THEME ENV.2012.6.1-1

EUPORIAS

(Grant Agreement 308291)

European Provision Of Regional Impact Assessment on a Seasonal-to-decadal timescale

Deliverable D*33.4*

Decision lab with relevant stakeholders

Deliverable Title	*Communicating levels of confidence and uncertainty*	
Brief Description	*This task will test the salience, effectiveness and usability of strategies for communicating confidence levels amongst* *EUPORIAS stakeholders in a decision lab.*	
WP number	33	
Lead Beneficiary	*Andrea Taylor, University of Leeds* *Suraje Dessai, University of Leeds*	
Contributors	*Andrea Taylor, University of Leeds* *Suraje Dessai, University of Leeds* *Carlo Buontempo, Met Office* *Mike Butts, DHI* *Laurent Dubus, EDF* *Ghislain Dubois, TEC* *Eroteida Sánchez, AEMET* *Christian Viel, Meteo France* *Jose Voces, AEMET*	
Creation Date	10/12/2015	
Version Number	2	
Version Date	30/12/2015	
Deliverable Due Date	31/12/2015	
Actual Delivery Date		
Nature of the Deliverable	X	*R – Report*
		P - Prototype
		D - Demonstrator
		O - Other
Dissemination Level/ Audience	X	*PU - Public*
		PP - Restricted to other programme participants, including the Commission services
		RE - Restricted to a group specified by the consortium, including the Commission services
		CO - Confidential, only for members of the consortium, including the Commission services

Version	Date	Modified by	Comments
1	10/12/2015	Andrea Taylor	
2	30/12/2015	Andrea Taylor	

Table of Contents

Contents

List of Tables

List of Figures

1. Executive Summary

The objective of this task was to test strategies for communicating levels of confidence and uncertainty in seasonal to decadal (S2D) climate predictions with relevant stakeholders. To do this we ran two online Decision Labs: a) an 'Abridged' Decision Lab with a general sample of European decision makers in relevant sectors, and b) a 'Full' Decision Lab with a group of participants who were already highly engaged with climate information.

After selecting a sub-sample of six of the communication strategies developed in Task 33.3, we presented 'Higher Skill' and 'Lower Skill' versions to Decision Lab participants. Four of these were shown to participants who indicated that they were experienced users of statistical information (Bubble Map, Violin Plot, Bar Graph and Table), while three were shown to participants who indicated that they were novice users of statistical information (Bar Graph, Confidence Index and Simple Table). In the Abridged Lab each participant was shown only one communication strategy, while in the Full Lab participants were presented with all those appropriate for their level of statistical expertise. For each type of communication strategy presented to them, participants were asked a series of questions to designed to ascertain: 1) their objective understanding of the information presented; 2) their subjective interpretation of what the predictions indicated; 3) how useful they perceived the predictions to be; 4) their overall preference or 'liking' of the communication strategies; and 5) how familiar the communication strategies were.

With respect to participants' objective understanding of the communication strategies, the findings of the two Decision Labs suggest that, for both experienced and novice users of statistical information, numeric representations of likelihood presented in tabular format tend to be the easiest to interpret when the information of interest is the likelihood of a particular tercile. However, they are less useful when other aspects of climate information are of more interest to users (e.g. ranges of values, spatial information). With regard to information about skill, the results of the two studies highlight areas where systematic misinterpretations of skill scores can take place. These include: skill scores being mistaken for tercile likelihoods, and failure to recognise that negative scores indicate that there is no skill. Hence, we recommend a) that care be taken not to place skill scores in areas of graphs and tables where they may be mistaken for likelihoods; and b) that where negative skill exists providers should take steps to emphasise this (e.g. with 'No Skill' warnings, or by presenting climatology only). Where users do not require precise numeric information about skill, providing descriptions in the form of evaluative categories (e.g. 'Good Skill', 'Some Skill', 'No Skill') may also help to avoid the conflation of likelihood with skill.

On examining participants' subjective interpretation of the communication strategies, we found that both experienced and novice users of statistical information tended to perceive Lower Skill predictions as being less useful than those with Higher Skill, and reported lower confidence in their subjective estimates of tercile likelihood when shown the Lower Skill predictions. However, we also find that, even for those communication strategies where objective understanding of information about skill was relatively high, subjective judgements of tercile likelihood were unduly influenced by predictions where no skill existed. Once again, this suggests that where skill does not exist providers should consider showing climatology only, or providing explicit warnings that climatology currently provides a better guide to future conditions.

With regard to participants' preference for different formats, we found that experienced users of statistical information did not, as a group, demonstrate a clear preference for one style of communication strategy over another. Novice users in the Abridged Lab did however demonstrate a preference for the Simple Table over the Confidence Index and Bar Graph. In keeping with earlier research we found that while preference tended to correspond with perceived familiarity (Taylor and Dessai, 2014; Taylor, Dessai & Bruine de Bruin, 2015a), it was not directly related to objective understanding (Lorenz, Dessai, Forster & Paavola, 2015). However, in the Abridged Lab, we did find some evidence to suggest that while preference is linked to perceived familiarity, perceived familiarity may be detrimental to understanding when similarities between communication strategies are superficial only.

Taken together the findings of both Decision Labs emphasise the need for climate service providers to give clear, explicit guidance as to how the communications that they provide should (and should not) be interpreted, taking into account the misconceptions that may arise.

2. Project Objectives

With this deliverable, the project has contributed to the achievement of the following objectives (DOW, Section B1.1):

No.	Objective	Yes	No
1	Develop and deliver reliable and trusted impact prediction systems for a number of carefully selected case studies. These will provide working examples of end to end climate-to-impacts-decision making services operation on S2D timescales.		X
2	Assess and document key knowledge gaps and vulnerabilities of important sectors (e.g., water, energy, health, transport, agriculture, tourism), along with the needs of specific users within these sectors, through close collaboration with project stakeholders.		X
3	Develop a set of standard tools tailored to the needs of stakeholders for calibrating, downscaling, and modelling sector-specific impacts on S2D timescales.		X
4	Develop techniques to map the meteorological variables from the prediction systems provided by the WMO GPCs (two of which (Met Office and MeteoFrance) are partners in the project) into variables which are directly relevant to the needs of specific stakeholders.		X
5	Develop a knowledge-sharing protocol necessary to promote the use of these technologies. This will include making uncertain information fit into the decision support systems used by stakeholders to take decisions on the S2D horizon. This objective will place Europe at the forefront of the implementation of the GFCS, through the GFCS's ambitions to develop climate services research, a climate services information system and a user interface platform.	X	
6	Assess and document the current marketability of climate services in Europe and demonstrate how climate services on S2D time horizons can be made useful to end users.		X

3. Detailed Report

3.1 Introduction

This report details the findings of two online Decision Labs undertaken to test a set of strategies for communicating levels of confidence and uncertainty in seasonal-to-decadal (S2D) climate predictions. The first, undertaken with a sample of European decision makers from public and private organisations, offers an exploration of comprehension and preference for different communication strategies, amongst a general sample of those working in climate and weather sensitive sectors. The second is a longer study undertaken with EUPORIAS stakeholders and participants invited to take part through the Climate Service Partnership mailing list, which explores the opinions and objective understanding of a group of highly engaged current and potential users of S2D predictions.

The objective of Work Package 33 of the EUPORIAS project is to develop strategies for communicating levels of confidence and uncertainty in S2D predictions. In the first phase of this work (Task 33.1) an assessment of user needs with respect to the communication of confidence and uncertainty in S2D predictions found that a) current users of S2D predictions perceived this information to be useful but comparatively difficult to access and understand, b) many users were not receiving information about forecast performance (e.g. skill, reliability) in a way that was easily interpreted as such, and c) preferences for certain communication strategies depended on one's comfort with using statistical information and existing familiarity with the format (see Taylor and Dessai, 2014; Taylor, Dessai & Bruine de Bruin, 2015a). This was supported by a subsequent review of existing approaches to communicating confidence and uncertainty (Task 33.2), which identified factors that can affect decision makers' understanding and utilisation of uncertain information, including: experience of using statistical information (e.g. Peters, 2008), elements of visual design (e.g. whether the use of colours is 'intuitive') (e.g. Kaye, Hartley & Hemming, 2012), tolerance for uncertainty (e.g. Ellsberg, 1990), desire for communications and tools that directly facilitate Act/Don't Act decisions (e.g. (McCown, 2012; McCown, Carberry, Dalgliesh, Foale, & Hochman, 2012). It also revealed that while many innovative ways of presenting information about likelihood and skill in seasonal climate predictions had been developed there had until that point been comparatively little empirical testing of comprehension and preference conducted with user groups (see Taylor, Dessai, Buontempo & Dubois, 2014 for full review).

Drawing on the key findings of these preceding tasks, a set of strategies for communicating levels of confidence in S2D predictions were developed for both advanced and novice users of statistical information (Task 33.3) (see the report on this task in Taylor et al., 2015b for full details of strategy development). Having developed these strategies to address the user preferences and difficulties identified in Tasks 33.1 and 33.2, the Decision Labs reported here (Task 33.4) were undertaken to assess 1) objective understanding of decision makers in relevant sectors (i.e. whether participants' interpretations of the communication strategies were consistent with what the communications are intended to convey); 2) participants' confidence in their interpretations of the predictions; 3) perceived usefulness of the communication strategies; and 4) participants' levels of preference for the communication strategies; and 5) the extent to which preference corresponds with objective understanding.

This report therefore proceeds as follows. In Section 3.2 we describe the communication strategies selected for inclusion in the two Decision Labs. This is followed by full reports on an Abridged Decision Lab conducted with a large sample of European decision makers in relevant sectors (Section 3.3), and a Full Decision Lab conducted with a core sample of stakeholders who are already highly engaged with climate information (Sections 3.4 and 3.5). It should be noted that while our analysis of the quantitative data obtained in these Decision Labs is reported in some detail for each set of research questions, in all cases 'non-technical' summaries are provided at the end of each relevant subsection. Finally, we discuss the overall findings of this task (Section 3.6) and summarise the key conclusions (Section 3.7)

Summary of Decision Lab Objectives

- To present a selection of strategies for communicating levels of confidence and uncertainty in S2D predictions to a) a large sample of European decision makers from relevant sectors (Study 1); and b) a smaller more specialist sample of highly engaged stakeholders.
- To examine how well the different formats are objectively understood and identify where misinterpretations occur.
- To examine how participants' subjectively interpret Higher Skill and Lower Skill predictions, and how useful they perceive these to be.
- To examine which of the communication strategies participants prefer.
- To assess whether greater objective understanding corresponds with stronger preference for particular communication strategies, and greater existing familiarity with them.

3.2 Selection of communication strategies for inclusion

The six different communication strategies presented in the Decision Labs were drawn from the selection of strategies formulated in Task 33.3 of this programme research. Whilst a wider range of visualisations and text-based communication formats were developed, time and resource constraints meant that it would not be feasible to systematically assess organisational responses to all of them. Hence, a subset of six different styles was chosen for inclusion in the Decision Labs. Selection was influenced by three main factors 1) the preferences expressed by participant in an earlier user needs survey (Task 33.1: See Taylor and Dessai, 2014; Taylor et al., 2015b); 2) a need to include formats suitable for those with less experience of using statistical information as well as formats suitable for those with existing expertise in this area; and 3) a wish to include formats that had a spatial or temporal element as well as those that consisted of a single tercile representation. We also sought to include a mixture of novel and more familiar formats. Of the six communication strategies chosen three were specifically intended for experienced users of statistical information (Bubble Map, Violin Plot, Table), while two were specifically intended for novice users of statistical information (Confidence Index, Simple Table). One visualisation, the Bar Graph, was presented to both advanced and novice users.

