.position = "none") +
  ggplot2::geom_vline(aes(xintercept = 0.5, colour = as.logical(outcome))) +
  ggplot2::xlab("Probability of Replication") +
  ggplot2::ylab(ggplot2::element_blank()) +
  ggplot2::scale_colour_brewer(palette = "Set1")
```

Computational details

The analyses and results in this paper were obtained using the following computing environment, versions of R and R packages:

```
R> devtools::session_info()
- Session info ---------------------------------------------------------------
 setting  value
 version  R version 4.2.2 (2022-10-31)
 os       macOS Monterey 12.6
 system   aarch64, darwin20
 ui       X11
 language (EN)
 collate  en_US.UTF-8
 ctype    en_US.UTF-8
 tz       Australia/Melbourne
 date     2023-03-31
 pandoc   2.19.2 @ /Applications/RStudio.app/Contents/Resources/app/quarto/bin/tools/ (v

- Packages -------------------------------------------------------------------
 package       * version    date (UTC) lib source
 abind           1.4-5      2016-07-21 [1] CRAN (R 4.2.0)
 aggreCAT      * 0.0.0.9004 2023-03-31 [1] local
 assertthat      0.2.1      2019-03-21 [1] CRAN (R 4.2.0)
 backports       1.4.1      2021-12-13 [1] CRAN (R 4.2.0)
 boot            1.3-28     2021-05-03 [1] CRAN (R 4.2.2)
 broom           1.0.1      2022-08-29 [1] CRAN (R 4.2.0)
 cachem          1.0.7      2023-02-24 [1] CRAN (R 4.2.2)
 callr           3.7.3      2022-11-02 [1] CRAN (R 4.2.0)
 car             3.1-1      2022-10-19 [1] CRAN (R 4.2.0)
 carData         3.0-5      2022-01-06 [1] CRAN (R 4.2.0)
 cellranger      1.1.0      2016-07-27 [1] CRAN (R 4.2.0)
 class           7.3-20     2022-01-16 [1] CRAN (R 4.2.2)
 cli             3.6.1      2023-03-23 [1] CRAN (R 4.2.0)
 coda            0.19-4     2020-09-30 [1] CRAN (R 4.2.0)
 colorspace      2.0-3      2022-02-21 [1] CRAN (R 4.2.0)
 cowplot         1.1.1      2020-12-30 [1] CRAN (R 4.2.0)
 crayon          1.5.2      2022-09-29 [1] CRAN (R 4.2.0)
 data.table      1.14.4     2022-10-17 [1] CRAN (R 4.2.0)
 DBI             1.1.3      2022-06-18 [1] CRAN (R 4.2.0)
 dbplyr          2.3.2      2023-03-21 [1] CRAN (R 4.2.0)
 DescTools       0.99.47    2022-10-22 [1] CRAN (R 4.2.0)
 devtools        2.4.5      2022-10-11 [1] CRAN (R 4.2.0)
 digest          0.6.31     2022-12-11 [1] CRAN (R 4.2.2)
 dplyr         * 1.1.1      2023-03-22 [1] CRAN (R 4.2.0)
 e1071           1.7-13     2023-02-01 [1] CRAN (R 4.2.0)
```

```
 ellipsis         0.3.2     2021-04-29 [1] CRAN (R 4.2.0)
 evaluate         0.18      2022-11-07 [1] CRAN (R 4.2.0)
 Exact            3.2       2022-09-25 [1] CRAN (R 4.2.0)
 expm             0.999-7   2023-01-09 [1] CRAN (R 4.2.0)
 fansi            1.0.3     2022-03-24 [1] CRAN (R 4.2.0)
 farver           2.1.1     2022-07-06 [1] CRAN (R 4.2.0)
 fastmap          1.1.1     2023-02-24 [1] CRAN (R 4.2.2)
 forcats        * 0.5.2     2022-08-19 [1] CRAN (R 4.2.0)
 fs               1.6.1     2023-02-06 [1] CRAN (R 4.2.0)
 gargle           1.2.1     2022-09-08 [1] CRAN (R 4.2.0)
 generics         0.1.3     2022-07-05 [1] CRAN (R 4.2.0)
 ggforce        * 0.4.1     2022-10-04 [1] CRAN (R 4.2.0)
 ggplot2        * 3.4.0     2022-11-04 [1] CRAN (R 4.2.0)
 ggpubr         * 0.6.0     2023-02-10 [1] CRAN (R 4.2.0)
 ggridges       * 0.5.4     2022-09-26 [1] CRAN (R 4.2.0)
 ggsignif         0.6.4     2022-10-13 [1] CRAN (R 4.2.0)
 gld              2.6.6     2022-10-23 [1] CRAN (R 4.2.0)
 glue             1.6.2     2022-02-24 [1] CRAN (R 4.2.0)
 googledrive      2.0.0     2021-07-08 [1] CRAN (R 4.2.0)
 googlesheets4    1.0.1     2022-08-13 [1] CRAN (R 4.2.0)
 gridExtra        2.3       2017-09-09 [1] CRAN (R 4.2.0)
 gt               0.8.0     2022-11-16 [1] CRAN (R 4.2.0)
 gtable           0.3.1     2022-09-01 [1] CRAN (R 4.2.0)
 haven            2.5.1     2022-08-22 [1] CRAN (R 4.2.0)
 hms              1.1.2     2022-08-19 [1] CRAN (R 4.2.0)
 htmltools        0.5.5     2023-03-23 [1] CRAN (R 4.2.2)
 htmlwidgets      1.6.1     2023-01-07 [1] CRAN (R 4.2.0)
 httpuv           1.6.9     2023-02-14 [1] CRAN (R 4.2.2)
 httr             1.4.5     2023-02-24 [1] CRAN (R 4.2.2)
 insight          0.19.0    2023-01-30 [1] CRAN (R 4.2.0)
 jsonlite         1.8.4     2022-12-06 [1] CRAN (R 4.2.2)
 kableExtra     * 1.3.4     2021-02-20 [1] CRAN (R 4.2.0)
 knitr          * 1.42      2023-01-25 [1] CRAN (R 4.2.0)
 labeling         0.4.2     2020-10-20 [1] CRAN (R 4.2.0)
 later            1.3.0     2021-08-18 [1] CRAN (R 4.2.0)
 lattice          0.20-45   2021-09-22 [1] CRAN (R 4.2.2)
 lifecycle        1.0.3     2022-10-07 [1] CRAN (R 4.2.0)
 lmom             2.9       2022-05-29 [1] CRAN (R 4.2.0)
 lubridate        1.9.0     2022-11-06 [1] CRAN (R 4.2.0)
 magrittr         2.0.3     2022-03-30 [1] CRAN (R 4.2.0)
 MASS             7.3-58.1  2022-08-03 [1] CRAN (R 4.2.2)
 Matrix           1.5-1     2022-09-13 [1] CRAN (R 4.2.2)
 memoise          2.0.1     2021-11-26 [1] CRAN (R 4.2.0)
 mime             0.12      2021-09-28 [1] CRAN (R 4.2.0)
 miniUI           0.1.1.1   2018-05-18 [1] CRAN (R 4.2.0)
```

```
 modelr          0.1.10    2022-11-11 [1] CRAN (R 4.2.0)
 munsell         0.5.0     2018-06-12 [1] CRAN (R 4.2.0)
 mvtnorm         1.1-3     2021-10-08 [1] CRAN (R 4.2.0)
 pillar          1.8.1     2022-08-19 [1] CRAN (R 4.2.0)
 pkgbuild        1.4.0     2022-11-27 [1] CRAN (R 4.2.0)
 pkgconfig       2.0.3     2019-09-22 [1] CRAN (R 4.2.0)
 pkgload         1.3.2     2022-11-16 [1] CRAN (R 4.2.0)
 png             0.1-8     2022-11-29 [1] CRAN (R 4.2.0)
 polyclip        1.10-4    2022-10-20 [1] CRAN (R 4.2.0)
 precrec         0.14.1    2023-01-08 [1] CRAN (R 4.2.0)
 prettyunits     1.1.1     2020-01-24 [1] CRAN (R 4.2.0)
 processx        3.8.0     2022-10-26 [1] CRAN (R 4.2.0)
 profvis         0.3.7     2020-11-02 [1] CRAN (R 4.2.0)
 promises        1.2.0.1   2021-02-11 [1] CRAN (R 4.2.0)
 proxy           0.4-27    2022-06-09 [1] CRAN (R 4.2.0)
 ps              1.7.2     2022-10-26 [1] CRAN (R 4.2.0)
 purrr         * 1.0.1     2023-01-10 [1] CRAN (R 4.2.0)