3.2.1 Communication strategies for advanced users of statistical information

Bubble Map

In our earlier user needs survey (Task 33.1), maps emerged as one of the communication strategies that were most highly favoured by participants. We therefore felt that it was important to include at least one map in this set of communication strategies. Based on prior work by Slingsby et al. (2009) and Jupp et al. (2012), Bubble Maps were identified as one way of combining likelihood and information about skill on the same map. In the development phase of this work package discussed in Taylor et al (2015b), several different iterations of the Bubble Map were created, each varying in complexity. This particular version was chosen as it featured both information about the likelihood of the most probable tercile (represented by the size and colour of the bubble) and information about skill (ROCSS for most likely tercile represented by the shading of each bubble) (see Figure 3.2.1).

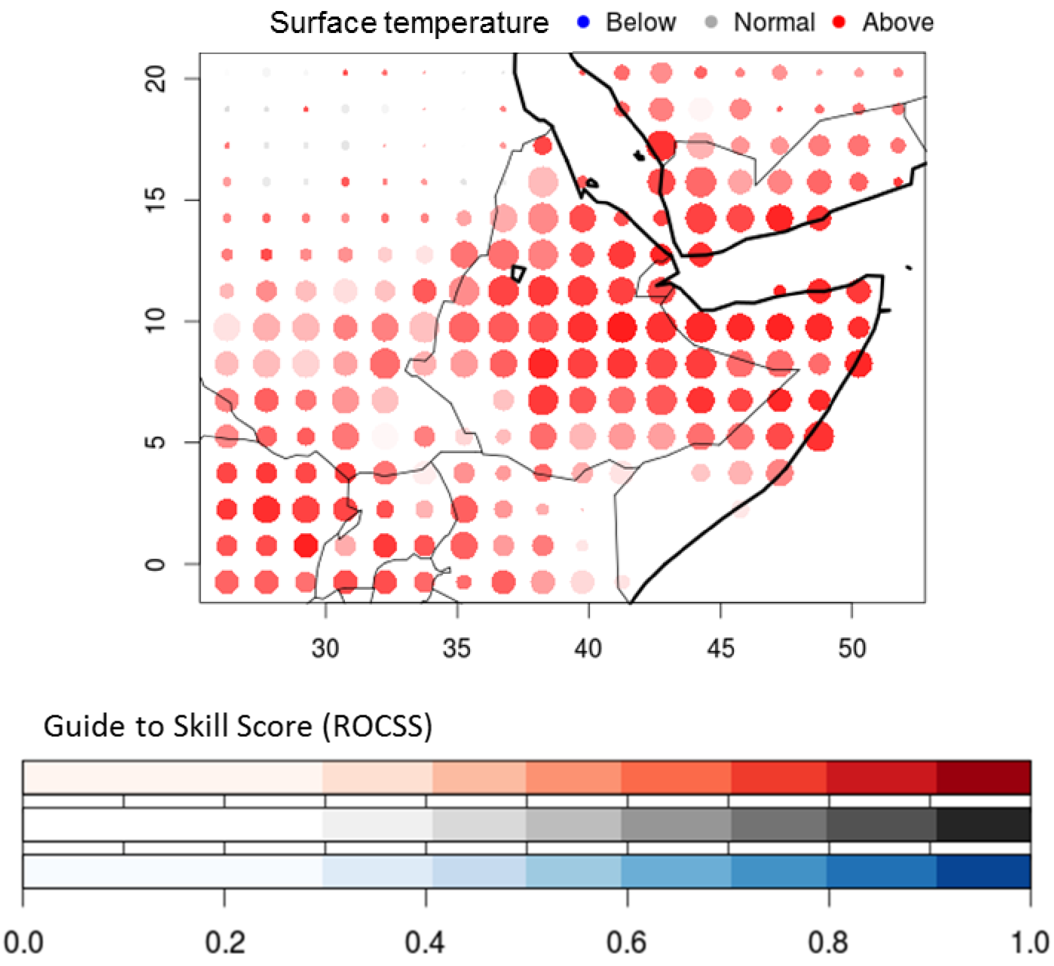


Figure 3.2.1 Bubble Map
On the map above, coloured bubbles show whether the forecast predicts that warmer than average (red), average (grey), or colder than average (blue) temperatures are more likely for each area. The size of the bubbles shows the predicted likelihood of the most likely category according to the forecast. Larger bubbles show greater likelihood, smaller bubbles show lower likelihood. A skill score is given for each bubble using ROCSS. A score above 0.5 shows good skill. A score between 0 and 0.5 shows some skill. A score below 0 shows no skill. On the map darker shades show greater skill, lighter shades show lower skill. Blank spaces show areas where there is no skill.

Violin Plot

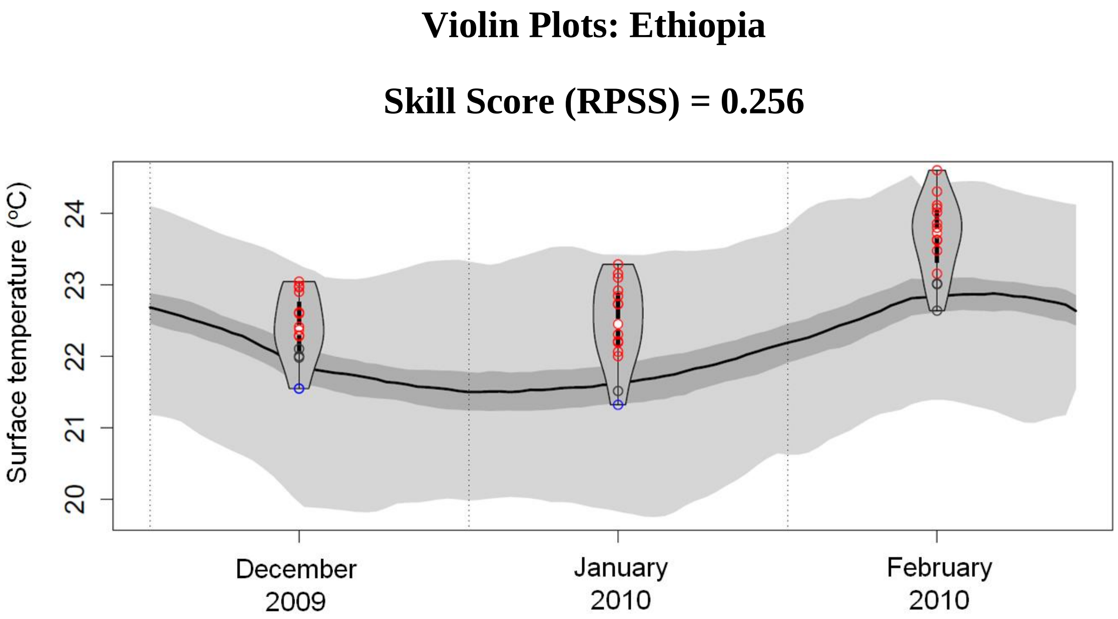
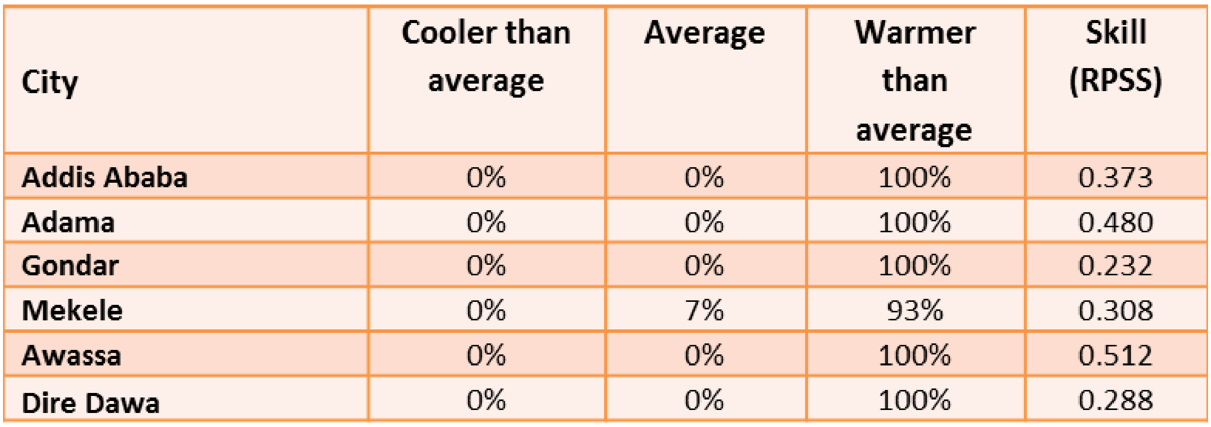


Figure 3.2.2 Violin Plot
The Violin plots above show the full probability distribution of the forecast for December 2009, January 2010, and February 2010. The shading in the background represents climatology. That is to say, which temperature ranges are average (darkest shade), warmer than average (lighter shade) and colder than average (lighter shade) for this season, based on what has been observed in the past 30 years. The colour coded circles represent each of the 15 members of the ensemble forecast used to make these predictions. Red circles show those that predict warmer than average temperatures, grey circles show those that predict average temperatures, and blue circles show those that predict colder than average temperatures. The white dots show the forecast median for each month. A skill score is given using RPSS. A score above 0.5 shows good skill. A score between 0 and 0.5 shows some skill. A score below 0 shows no skill.

In Task 33.1 It was found that those respondents who were most comfortable using statistical information tended to favour communication strategies that represented forecast dispersion or ‘spread’, with some noting a preference for full probability density functions (PDFs). This particular visualisation (Figure 3.2.2) was selected over versions that featured a standard box plot or dots representing ensemble members as it provided a full PDF and could represent modalities in the distribution. Dots representing each ensemble member were however overlaid on the violin plots to provide information about the precise number of ensemble members falling into each tercile. The skill for this visualisation was given as an RPSS score. As with the Bar Graph negative skill scores (RPSS ≤ 0) were coded red.

Table

The Table (Figure 3.2.3), based on a Dengue forecast by Lowe et al. (2014), was included in the selection as several participants in our earlier user needs survey (Task 33.1) noted that they liked to receive information about uncertainty in numeric as well and graphical forms, with tables being a commonly mentioned format. This table displays the predicted probability of upper, middle and lower terciles, along with a skill score (RPSS) for each city included.