 R2jags          0.7-1     2021-08-05 [1] CRAN (R 4.2.0)
 R2WinBUGS       2.1-21    2015-07-30 [1] CRAN (R 4.2.0)
 R6              2.5.1     2021-08-19 [1] CRAN (R 4.2.0)
 RColorBrewer    1.1-3     2022-04-03 [1] CRAN (R 4.2.0)
 Rcpp            1.0.10    2023-01-22 [1] CRAN (R 4.2.2)
 readr         * 2.1.3     2022-10-01 [1] CRAN (R 4.2.0)
 readxl          1.4.1     2022-08-17 [1] CRAN (R 4.2.0)
 remotes         2.4.2     2021-11-30 [1] CRAN (R 4.2.0)
 reprex          2.0.2     2022-08-17 [1] CRAN (R 4.2.0)
 rfUtilities     2.1-5     2019-10-03 [1] CRAN (R 4.2.0)
 rjags           4-13      2022-04-19 [1] CRAN (R 4.2.0)
 rlang           1.1.0     2023-03-14 [1] CRAN (R 4.2.0)
 rmarkdown       2.18      2022-11-09 [1] CRAN (R 4.2.0)
 rootSolve       1.8.2.3   2021-09-29 [1] CRAN (R 4.2.0)
 rstatix         0.7.2     2023-02-01 [1] CRAN (R 4.2.0)
 rstudioapi      0.14      2022-08-22 [1] CRAN (R 4.2.0)
 rvest           1.0.3     2022-08-19 [1] CRAN (R 4.2.0)
 scales          1.2.1     2022-08-20 [1] CRAN (R 4.2.0)
 sessioninfo     1.2.2     2021-12-06 [1] CRAN (R 4.2.0)
 shiny           1.7.4     2022-12-15 [1] CRAN (R 4.2.0)
 stringi         1.7.8     2022-07-11 [1] CRAN (R 4.2.0)
 stringr       * 1.5.0     2022-12-02 [1] CRAN (R 4.2.0)
 svglite         2.1.1     2023-01-10 [1] CRAN (R 4.2.0)
 systemfonts     1.0.4     2022-02-11 [1] CRAN (R 4.2.0)
 tibble        * 3.2.1     2023-03-20 [1] CRAN (R 4.2.0)
 tidyr         * 1.3.0     2023-01-24 [1] CRAN (R 4.2.0)
 tidyselect      1.2.0     2022-10-10 [1] CRAN (R 4.2.0)
 tidyverse     * 1.3.2     2022-07-18 [1] CRAN (R 4.2.0)
```

```
 timechange       0.1.1      2022-11-04 [1] CRAN (R 4.2.0)
 tinytex        * 0.42       2022-09-27 [1] CRAN (R 4.2.0)
 tweenr           2.0.2      2022-09-06 [1] CRAN (R 4.2.0)
 tzdb             0.3.0      2022-03-28 [1] CRAN (R 4.2.0)
 urlchecker       1.0.1      2021-11-30 [1] CRAN (R 4.2.0)
 usethis          2.1.6      2022-05-25 [1] CRAN (R 4.2.0)
 utf8             1.2.2      2021-07-24 [1] CRAN (R 4.2.0)
 vctrs            0.6.1      2023-03-22 [1] CRAN (R 4.2.0)
 viridisLite      0.4.1      2022-08-22 [1] CRAN (R 4.2.0)
 webshot          0.5.4      2022-09-26 [1] CRAN (R 4.2.0)
 withr            2.5.0      2022-03-03 [1] CRAN (R 4.2.0)
 xfun             0.34       2022-10-18 [1] CRAN (R 4.2.0)
 xml2             1.3.3      2021-11-30 [1] CRAN (R 4.2.0)
 xtable           1.8-4      2019-04-21 [1] CRAN (R 4.2.0)
 yaml             2.3.7      2023-01-23 [1] CRAN (R 4.2.0)

 [1] /Library/Frameworks/R.framework/Versions/4.2-arm64/Resources/library

------------------------------------------------------------------------------
```

Acknowledgments

This project is sponsored by the Defense Advanced Research Projects Agency (DARPA) under cooperative agreement No.HR001118S0047. The content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

(2015). "Estimating the reproducibility of psychological science." *Science*, **349**(6251), aac4716. doi:10.1126/science.aac4716.