City	Cooler than average	Average	Warmer than average	Skill (RPSS)
Addis Ababa	0%	0%	100%	0.373
Adama	0%	0%	100%	0.480
Gondar	0%	0%	100%	0.232
Mekele	0%	7%	93%	0.308
Awassa	0%	0%	100%	0.512
Dire Dawa	0%	0%	100%	0.288

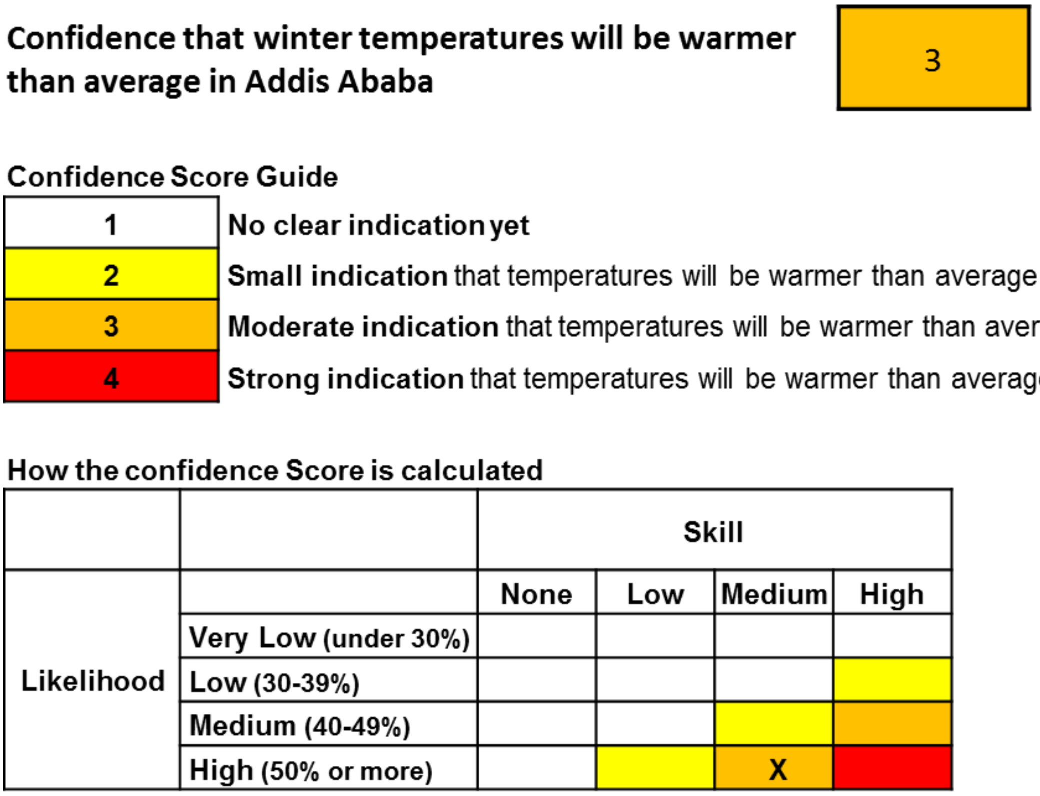
Figure 3.2.3 Table
The table above shows the predicted likelihood of temperatures being warmer than average, average, or colder than average for six locations in Ethiopia. A skill score is given for each location using RPSS. A score above 0.5 shows good skill. A score between 0 and 0.5 shows some skill. A score below 0 shows no skill.

3.2.3 Communication strategies for novice users of statistical information

Two formats were developed specifically for those with less experience of using statistical information: the Confidence Index (Figure 3.2.4) and the Simple Table (Figure 3.2.5) both made use of evaluative categories (i.e. None, Low, Medium, High) to describe skill. These categories were chosen over numeric representations as prior work on risk communication suggests that evaluative categories can be helpful to non-expert decision makers who may struggle to identify how 'good' or 'poor' numeric scores are (see for instance Peters et al., 2009). Both of these formats were first piloted with members of the European public in a large scale survey, with amendments being made on the basis of this initial feedback.

Confidence Index

The Confidence Index (Figure 3.2.4) is adapted from the index used by MeteoFrance to rate confidence in weather forecasts. This Confidence Index uses a colour-coded score of 1-4 to show how strongly the prediction indicates that a particular event will be (in this case warmer than average temperatures). This is done by weighting the predicted likelihood of the event by the skill of the prediction (RPSS is used to represent skill here, but the categorical nature of the prediction means that ROCSS could also be used. In this example the thresholds for different likelihoods (i.e. ‘very low’, ‘low’. ‘medium’, ‘high’) are subjective. However, it is intended that these would, in practice, be user defined.



Confidence that winter temperatures will be warmer than average in Addis Ababa 🟧 3

Confidence Score Guide

1	**No clear indication yet**
🟨 2	**Small indication** that temperatures will be warmer than average
🟧 3	**Moderate indication** that temperatures will be warmer than average
🟥 4	**Strong indication** that temperatures will be warmer than average

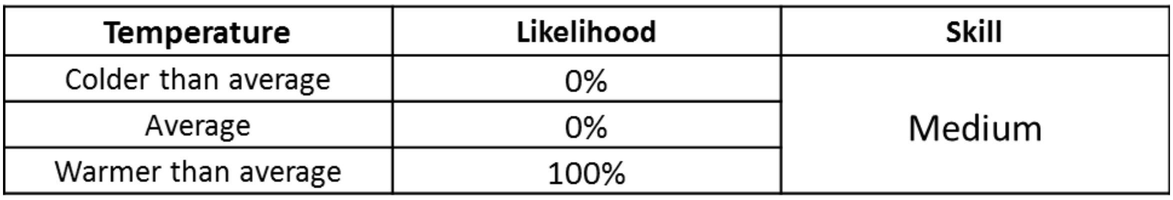
How the confidence Score is calculated

		Skill			
Likelihood		None	Low	Medium	High
	Very Low (under 30%)				
	Low (30-39%)				🟨
	Medium (40-49%)			🟨	🟧
	High (50% or more)		🟨	🟧 X	🟥

Figure 3.2.4 Confidence Index
The Confidence Index shows how strongly the forecast indicates that temperatures will cross a certain threshold (in this case being warmer than average). This is done by combining information about how likely the forecast predicts warmer than average temperatures to be, with information about how well the forecast performs for this season (‘Skill’).

Simple Table

The ‘Simple Table’ (Figure 3.2.5) is a simplified version of the Table for users with greater experience of using statistics. It shows a forecast for one region only, and features a verbal category (i.e. None, Low, Medium, High) derived from RPSS rather than a numeric value to indicate skill.



Temperature	Likelihood	Skill
Colder than average	0%	Medium
Average	0%	
Warmer than average	100%	

Figure 3.2.5 Simple Table
The table above shows the predicted likelihood of winter temperatures being warmer than average, average and colder than average for Addis Ababa. The forecast's performance is shown using a ‘Skill’ rating. Where there is no skill for this time of year this rating is ‘None’.

3.2.1 Communication strategies for all participants

The tercile Bar Graph depicted in Figure 3.2.6 below was selected for inclusion as it represented a format that is already widely used to display seasonal climate predictions (see the public facing Meteo Swiss website for one example). Information about the skill for each tercile (ROCSS) was reported under each of the three bars. In an effort make skill level more salient, scores were colour-coded as red = no skill (ROCSS ≤ 0), grey = some skill (ROCSS > 0 < 0.5), and blue = good skill (ROCSS ≥ 0.5).

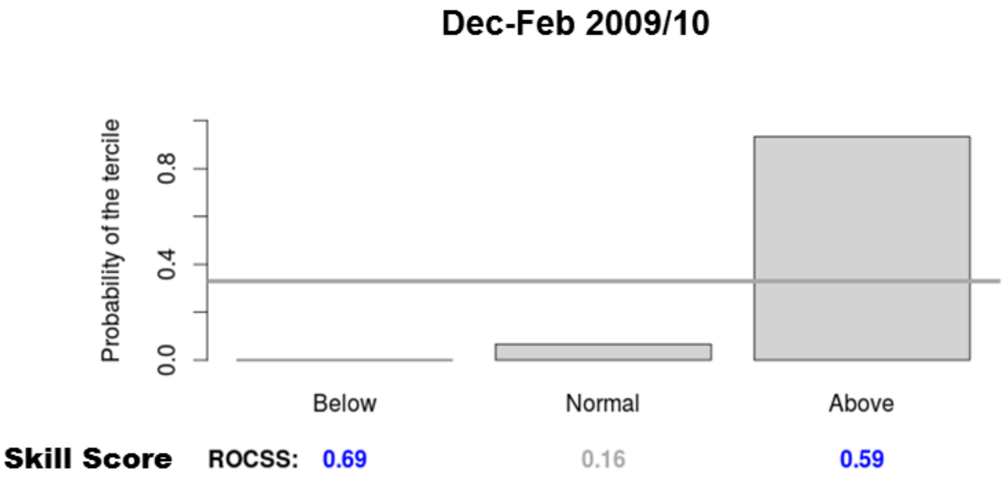


Figure 3.2.6 Bar Graph
The bar graph above shows the predicted likelihood of temperatures being colder than average (Below), average (Normal), and warmer than average (Above). The line going across the graph shows what the graph would look like if all conditions were equally likely. A skill score for each category is given using ROCSS. A score above 0.5 shows good skill (blue). A score between 0 and 0.5 shows some skill (grey). A score below 0 shows no skill (red).

3.3 Abridged Decision Lab

3.3.1 Objective

The objective of this Decision Lab was to test the six selected communication strategies with a large sample of European decision makers in climate and weather sensitive sectors.

3.3.2 Method

3.3.2.1 Participants

284 Participants took part in the study with 264 providing full completes. Participants were recruited through four different channels: the Climate UK network (*n*=5), a European Decision Makers mailing list obtained through Experian (*n*=14), LinkedIn groups with an interest in climate risk management (*n*=11), and a European Business Decision Maker panel managed by Research Now (*n*=254: UK = 65, France = 61, Spain = 64, Germany = 64).

Participants recruited through Research Now were asked a series of screening questions about their employment status, sector, and whether they made weather sensitive decisions in their work. In these instances, only those who reported that they were currently employed in a sector represented in the EUPORIAS project and were responsible for making weather sensitive decisions were invited to continue with the Decision Lab. These screening procedures were not used for participants recruited through the other channels. A breakdown of the sectors represented in the sample is provided in Table 3.3.1. The table also details the countries in which participants' reported that their organisations were based.

Prior to starting the Decision Lab participants were asked to indicate their level of statistical expertise on a 5 point scale:

1. I do not have much experience of using statistical or mathematical information (n= 54)
2. I am comfortable using basic statistical information (e.g. means, percentages) (n=108)
3. I am comfortable using more complex statistical information (e.g. confidence ranges, standard deviations, probability distributions) (n=67)
4. I am comfortable using common statistical tests (e.g. t-tests, correlations) (n=43)
5. I am comfortable using advance statistical techniques (e.g. Monte Carlo Simulations, structural equation modelling) (n=12)

Those who selected 1 or 2 were classified as being novice users of statistical information (n = 162). Those who selected 3, 4 or 5 were classified as advanced users (n = 122). This information was used to determine which communication strategies to present to participants.

Table 3.3.1 Breakdown of the sectors and countries represented in the sample

Sector	**Total number of participants (*n=284*)**	**Total number of participants who completed the study (*n=264*)**
Transport	57	51
Health	49	48
Government/Local government	46	46
Tourism	30	30
Agriculture/Farming	22	20
Finance/Insurance	19	17
Energy	14	14
Aerospace/Aviation	10	8
Water	10	9
Distribution and Logistics	9	8
Research	5	4
Consultancy	2	-
Emergency response	2	1
Forestry	2	2
Utilities (other than energy and water)	2	2
Construction and maintenance	1	1
Education	1	1
Fisheries	1	1
Publishing	1	1
Country		
UK	82	74
Germany	65	59
France	64	63
Spain	64	62
USA	2	1
Belgium	1	1
Denmark	1	1
Italy	1	-
Japan	1	1
Netherlands	1	1
Republic of Ireland	1	1
Sweden	1	-

3.3.2.2 Design

The Decision Lab branched so that novice users of statistical information received different communication strategies than advanced users, with the bar graph being the only format that was shown to both groups.

Advanced Users

For experienced users of statistical information a 4x2 mixed factorial design was used, with Format (Bubble Map vs. Violin Plot vs. Table vs. Bar Graph) as a between groups factor, and Skill (Higher Skill vs. Lower Skill) as a within groups factor. **That is to say that each participant classified as being an experienced user of**

statistical information was shown a Higher Skill prediction and a Lower Skill prediction in one of the four different Formats.

Novice users

For novice users of statistical information a 3x2 mixed factorial design was used, with Format (Bar graph vs. Confidence Index vs. Simple Table) as a between groups factor, and Skill (Higher Skill vs. Lower Skill) as a within groups factor. **That is to say that each participant classified as being a novice user of statistical information was shown a Higher Skill prediction and a Lower Skill prediction in one of the three different Formats**.