Alipourfard N, Arendt B, Benjamin DM, Benkler N, Bishop MM, Burstein M, Bush M, Caverlee J, Chen Y, Clark C, Dreber A, Errington TM, Fidler F, Fox N, Frank A, Fraser H, Friedman S, Gelman B, Gentile J, Gordon M, Griffin C, Gulden T, Hahn K, Hartman R, Holzmeister F, Hu X, Johannesson M, Kezar L, Kline Struhl M, Kuter U, Kwasnica A, Lee D, Lerman K, Liu Y, Loomas Z, Luis B, Magnusson I, Bishop M, Miske O, Mody F, Morstatter F, Nosek BA, Parsons S, Pennock D, Pi H, Pujara J, Rajtmajer S, Ren X, Salinas A, Selvam R, Shipman F, Silverstein P, Sprenger A, Squicciarini A, Stratman S, Sun K, Tikoo S, Twardy CR, Tyner A, Viganola D, Wang J, Wilkinson D, Wintle B (2021). "Systematizing Confidence in Open Research and Evidence (SCORE)." doi:10.31235/osf.io/46mnb.

Arlidge WN, Alfaro-Shigueto J, Ibanez-Erquiaga B, Mangel JC, Squires D, Milner-Gulland

EJ (2020). "Evaluating elicited judgments of turtle captures for data-limited fisheries management." *Conservation Science and Practice*, **2**(5).

Bache SM, Wickham H (2020). *magrittr: A Forward-Pipe Operator for R*. R package version 2.0.1, URL https://CRAN.R-project.org/package=magrittr.

Bryan J, Hester J, Robinson D, Wickham H (2021). *reprex: Prepare Reproducible Example Code via the Clipboard*. R package version 2.0.0, URL https://CRAN.R-project.org/package=reprex.

Camerer C, Dreber A, Holzmeister F, Ho T, Huber J, Johannesson M, Kirchler M, Nave G, Nosek BA, Pfeiffer T, Altmejd A, Buttrick N, Chan T, Chen Y, Forsell E, Gampa A, Heikensten E, Hummer L, Taisuke I, Isaksson S, Manfredi D, Rose J, Wagenmakers E, Wu H (2018). "Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015." *naturecom*.

Ebersole CR, Atherton OE, Belanger AL, Skulborstad HM, Allen JM, Banks JB, Baranski E, Bernstein MJ, Bonfiglio DB, Boucher L, Brown ER, Budiman NI, Cairo AH, Capaldi CA, Chartier CR, Chung JM, Cicero DC, Coleman JA, Conway JG, Davis WE, Devos T, Fletcher MM, German K, Grahe JE, Hermann AD, Hicks JA, Honeycutt N, Humphrey B, Janus M, Johnson DJ, Joy-Gaba JA, Juzeler H, Keres A, Kinney D, Kirshenbaum J, Klein RA, Lucas RE, Lustgraaf CJ, Martin D, Menon M, Metzger M, Moloney JM, Morse PJ, Prislin R, Razza T, Re DE, Rule NO, Sacco DF, Sauerberger K, Shrider E, Shultz M, Siemsen C, Sobocko K, Weylin Sternglanz R, Summerville A, Tskhay KO, van Allen Z, Vaughn LA, Walker RJ, Weinberg A, Wilson JP, Wirth JH, Wortman J, Nosek BA (2016). "Many Labs 3: Evaluating participant pool quality across the academic semester via replication." *Journal of Experimental Social Psychology*, **67**, 68–82. ISSN 0022-1031. `doi:https://doi.org/10.1016/j.jesp.2015.10.012`. Special Issue: Confirmatory, URL https://www.sciencedirect.com/science/article/pii/S0022103115300123.

Fraser H, Bush M, Wintle B, Mody F, Smith ET, Hanea A, Gould E, Hemming V, Hamilton DG, Rumpff L, Wilkinson D, Pearson R, Singleton Thorn F, Ashton R, Willcox A, Gray C, Head A, Ross M, Groenewegen R, Marcoci A, Vercammen A, Parker T, Hoekstra R, Nakagawa S, Mandel D, van Ravenzwaaij D, McBride M, Sinnot R, Vesk P, Burgman M, Fidler F (2021). "Predicting reliability through structured expert elicitation with repliCATS (Collaborative Assessments for Trustworthy Science)." `doi:10.31222/osf.io/2pczv`.

Goossens L, Cooke R, Hale A, Rodic-Wiersma L (2008). "Fifteen years of expert judgement at TUDelft." *Safety Science*, **46**(2), 234–244.

Gordon M, Viganola D, Bishop M, Chen Y, Dreber A, Goldfedder B, Holzmeister F, Johannesson M, Liu Y, Twardy C, Wang J, Pfeiffer T (2020). "Are replication rates the same across academic fields? Community forecasts from the DARPA SCORE programme." *Royal Society Open Science*, **7**(7), 200566. `doi:10.1098/rsos.200566`.

Gould E, Willcox A, Fraser H, Singleton Thorn F, Wilkinson DP (2021). "Using model-based predictions to inform the mathematical aggregation of human-based predictions of replicability." `doi:10.31222/osf.io/f675q`.

Hanea A, Wilkinson DP, McBride M, Lyon A, van Ravenzwaaij D, Singleton Thorn F, Gray CT, Mandel DR, Willcox A, Gould E, et al (2021). "Mathematically aggregating experts' predictions of possible futures." *PLoS ONE*, **16**(9). doi:https://doi.org/10.1371/journal.pone.0256919.

Hemming V, Burgman M, Hanea A, McBride M, Wintle B (2017). "A practical guide to structured expert elicitation using the IDEA protocol." *Methods in Ecology and Evolution*, **9**(1), 169–180. doi:10.1111/2041-210x.12857. URL http://dx.doi.org/10.1111/2041-210x.12857.

Henry L, Wickham H (2020). *purrr: Functional Programming Tools*. R package version 0.3.4, URL https://CRAN.R-project.org/package=purrr.

Isager PM, van Aert R, Bahnik S, Brandt M, Desoto K, Ginner-Sorolla R, Krueger J, Perugini M, Ropovik I, van't Veer A, Vranka M, Lakens D (2020). "Deciding what to replicate: A formal definition of "replication value" and a decision model for replication study selection." *Journal of Informetrics*, **13**(2), 635–642. doi:https://doi.org/10.1037/met0000438.

Klein RA, Ratliff KA, Vianello M, Adams Jr RB, Bahnic S, Bernstein MJ, Bocian K, Brandt M, Brooks B, Brumbaugh CC, Cermalcilar Z, Chandler J, Cheong W, Davis WE, Devos T, Eisner M, Frankowska N, Furrow D, Galiani EM, Hasselman F, Hicks JA, Hovermale J, Hunt S, Huntsinger JR, IJzerman H, John MS, Joy-Gaba JA, Barry Kappes H, Kreuger LE, Kurtz J, Levitan CA, Mallet RK, Morris WL, Nelson AJ, Nier JA, Packard G, Pilati R, Rutchick AM, Schmidt K, Skorinko JL, Smith R, Steiner TG, Storbeck J, Van Swol LM, Thompson D, van 't Veer A, Vaughn LA, Vranka M, Wichman AL, Woodzicka JA, Nosek BA (2014). "Investigating Variation in Replicability." *Social Psychology*, **45**(3), 142–152.

Klein RA, Vianello M, Hasselman F, Adams BG, Reginald B Adams J, Alper S, Aveyard M, Axt JR, Babalola MT, Štěpán Bahník, Batra R, Berkics M, Bernstein MJ, Berry DR, Bialobrzeska O, Binan ED, Bocian K, Brandt MJ, Busching R, Rédei AC, Cai H, Cambier F, Cantarero K, Carmichael CL, Ceric F, Chandler J, Chang JH, Chatard A, Chen EE, Cheong W, Cicero DC, Coen S, Coleman JA, Collisson B, Conway MA, Corker KS, Curran PG, Cushman F, Dagona ZK, Dalgar I, Rosa AD, Davis WE, de Bruijn M, Schutter LD, Devos T, de Vries M, Doğulu C, Dozo N, Dukes KN, Dunham Y, Durrheim K, Ebersole CR, Edlund JE, Eller A, English AS, Finck C, Frankowska N, Ángel Freyre M, Friedman M, Galliani EM, Gandi JC, Ghoshal T, Giessner SR, Gill T, Gnambs T, Ángel Gómez, González R, Graham J, Grahe JE, Grahek I, Green EGT, Hai K, Haigh M, Haines EL, Hall MP, Heffernan ME, Hicks JA, Houdek P, Huntsinger JR, Huynh HP, IJzerman H, Inbar Y, Åse H Innes-Ker, Jiménez-Leal W, John MS, Joy-Gaba JA, Kamiloğlu RG, Kappes HB, Karabati S, Karick H, Keller VN, Kende A, Kervyn N, Knežević G, Kovacs C, Krueger LE, Kurapov G, Kurtz J, Lakens D, Lazarević LB, Levitan CA, Neil A Lewis J, Lins S, Lipsey NP, Losee JE, Maassen E, Maitner AT, Malingumu W, Mallett RK, Marotta SA, Međedović J, Mena-Pacheco F, Milfont TL, Morris WL, Murphy SC, Myachykov A, Neave N, Neijenhuijs K, Nelson AJ, Neto F, Nichols AL, Ocampo A, O'Donnell SL, Oikawa H, Oikawa M, Ong E, Orosz G, Osowiecka M, Packard G, Pérez-Sánchez R, Petrović