3.3.2.3 Communication strategies

As previously noted six types of communication strategy were chosen for inclusion in the Decision Lab: Bubble Map, Violin Plot, Table, Confidence Index and Simple Table (Figures 3.2.1-6). Two versions of each Format were produced. One showing a temperature forecast for a region of the world where prediction skill is comparatively higher (Ethiopia: Higher Skill), and another for a region of the world where skill is comparatively lower (Iberian Peninsula: Lower Skill). All communication strategies were based on the same underlying surface temperature dataset[1] retrieved from ECOMS-UDG (https://meteo.unican.es/trac/wiki/udg/ecoms). Predictions are retrieved from System 4 (15 ensemble members) and observations from WATCH-Forcing-Data-ERA-Interim (WFDEI) (Weedon et al., 2014). The time period considered for these plots is 1982 to 2010. Plots are for northern hemisphere winter (December to February).

3.3.2.4 Measures

The following measures are listed in the order that they appeared to participants.

Objective understanding

Three questions measuring objective understanding of the information about likelihood and skill were common to all communication strategies. For each prediction shown participants were asked to indicate:

1. The probability of temperatures being above average for a particular timeframe and/or region according to the prediction
2. The probability of temperatures being below average for a particular timeframe and/or region according to the prediction
3. The skill of the prediction

In addition to this, visualisation specific questions were asked for the Bubble Map and Violin Plot. For the Bubble Map, participants were asked questions about their understanding of the spatial elements of the map (i.e. which tercile was predicted to be most likely for a specific region). For the Violin Plot, participants were asked about their understanding of the forecast spread. It should be noted that in scoring

[1] It should be noted that in the case of the Simple Table the predicted probability of ‘Below Average’ and ‘Above Average’ terciles for Barcelona differed from the source prediction.

participants responses to questions about likelihood, greater flexibility in what constituted a ‘correct response’ was permitted for the Bubble Map, Violin Plot and Bar Graph, where precise numeric values were not given.

Subjective interpretation

For both the Higher Skill and Lower Skill predictions, participants were asked three questions about their subjective interpretation of the predictions:

1. Looking at the forecast and its skill how likely do you think that it is that temperatures will be Warmer than average? *(1 = very unlikely, 10 = very likely)*
2. How confident are you in this judgement? *(1 = not confident at all, 10 = very* confident)
3. How useful do you think that this type of forecast would be for decision making in your organisation? *(1 = not useful at all, 10 = very useful)*

Subjective preference

To measure subjective preference participants were asked to rate their agreement with five statements *(1=strongly disagree, 5=strongly agree).*

1. “I like this type of [FORMAT]”
2. “I find this type of [FORMAT] easy to understand”
3. “This type of [FORMAT] provides useful information”
4. “I would use this type of [FORMAT] in my decision making”
5. “I would share this type of [FORMAT] with other people in my organisation for them to use in their decision making”

Familiarity

Familiarity was measured using level of agreement *(1=strongly disagree, 5=strongly agree)* with the statement:

> “I already use this type of [FORMAT] in my work”

3.3.2.4 Procedure

Participants were invited to take part in the study via social media, mailing lists, direct approach by email, and approach by the sampling company Research Now. Those who clicked on the link provided were directed to a Qualtrics survey where they were told about the aims of the study, and that all of the data gathered would be anonymised before being reported. Those who indicated that they wished to proceed were asked preliminary questions about their sector and statistical expertise. Depending on their self-reported level of statistical expertise, they were then randomly assigned to either one of the four formats for experienced users of statistical information, or one of the three formats for novice users.

Participants were presented first with the Higher Skill prediction, and asked to respond to questions about their objective understanding and subjective interpretation of the information provided. They were then presented with the Lower

Skill prediction, and asked a comparable set of questions about objective understanding and subjective interpretation. After completing these questions, they were asked to rate their opinion of the format.

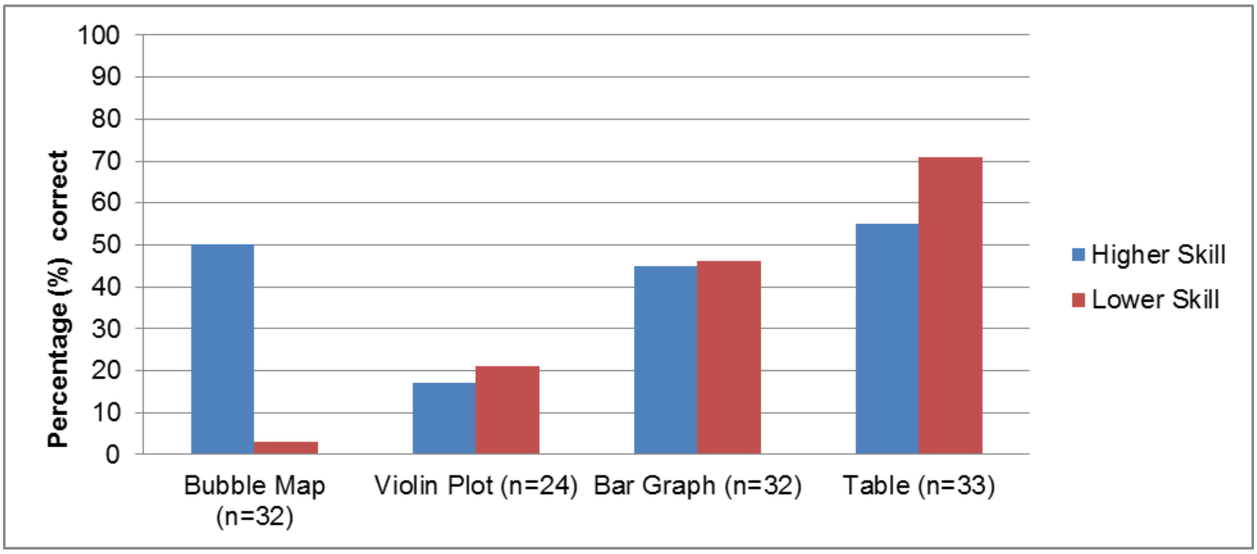
3.3.3 Abridged Decision Lab: Objective understanding

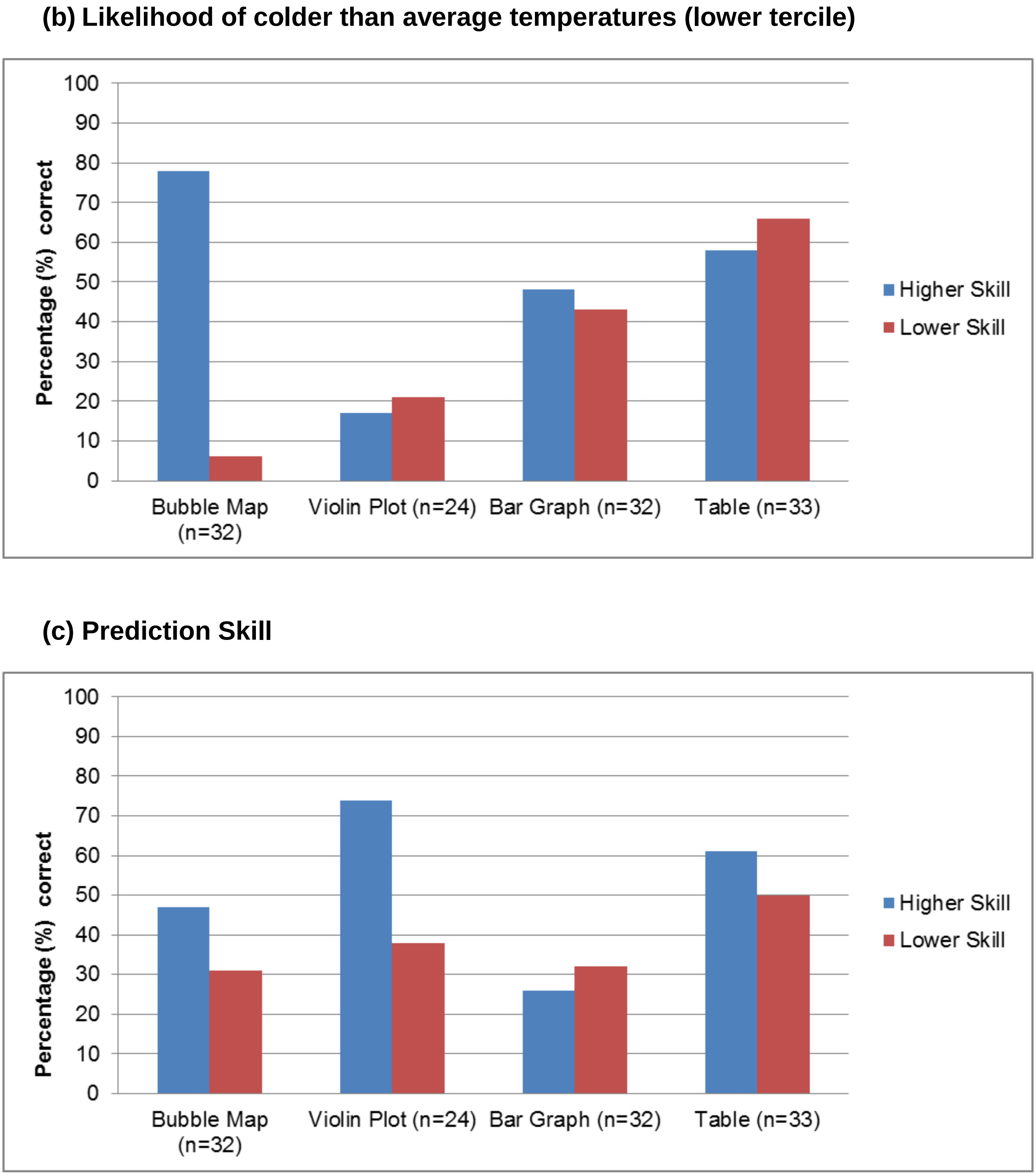
As noted in Section 3.3.2, for each prediction participants were asked questions designed to assess their objective understanding of the content. To facilitate comparison of the communication strategies, three of them were common to all formats: a) likelihood of warmer than average temperatures; b) likelihood of colder than average temperatures; and c) prediction skill. For the Bubble Map and Violin Plot additional questions about participants' understanding of their unique characteristics were included. In this set of analyses we compare how many of the questions common to all predictions were answered correctly by experience users of statistical information (3.3.3.1) and novice users of statistical information (3.3.3.2), before going on to examine responses to each of the communication strategies in detail (3.3.3.3-8) for the purpose of identifying where potential misunderstandings lie and how these might be resolved.

3.3.3.1 Objective understanding: Communication strategies for experienced users of statistical information

Figures 3.3.1a-c show the proportion of participants who answered each of the three questions common to all communication strategies correctly for both the Higher Skill and Lower Skill versions that they were presented with.

(a) Likelihood of warmer than average temperatures (upper tercile)





(b) Likelihood of colder than average temperatures (lower tercile)

(c) Prediction Skill

Figure 3.3.1 Proportion of experienced users of statistical *information* correctly answering questions about the (a) predicted likelihood of warmer than average temperatures (upper tercile), (b) predicted likelihood of colder than average temperatures (lower tercile), and (c) prediction skill, for both the Higher Skill (Ethiopia) and Lower Skill (Iberian Peninsula) visualsations.

Likelihood

We see that, for the questions about the likelihood of temperatures being warmer than average (upper tercile) and colder than average (lower tercile), the proportion of correct responses was higher for the Table than the Bar Graph, and higher for the Bar Graph than the Violin Plot. This pattern was observed for both the Higher Skill and Lower Skill predictions. When it came to the Bubble Map however, it can be seen that participants responded more accurately to the Higher Skill prediction than the Lower Skill prediction.

To assess whether these differences were statistically significant a pair of 2x4 mixed factorial[2] analysis of variance (ANOVA) tests were conducted, with Correct Response as a dependent variable, Format as a between groups variable, and Prediction Skill as a repeated measures variable (see Sheffe (1965, 1999) for a description of this procedure, and Lunney (1970) for a discussion of its application in cases where dependent variables are dichotomous)[3]. A significant main effect of Format was found for the questions concerning both the likelihood of warmer than average temperatures (*F(3, 112)*=8.5, *p* < .001) and the likelihood of colder than average temperatures (*F(3, 112)*=6.1*, p* =.001). Likewise, a significant interaction[4] was found between Format and Prediction Skill (warmer than average: *F(3, 112)*=9.6, *p* < .001; colder than average: *F(3, 112)*=18.5, *p* < .001). An inspection of post-hoc Bonferroni tests showed that for warmer than average temperatures, correct responses were higher overall for those shown the Table than those shown the Bubble Map (*p* < .001) or Violin Plot (*p* = .001), while correct responses were higher for those shown the Bar Graph than the Violin Plot (*p*=.04). For the question about colder than average temperatures however the main effect of Format was less pronounced, with correct responses for the Table being higher than for the Violin Plots only (*p*<.001). Only a marginally significant difference between correct responses for the Bar Graph and responses for the violin Plot were found (*p*=.07).

[2] A mixed-factorial test is one that allows the inclusion of both between groups and repeated measures factors. In experimental social science a between groups factor refers to an independent variable where participants are assigned to a single condition or ‘group’ (in this case each participant was assigned a single visualisation format). A repeated measures or ‘within groups’ factor refers to an independent variable where participants take part in all condition (in this case each participant who completed the study was presented with both a Higher Skill and Lower Skill prediction)

[3] Like linear regression, Analysis of Variance (ANOVA) falls within the General Linear Model. However, the fact that it allows direct comparisons to be made between multiple experimental conditions makes it more suitable for the exploratory analysis of data from social scientific experiments than regression.

[4] Interaction effects refer to cases where the effect of one independent variable is contingent on the value of another independent variable. In this case we find that the affect of visualisation format on rate of correct response is contingent on the whether a Higher Skill or Lower Skill prediction is being shown.

In Summary

Taken together these results suggest that for this group the Table and Bar Graph were the visualisations that best facilitated the precise interpretations of information about the likelihood of particular terciles; with the Violin Plot eliciting fewer correct responses. Interestingly however, we found that for the Bubble Map, interpretation of the likelihood information presented was far more accurate from the Higher Skill visualisation than the Lower Skill visualisation, suggesting that participants may have not understood that the predominance of 'black space' should be interpreted as the prediction indicating that all terciles should be considered equally likely.

Skill

In contrast to the questions about likelihood, it was found the proportion of correct responses to the questions about skill was highest for the Violin Plot and Table and lowest for the Bar Graph. A 2x4 mixed factorial ANOVA confirmed that there was a main effect of Format on correct responses to this question (*F(3, 112)*=3.4 , *p* =.02). Subsequent post-hoc Bonferroni tests showed that those presented with the Table were more accurate in their judgement of skill than those shown the Bar graph (*p* = .04), while a marginally significant difference between the Violin Plot and Bar Graph was found (*p*=.08). Additionally, it was found that the level of skill inherent in the prediction (Higher Skill versus Lower Skill) had a significant main effect on the accuracy of participants' interpretation of the Skill Information, with interpretation being more accurate for the Higher Skill predictions than the Lower Skill predictions.

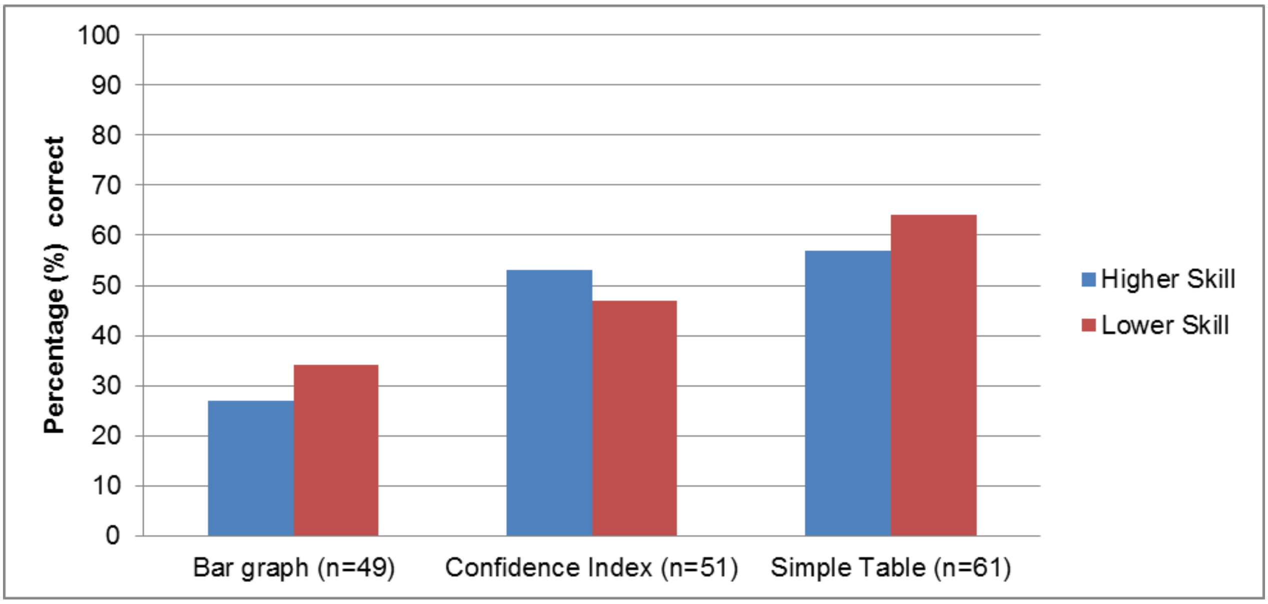
In Summary

We find that the accurate interpretation of information about skill was highest in those cases where skill scores were above 0 (i.e. in the Higher Skill visualisations). As this was found for all visualisations, it would appear to suggest that participants had difficulty interpreting than a negative skill value meant that there was no skill. We also find that responses were least accurate when three different scores were given (see 3.3.3.5 for a discussion of possible reasons for this).

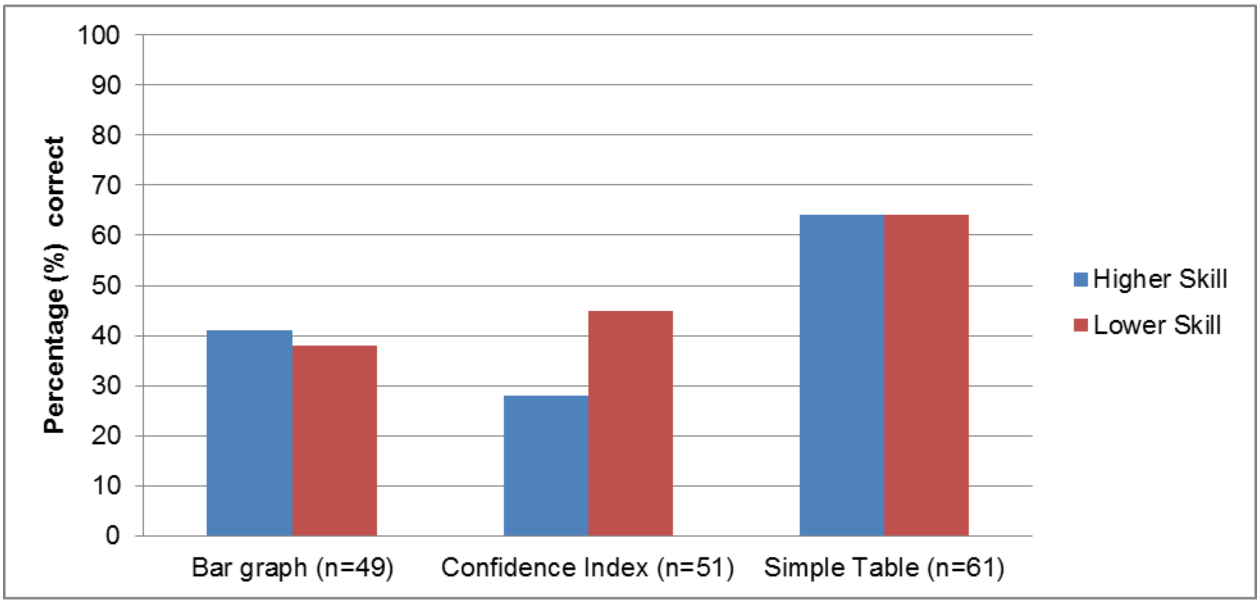
3.3.3.2 Objective understanding: Communication strategies for novice users of statistical information

For those participants classified as novice users of statistical information Figures 3.3.2a-c show the proportion who answered each of the questions about likelihood and skill correctly.

(a) Likelihood of warmer than average temperatures (upper tercile)



(b) Likelihood of colder than average temperatures (lower tercile)



(c) Prediction skill

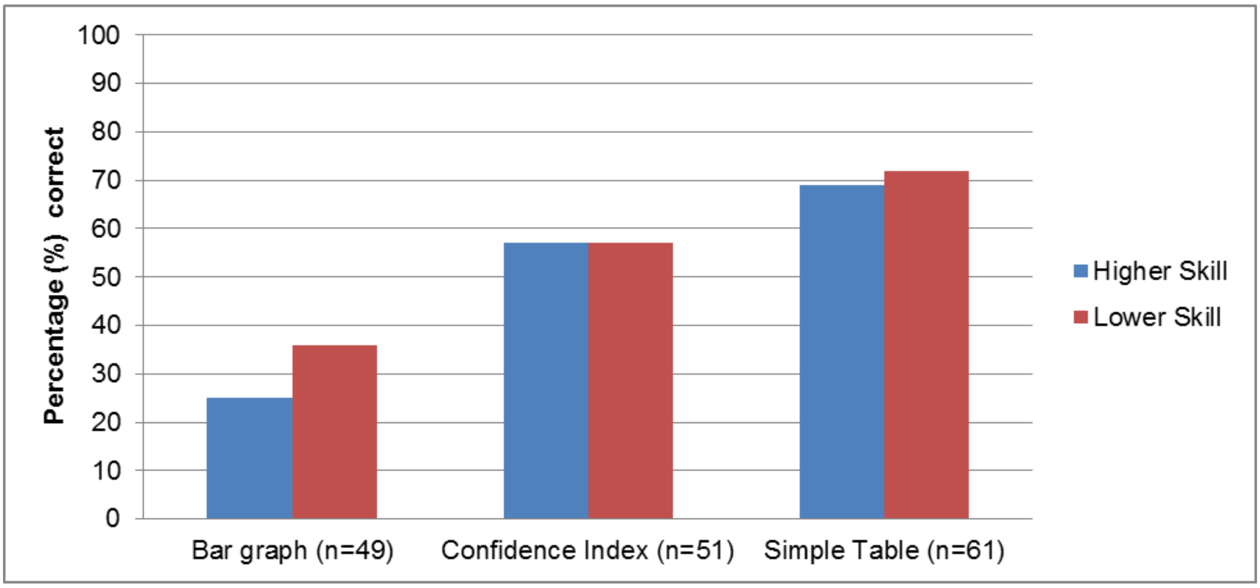


Figure 3.3.2 Proportion novice users of statistical information correctly answering questions about (a) the predicted likelihood of warmer than average temperatures (upper tercile), (b) the predicted likelihood of colder than average temperatures (lower tercile), and (c) prediction skill, for the Higher Skill (Ethiopia) and Lower Skill (Iberian Peninsula) visualsations.

Likelihood

As can be seen from Figures 3.3.2a and 3.3.2b a clear pattern was found with respect to the accuracy of participants' interpretation of the information about the predicted likelihood of warmer than average (upper tercile) and colder than average (lower tercile) temperatures. In all cases the Simple Table was found to elicit the greatest number of correct responses. The Confidence Index meanwhile was found to elicit a greater number of correct responses than the Bar Graph when the question concerned warmer than average temperatures (Figure 3.3.2a). However, accuracy appeared to diminish when the question was about the likelihood of colder than average temperatures. In this latter case, many participants assigned to the Confidence Index did not recognise that this particular format did not actually provide any information about the likelihood of temperatures being colder than average.