B, Pilati R, Pinter B, Podesta L, Pogge G, Pollmann MMH, Rutchick AM, Saavedra P, Saeri AK, Salomon E, Schmidt K, Schönbrodt FD, Sekerdej MB, Sirlopú D, Skorinko JLM, Smith MA, Smith-Castro V, Smolders KCHJ, Sobkow A, Sowden W, Spachtholz P, Srivastava M, Steiner TG, Stouten J, Street CNH, Sundfelt OK, Szeto S, Szumowska E, Tang ACW, Tanzer N, Tear MJ, Theriault J, Thomae M, Torres D, Traczyk J, Tybur JM, Ujhelyi A, van Aert RCM, van Assen MALM, van der Hulst M, van Lange PAM, van 't Veer AE, Vásquez-Echeverría A, Vaughn LA, Vázquez A, Vega LD, Verniers C, Verschoor M, Voermans IPJ, Vranka MA, Welch C, Wichman AL, Williams LA, Wood M, Woodzicka JA, Wronska MK, Young L, Zelenski JM, Zhijia Z, Nosek BA (2018). "Many Labs 2: Investigating Variation in Replicability Across Samples and Settings." *Advances in Methods and Practices in Psychological Science*, **1**(4), 443–490. doi:10.1177/2515245918810225.

Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H (2020). *vegan: Community Ecology Package*. R package version 2.5-7, URL https://CRAN.R-project.org/package=vegan.

Pearson R, Fraser H, Bush M, Mody F, Widjaja I, Head A, Wilkinson DP, Sinnott R, Wintle B, Burgman M, Fidler F, Vesk P (2021). "Eliciting group judgements about replicability: a technical implementation of the IDEA Protocol." URL http://hdl.handle.net/10125/70666.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Su YS, Yajima M (2020). *R2jags: Using R to Run 'JAGS'*. R package version 0.6-1, URL https://CRAN.R-project.org/package=R2jags.

Sutherland WJ, Dicks LV, Everard M, Geneletti D (2018). "Qualitative methods for ecologists and conservation scientists." *Methods in Ecology and Evolution*, **9**(1), 7–9. doi:https://doi.org/10.1111/2041-210X.12956.

Wickham H (2014). "Tidy data." *Journal of Statistical Software*, **59**(10).

Wickham H (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN 978-3-319-24277-4. URL https://ggplot2.tidyverse.org.

Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). "Welcome to the tidyverse." *Journal of Open Source Software*, **4**(43), 1686. doi:10.21105/joss.01686.

Wickham H, François R, Henry L, Müller K (2021). *dplyr: A Grammar of Data Manipulation*. R package version 1.0.6, URL https://CRAN.R-project.org/package=dplyr.

Wickham H, Grolemund G (2017). *R for Data Science: Import, Tidy, Transform, Visualize, and Model Data*. 1 edition. O'Reilly Media. ISBN 1491910399. URL http://r4ds.had.co.nz/.

Wilke CO (2021). *ggridges: Ridgeline Plots in "ggplot2"*. R package version 0.5.3, URL https://CRAN.R-project.org/package=ggridges.

Wintle B, Mody F, Smith ET, Hanea A, Wilkinson DP, Hemming V, Bush M, Fraser H, Singleton Thorn F, McBride M, Gould E, Head A, Hamilton D, Rumpff L, Hoekstra R, Fidler F (2021). "Predicting and reasoning about replicability using structured groups." doi:10.31222/osf.io/vtpmb.

Yenni GM, Christensen EM, Bledsoe EK, Supp SR, Diaz RM, White EP, Ernest SKM (2019). "Developing a modern data workflow for regularly updated data." *PLOS Biology*, **17**(1), 1–12. doi:10.1371/journal.pbio.3000125. URL https://doi.org/10.1371/journal.pbio.3000125.

Affiliation:

Elliot Gould[3]
E-mail: elliot.gould (at) unimelb.edu.au

Charles T. Gray

Aaron Willcox

Rose O'Dea

Rebecca Groenewegen

David P. Wilkinson

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd

[3]School of Ecosystem and Forest Sciences, University of Melbourne