This interpretation was supported by pair of 3x2 mixed-factorial ANOVAS, with Correct Response as a dependent variable, Format as a between groups variable, and Prediction Skill as a repeated measures variable. A main effect of Format was found for both "likelihood of warmer than average temperatures" (upper tercile) (*F(2, 151)*=6.6, *p*=.002) and "likelihood of colder than average temperatures" (lower tercile) (*F(2, 151)*=6.5, *p*=.002). For questions regarding the likelihood of warmer than average temperatures, post-hoc Bonferroni tests show that those shown the Simple Table performed better than those shown the Bar Graph (*p*=.001). Those shown the Confidence Index also performed better than those shown the Bar Graph, although the difference was only marginally significant (*p*=.06). For the questions regarding the likelihood of colder than average temperatures, those shown the

Simple Table performed significantly better than those shown either the Bar Graph (*p*=.02) or Confidence Index (*p*=.004).

In Summary
Overall, we find that when it came to understanding information about predicted likelihood, the Simple Table was the most well understood of the communication strategies shown to novice users of statistical information, followed by the Confidence Index; although, many participants did not correctly identify that the Confidence Index did not provide any information about the likelihood of temperatures being colder than normal. The Bar Graph was the least accurately interpreted of the three communication strategies (reasons for this, along with potential solutions are discussed in 3.3.3.5).

Skill

As can be seen in Figure 3.3.2c, rate of correct response to questions about skill was highest for the Simple Table and lowest for the Bar Graph A 2x3 mixed factorial ANOVA confirmed these difference to be statistically significant (*F(2, 151)*=18.9, *p*<.001), with post-hoc Bonferroni tests showing that those shown the Simple Table performed better on the skill questions than those shown the Bar Graph (*p* <.001). The difference between those shown the Simple Table and those shown the Confidence Index was marginally significant (*p*=.07). Those who were shown the Confidence Index meanwhile, performed better on the skill questions than those shown the Bar Graph (*p*=.001).

In Summary
For novice users of statistical information, the Simple Table appears to have been the most effective of the three visualisations for conveying skill, followed by the Confidence Index. For these participants the Bar Graph appears to have been the most difficult to interpret when it came to information about skill. This may be due to the fact that two former visualisations presented information about skill as a category (e.g. High, Medium, Low, None) rather than a numeric value.

3.3.3.3 Objective Understanding: Bubble Map

Spatial Elements

In addition to the questions regarding likelihood and skill reported above, those participants who were shown the Bubble Map were asked questions regarding their interpretation of the spatial components of the graph. For both the Higher Skill version of the map (Horn of Africa) and the Lower Skill version of the map (Iberian Peninsula) 50% of participants gave accurate responses, correctly identifying regions where warmer than average or colder than average temperatures were predicted.

General Considerations

As detailed above, it was found that those presented with the Bubble Map responded more accurately to questions about the predicted likelihood of terciles when skill was high than when it was low. That is to say that most participants were not aware that where ‘white space’ was present one should assume that all terciles are equally likely. Indeed, an examination of the pattern of incorrect responses to this visualisation showed that many participants appear to have interpreted ‘white space’ as indicating a 0-10% probability of both upper and lower terciles. This suggests that it needs to be explicitly stated that where white space exists then all terciles should be considered to be equally likely. The interpretation of this style of map may also be assisted by providing an additional key that illustrates what different sized bubbles represent.

3.3.3.4 Objective Understanding: Violin Plot

Range

In addition to the questions about likelihood and skill, participants were asked questions about ‘spread’ of the forecast (i.e. which values temperatures were expected to fall between based on the spread of the prediction). This question was answered correctly by 38% of participants for the Higher Skill prediction and 50% for the Lower Skill prediction.

General Considerations

Of the four communication strategies presented to experienced users of statistical information the Violin Plot was the one that participants appeared to find difficult to interpret when it came to estimating the likelihood of upper and lower terciles; as was demonstrated both by the spread of incorrect responses to the likelihood questions, and the fact that the study ‘drop out’ rate was higher for this communication strategy than for others. While it is possible to extrapolate precise likelihoods from this visualisation (i.e. by examining the colour coded ensemble members), our findings suggest that it is more suited for situations where a range of values rather than precise categorical probabilities is required.

When it came to extrapolating information about skill from this diagram, it was found that more participants correctly interpreted the meaning of the skill score (RPSS) for the Higher Skill than the Lower Skill prediction. This indicates that some participants did not realise that a negative value indicated No Skill. This suggests that, where negative skill exists, explicitly stating that there is ‘No Skill’ may be necessary.

3.3.3.5 Objective Understanding: Bar Graph

Amongst experienced users of statistical information, the Bar Graph elicited the second highest rate of correct response to the questions on predicted likelihood after the Table. For novice users however it was the least well understood communication strategy. An inspection of the pattern of incorrect response to these questions suggested that some participants mistook the Skill scores under each bar for likelihood information, while others interpreted the line representing climatology as an indicator of probability. Taken together this suggests that skill scores should be either repositioned to avoid confusion or replaced by categorical indicators of skill (e.g. None, Low, Medium, and High), and that climatology should not be included in this way.

While both experienced and novice users of statistical information performed moderately well in extracting information about tercile likelihood from the Bar Graph, it was the communication strategy that elicited the lowest rate of correct responses to the questions about skill. One possible reason for this is that– as mentioned above – some participants mistook the ROCSS values for information about likelihood. However, the pattern of incorrect responses to questions about skill also suggests that some may have struggled to interpret the fact that – in the case of the Higher Skill prediction – skill scores were better for the upper and lower terciles (Good Skill) than the middle tercile (Some Skill). This highlights a potential trade-off between providing detailed information about the performance of the forecasting system, which may differ between terciles, and providing a more limited amount of information (i.e.. a single skill score).

3.3.3.6 Objective Understanding: Table

Of the four communication strategies presented to experienced users of statistical information, the Table was that which elicited the highest proportion of correct responses to the questions on predicted likelihood and skill. However, responses to the questions on skill suggest that some participants had difficulty interpreting negative skill scores, incorrectly perceiving these to indicate that there was ‘Some Skill’ as oppose to ‘No Skill’. Again, this may suggest that, where negative skill exists, explicitly stating that there is ‘No Skill’ may be necessary.

3.3.3.7 Objective Understanding: Confidence Index

Amongst novice users of statistical information the Confidence Index appears to have been better understood than the Bar Graph, but less well understood than the Simple Table. An inspection of the responses given to the questions about the likelihood of colder than average temperatures, indicates that while this information was not provided (making “the forecast does not provide this information” the correct response), some participants made inferences based on the information provided about warmer than average temperatures. This suggests that this style of communication strategy may only be suitable for situations where a binary split between Event/Not Event is present.

3.3.3.8 Objective Understanding: Simple Table

As with more complex version presented to experienced users of statistical information, the Simple Table emerged as the communication strategy that elicited the higher proportion of accurate responses when it came to questions about likelihood and skill. Once again, this suggest that using evaluative categories (e.g. None, Low, Medium, High) to refer to skill may be a more effective way to communicate forecast performance to those who are less comfortable with using statistical information, especially in cases where skill scores are negative.

3.3.4 Abridged Decision Lab: Subjective interpretation of predictions

For each communication strategy that they were shown, participants were asked to indicate, on a scale of 1-10, how likely they thought that warmer than average temperatures would be observed for the forecast period, taking into account the forecast and its skill. They were then asked to rate how confident they were in this judgement. With the exception of the Bubble Map, where there was some skill in a small number of regions, the predictions for the Iberian Peninsula (Lower Skill) showed negative skill scores (i.e. No Skill). Hence, for these Lower Skill predictions, judgements of likelihood should be roughly consistent across all formats (i.e. with participants discounting the information about likelihood provided by the prediction). As can be seen from Tables 3.3.2 and 3.3.3 however this was not the case. Amongst experienced users of statistical information, we find that those presented with the Violin Plot tended to judge warmer than average temperatures to be more likely that those presented with the Bubble Map. Amongst novice users, we found that those presented with the Simple Table judged warmer than average temperatures to be more likely than those shown the Confidence Index and Bar Graph.

Table3.3.2 Mean (SD) subjectively perceived likelihood of temperatures being warmer than average for the Higher Skill and Lower Skill predictions amongst experienced users of statistical information

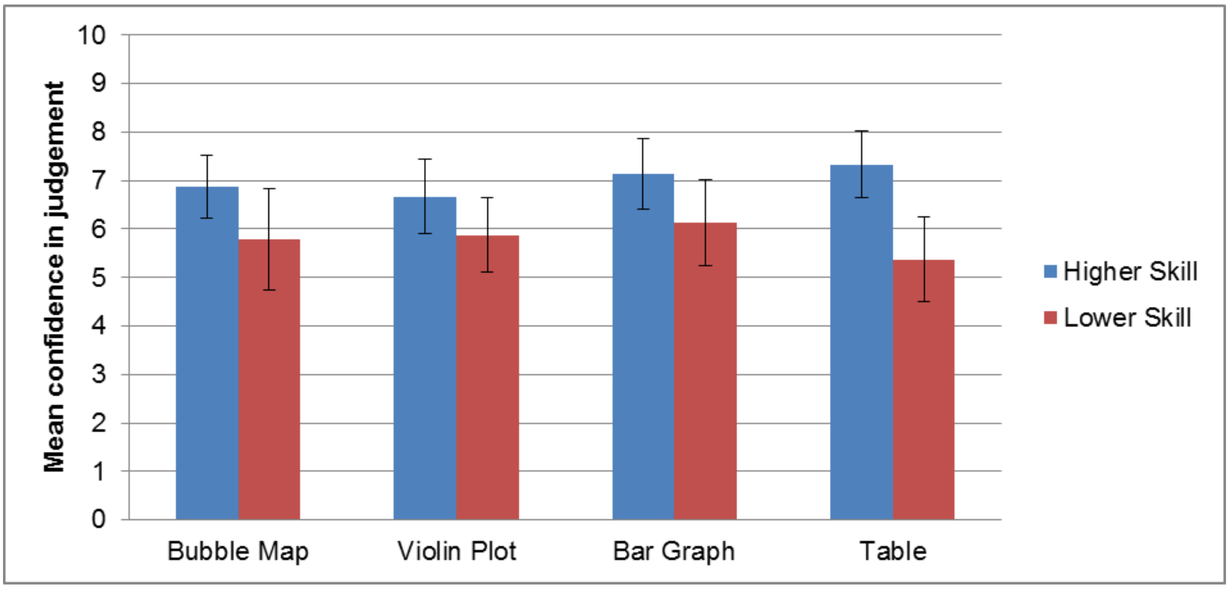
	Bubble Map	Violin Plot	Bar Graph	Table
Higher Skill (Ethiopia)	7.8 (1.4)	7.0 (2.0)	7.8 (2.2)	6.9 (2.5)
Lower Skill (Iberian Peninsula)	3.4 (2.8)	5.1 (2.0)	4.6 (2.5)	4.7 (2.1)

Table 3.3.3 Mean (SD) subjectively perceived likelihood of temperatures being warmer than average for the Higher Skill and Lower Skill predictions amongst novice users of statistical information

	Bar graph	Confidence Index	Simple Table
Higher Skill (Ethiopia)	7.4 (2.1)	6.8 (1.5)	8.2 (2.3)
Lower skill (Iberian Peninsula)	4.4 (2.6)	3.8 (2.3)	5.7 (2.0)

When asked about their confidence in this judgement however, it was found participants were less confident in their judgements when it came to the Lower Skill prediction than the Higher Skill prediction. This is illustrated in Figures 3.3.3a-b.

(a) Confidence in subjective judgement of likelihood amongst *e*xperienced users of statistical information



(b) Confidence in subjective judgement of likelihood amongst novice users of statistical information

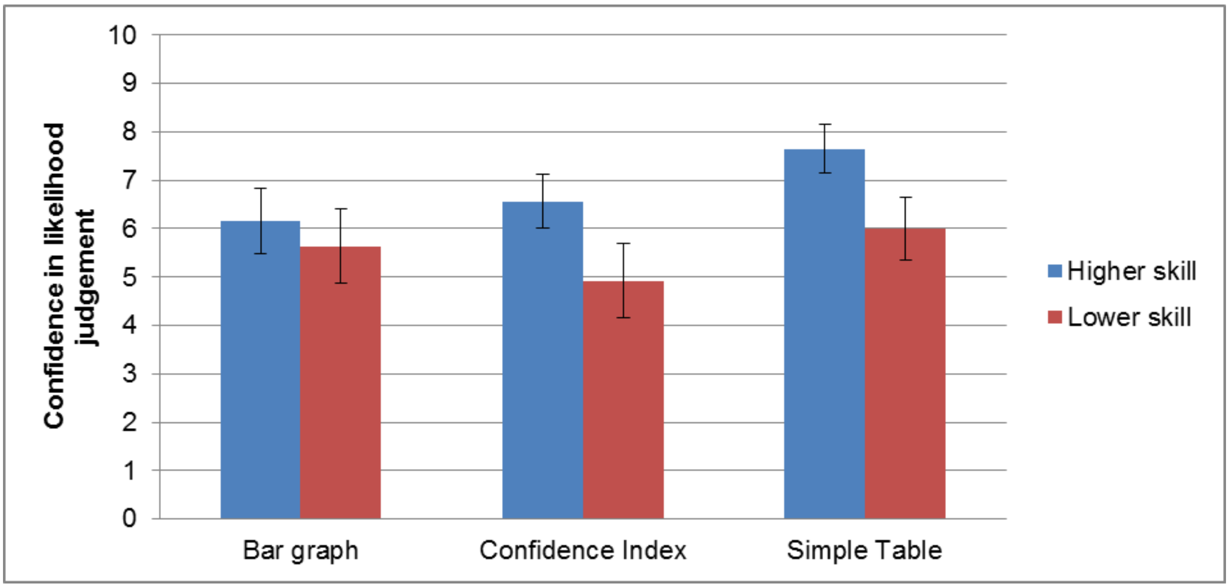


Figure 3.3.3 Mean confidence in subjective judgement of likelihood (1=Not confident at all, 10=Very confident) amongst (a) experienced users of statistical information and (b) novice users of statistical information. Error bars represent 95% confidence interval around the mean.

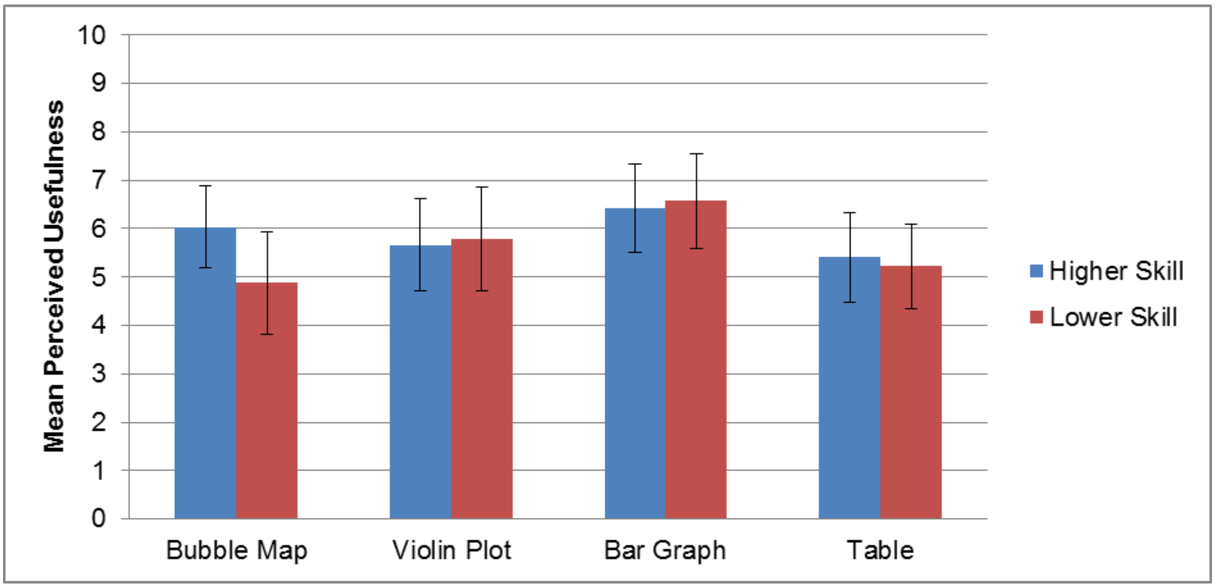
In Summary

In summary, our findings suggest that even when skill was negative our participants still used the predicted probabilities on the visualisations in making subjective judgements about how likely it was that temperatures would be warmer than average. However, confidence in these judgements was lower when it came to Lower Skill visualisations than the Higher Skill visualisations.

3.3.5 Abridged Decision Lab: Perceived Usefulness

For both the Higher Skill and Lower Skill predictions participants were asked to rate how useful they thought that the type of information provided by the predictions would be for decision making within their organisation. Mean ratings of Perceived Usefulness are illustrated in Figures 3.3.4a-b.

(a) Perceived Usefulness amongst experienced users of statistical information



(b) Perceived Usefulness amongst novice users of statistical information

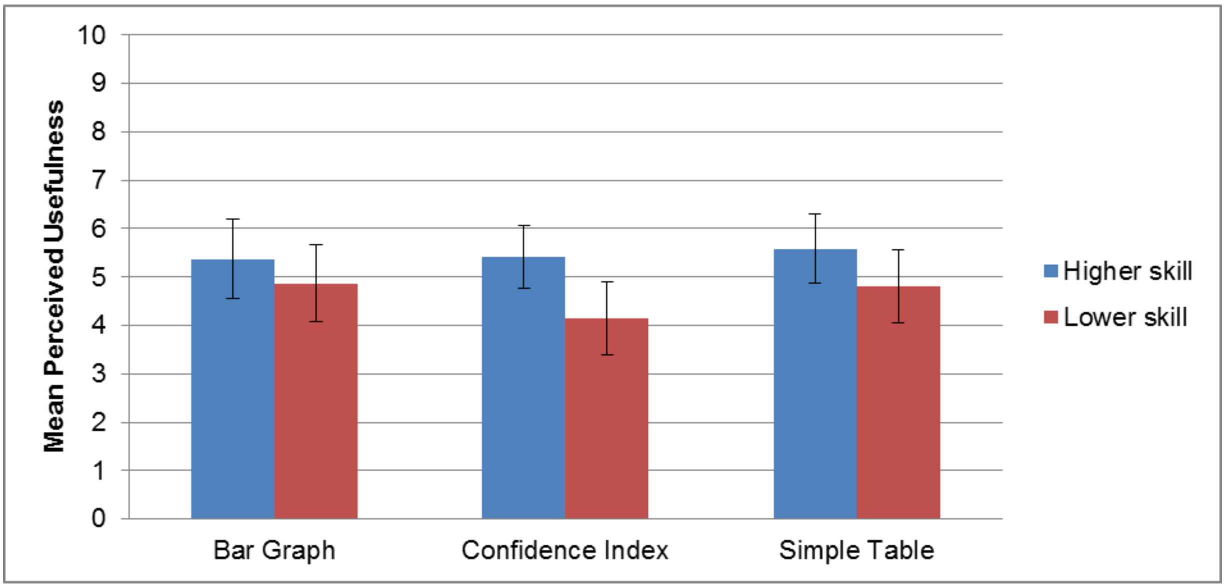


Figure 3.3.4 Mean perceived Usefulness of each of the communication strategies for decision making in one's organisation for (a) experienced and (b) novice users of statistical information. Error bars represent 95% confidence interval around the mean.

A pair of 2x4 mixed factorial ANOVA tests indicated that amongst experienced (*F(1,111)*=3.9, *p*=.05) and novice users of statistical information (*F(1,151)*=37.3, *p*<.001), formats depicting Higher Skill predictions were perceived as more useful for decision making than those depicting Lower Skill.

In Summary

In summary, Higher Skill visualisations were perceived as more Useful that Lower Skill visualisations.

3.3.6 Abridged Decision Lab: Preference and Familiarity

Participants' subjective preference for the different communication strategies was measured using the five items described in Section 3.2.2. As there was strong internal consistency between responses to these items (Cronbach's alpha > 0.90)[5], a summary measure of Preference was created by taking their mean. Familiarity was measured by agreement with statement *"I already use [FORMATS] like this in my work"*. Mean preference scores are report for experienced and novice users of statistical information in Tables 3.3.4 and 3.3.5 respectively.

[5] The Cronbach's alpha reliability coefficient measures the inter-item homogeneity (or 'consistency') between items on psychometric measurement scales (Cronbach, 1951). Values above 0.7 are deemed to represent an acceptable level of internal consistency for items to be combined into a single measure (Nunnaly, 1978)

Experienced users of statistical information

For experienced users of statistical information, mean ratings of Preference and Familiarity can be found in Table 3.3.4. While Preference ratings were above the scale mid-point of 3, ratings of Familiarity were below it, indicating that the communication strategies presented were relatively unfamiliar to participants.

Mean Preference ratings were slightly higher for the Bubble Map and Table than the Violin Plot and Bar Graph, while mean Familiarity ratings were highest for the Table and Violin Plot. However, a Multivariate Analysis of Variance (MANOVA)[6] test with both Preference and Familiarity entered as dependent variables, indicated that these differences were not statistically significant (*λ*=1.0, F*(6, 216)*=.4, *p*=.90).

Table 3.3.4 Mean ratings of Preference and Familiarity amongst experienced users of statistical information

	Bubble Map (*n*=32) Mean (SD)	Violin Plot (*n*=23) Mean (SD)	Bar Graph (*n*=27) Mean (SD)	Table (*n*=31) Mean (SD)
Preference	3.4 (1.0)	3.2 (1.0)	3.3 (1.0)	3.4 (0.9)
Familiarity	2.2 (1.2)	2.4 (1.2)	2.3 (1.3)	2.5 (1.1)

In Summary

For those with Greater Statistical Experience, ratings of Preference and Familiarity did not significantly differ between the four different types of visualisation.

Novice users of statistical information

Mean ratings of Preference and Familiarity amongst novice users of statistical information are detailed in Table 3.3.5. A MANOVA test showed a significant overall effect of Format (*λ*=0.9, F*(4, 294)*=2.7, *p*=.03), with follow up ANOVAs showing a significant main effect for Preference (F*(2, 147)*=4.8, *p*=.004), but only a marginally significant effect for Familiarity (F*(2, 147)*=2.7, *p*=.09). Post-hoc Bonferroni tests showed that Preference ratings for the Simple Table were higher than Preference ratings for the Bar Graph (*p*=.01) and the Confidence Index (*p*=.08), although the latter difference was only marginally significant.

[6] The Multivariate Analysis of Variance (MANOVA) test represents an extension of the Analysis of Variance (ANOVA), which allows for the simultaneous analysis of the effect of one or more independent variables two or more dependent variables whilst avoiding the inflated risk of Type 1 errors that arises from performing multiple tests (see for instance Bray & Maxwell, 1985).

Table 3.3.5 Mean ratings of Preference and Familiarity amongst novice users of statistical information

	Bar Graph (*n*=47) Mean (SD)	Confidence Index (*n*=49) Mean (SD)	Simple Table (*n*=54) Mean (SD)
Preference	2.8 (1.2)	3.0 (1.0)	3.4 (0.8)
Familiarity	2.1 (1.3)	2.0 (1.2)	2.5 (1.2)

In Summary
For novice users of statistical information, the Simple Table was preferred to the Bar Graph and Confidence Index.

3.3.7 Abridged Decision Lab: Are Preference and Familiarity Associated with Objective Understanding?

Prior research examining user perceptions of climate visualisations has failed to find an associated between preference and objective understanding (see for instance Lorenz et al., 2015). That is to say that those communication strategies that users prefer may not necessarily be those that they best understand. As such a disparity could pose a challenge for forecast providers, we examined whether this was the case amongst our sample. To Begin, we examined the simple correlations between Objective Understanding (sum of correct responses), Preference, Familiarity, Confidence in Judgements of Likelihood (for Higher Skill and Lower Skill predictions), and perceived Usefulness for decision making in one’s own organisation (for Higher Skill and Lower Skill predictions). This is shown in Table 3.3.6 below. As one can see, Preference has a moderate to strong positive associated with Familiarity, Confidence in judgements of likelihood, and Perceived Usefulness, but not Objective Understanding. Indeed, Objective Understanding was negatively associated with Familiarity, suggesting that those who reported greater familiarity with the communication strategies were more prone to misinterpret them.

Table 3.3.6 Correlation (Pearson's *R*) between measures of Objective Understanding, Preference, Familiarity, Confidence in judgement and Perceived Usefulness across all communication strategies

	Familiarity	Preference	Confidence in judgement (Higher Skill)	Confidence in judgement (Lower Skill)	Perceived usefulness (Higher Skill)	Perceived Usefulness (Lower Skill)
Objective Understanding	-.17**	.07	.33***	-.09	-.03	-.15*
Familiarity		.52***	.10	.18***	.39***	.47***
Preference			.35***	.29***	.57***	.55***
Confidence in judgement (Higher Skill)				.48***	.31***	.25***
Confidence in judgement (Lower Skill)					.27***	.55***
Perceived usefulness (Higher Skill)						.74***

*Significant at *p*≤.05 **Significant at *p*≤.01 ***Significant at *p*≤.001

Our initial correlational analysis showed that while Preference did not directly correspond with greater Objective Understanding, it did correspond with greater perceived Familiarity. As a negative correlation was found to exist between Familiarity and Objective Understanding, this raised the question of whether Familiarity may be suppressing a positive relationship between Preference and Objective Understanding. To examine whether this was the case a hierarchical linear regression analysis was performed; with Preference as a criterion variable and Objective Understanding (Block 1) and Familiarity (Block 2) entered as predictors. This analysis is reported in Table 3.3.7 below.

When entered alone in Block 1, Objective Understanding was not associated with Preference. However, when Familiarity was controlled for in Block 2 a significant positive association was found between Objective Understanding and Preference. This suggests that while greater understanding does correspond with a stronger preference for particular communication strategies, this effect is suppressed by perceived familiarity.

Table 3.3.7. Hierarchical linear regression examining the extent to which Objective Understanding and Familiarity predict Preference (Unstandardized B and Standardised β coefficients reported)

	Block 1		Block 2	
	B (SE)	*β*	*B (SE)*	*β*
Objective Understanding	.04 (.03)	.07	.09 (.03)	.14**
Familiarity	-	-	.46 (.04)	.55***
ANOVA	*F(1,261)=1.4*		*F(2,260)=54.7*	
ΔR^2	.01		.30	

*Significant at *p*≤.05 **Significant at *p*≤.01 ***Significant at *p*≤.001

In Summary

We find that amongst this sample a stronger preference for particular communication strategies corresponds with greater perceived familiarity, greater confidence in one’s subjective interpretation of what the visualisations show, and greater perceived usefulness. We also find that, all else being equal, stronger preference ratings corresponded with better objective understanding. However, this is obscured by the fact that familiarity was negatively related to objective understanding. This suggests that users who perceive new communication formats to be similar to ones that they already use may be prone to misinterpreting them as a result (i.e. mistaking novel elements for something else).

3.4 Full Decision Lab

3.4.1 Objective

The objective of this study was to assess responses to the six types of communication strategy selected for inclusion in this study amongst a small but highly engaged sample of participants.

3.4.2 Method

3.4.2.1 Participants

95 Participants took part in the study with 58 providing full completes. Participants were primarily recruited through the EUPORIAS stakeholder mailings lists, EUPORIAS Twitter feed, and the Climate Service Partnership mailing list (www.climate-service-center.de), with some participants passing the survey link to colleagues. In recruiting participants we targeted European organisations in key climate sensitive sectors. However, we also received responses from both organisations outside of Europe and meteorological/research institutions. We include all responses in our analysis, but control for institutional differences. A breakdown of the sectors represented in the sample can be found in Table 3.4.1.

Table 3.4.1 Breakdown of the sectors and countries represented in the sample

Sector	Total number of participants (*n=95*)	Total number of participants who completed the study (*n=58*)
Research	31	15
Agriculture and Farming	11	9
Meteorology /Climate services	11	8
Water	10	8
Consultancy	8	4
Central and local government	6	4
Health	4	
Energy	3	1
Environmental monitoring/services/adaptation	3	3
Transport	3	2
Emergency response	1	
Finance	1	1
Non-profit	1	1
Wine	1	1
Other	1	1
Region		
Europe	49	31
South America	12	8
Africa	9	4
North America	9	5
Australia	4	3
Asia	2	1
No Information	10	6

At the start of the study, participants were asked to indicate their level of statistical expertise on a 5 point scale:

1. I do not have much experience of using statistical or mathematical information (n=1)
2. I am comfortable using basic statistical information (e.g. means, percentages) (n=10)
3. I am comfortable using more complex statistical information (e.g. confidence ranges, standard deviations, probability distributions) (n=35)
4. I am comfortable using common statistical tests (e.g. t-tests, correlations) (n=30)
5. I am comfortable using advance statistical techniques (e.g. Monte Carlo Simulations, structural equation modelling) (n=19)

Those who selected 1 or 2 were classified as novice users of statistical information (n = 11). Those who selected 3, 4 or 5 were classified as experienced users of statistical information (n = 84). This information was used to determine which communication strategies to present to participants.

3.4.2.2 Design

The Decision Lab branched so that those participants classified as novice users of statistical information received different communication strategies than experienced users, with the bar graph being the only format that was shown to both groups. **In contrast to the Abridged Decision Lab participants in this study were presented with all communication strategies judged appropriate to their self-reported statistical expertise.**

Novice users of statistical information

A 3x2 repeated measures design was used, with each participant being presented with two predictions (Higher Skill and Lower Skill) for each of the three Formats for novice users of statistical information (Bar Graph, Confidence Index, and Simple Table).

Experienced users of statistical information

A 4x2 repeated measures design was used, with each participant being presented with two predictions (Higher Skill and Lower Skill) for each of the four Formats for experienced users of statistical information (Bubble Map, Violin Plot, Bar Graph, and Table).

3.4.2.3 Communication strategies

The Full Decision Lab utilised the same communication strategies as the Abridge Decision Lab: Bubble Map, Violin Plot, Table, Confidence Index and Simple Table (Figures 3.2.1-6). Two versions of each Format were presented. One showing a temperature forecast for a region of the world where prediction skill is comparatively higher (Ethiopia: Higher Skill), and another for a region of the world where skill is comparatively lower (Iberian Peninsula: Lower Skill). All communications were based on the same underlying surface temperature dataset retrieved from ECOMS-UDG (https://meteo.unican.es/trac/wiki/udg/ecoms). Predictions were retrieved from System 4 (15 ensemble members) and observations from WFDEI (Weedon et al., 2014). The time period considered for these plots is 1982 to 2010. Plots are for northern hemisphere winter (December to February).

3.4.2.4 Measures

The following measures are listed in the order that they appeared for each Format. Note that the questions themselves are identical to those used in the Abridged Decision Lab

Objective understanding

Three questions measuring objective understanding of the information about likelihood and skill were common to all communication strategies. For each prediction participants were asked to indicate:

1. The probability of temperatures being above average for a particular timeframe and/or region according to the prediction
2. The probability of temperatures being below average for a particular timeframe and/or region according to the prediction
3. The skill of the prediction

In addition to this, visualisation specific questions were asked for the Bubble Map and Violin Plot. For the Bubble Map, participants were asked questions about their understanding of the spatial elements of the map (i.e. which tercile was predicted to be most likely for a specific region). For the Violin Plot, participants were asked about their understanding forecast spread. It should be noted that in scoring participants' responses to questions about likelihood, greater flexibility in what constituted a 'correct response' was permitted for the Bubble Map, Violin Plot and Bar Graph, where precise numeric values were not given.

Subjective interpretation

For both the Higher Skill and Lower Skill predictions, participants were asked three questions about their subjective interpretation of the predictions:

1. Looking at the forecast and its skill how likely do you think that it is that temperatures will be Warmer than average? *(1 = very unlikely, 10 = very likely)*
2. How confident are you in this judgement? *(1 = not confident at all, 10 = very confident)*
3. How useful do you think that this type of forecast would be for decision making in your organisation? *(1 = not useful at all, 10 = very useful)*

Subjective preference

To measure subjective preference participants were asked to rate their agreement with five statements *(1=strongly disagree, 5=strongly agree).*

1. “I like this type of [FORMAT]”
2. “I find this type of [FORMAT] easy to understand”
3. “This type of [FORMAT] provides useful information”
4. “I would use this type of [FORMAT] in my decision making”
5. “I would share this type of [FORMAT] with other people in my organisation for them to use in their decision making”

Familiarity

Familiarity was measured using level of agreement *(1=strongly disagree, 5=strongly agree)* with the statement:

“I already use this type of [FORMAT] in my work”

Liked and Disliked Elements

To gain a more detailed insight into particular elements of the communication strategies that participants may or may not like, we gave participants the chance to respond to the following (optional) open-ended questions:

1. What, if anything, do you like about this type of [FORMAT]?
2. What, if anything, do you not like about this type of [FORMAT]

Potential use in Decision Making

At the end of the study participants were asked to respond to the following open-ended question:

> If this type of map was used to show forecasts for events that your organisation is interested in, how would you use it?

3.3.2.4 Procedure

On following the invitation link to the study participants were forwarded to a Qualtrics survey where they were told about the aims of the study, and that all of the data gathered would be anonymised before being reported. Those who indicated that they wished to proceed were asked preliminary questions about their sector and statistical expertise. They were then shown a screen providing more detail about the forecasts to be shown, and an outline of what forecast skill is.

Participants were then presented with two predictions (Higher Skill and Lower Skill) for each of the 3 or 4 formats they had been assigned depending on their level of statistical expertise. To ensure that our analysis was not confounded by order effects[7] the order in which the different Formats were presented was randomised.

[7] The term ‘order effects’ refers to any instance where the order in which stimuli (in this case visualisations) are shown to experimental participants affects their responses. For instance, repetition may lead to either practice effects, whereby performance on objective measures of understanding improves over time, or fatigue effects, whereby participants attend to less information over time. By randomising the order of presentation we ensure that any difference found in participants’ responses to the different visualisation Formats cannot be attributed to the order in which they were presented.

For each Format, participants were presented first with the Higher Skill prediction, and asked to respond to questions about their objective understanding and subjective interpretation of the information provided. They were then presented with the Lower Skill prediction, and asked a comparable set of questions about objective understanding and subjective interpretation. After completing these questions, they were asked to rate their opinion of the format, note any elements that they liked or disliked, and describe how they could potentially use information presented in that format in their organisational decision making.

3.4.3 Full Decision Lab: Objective understanding

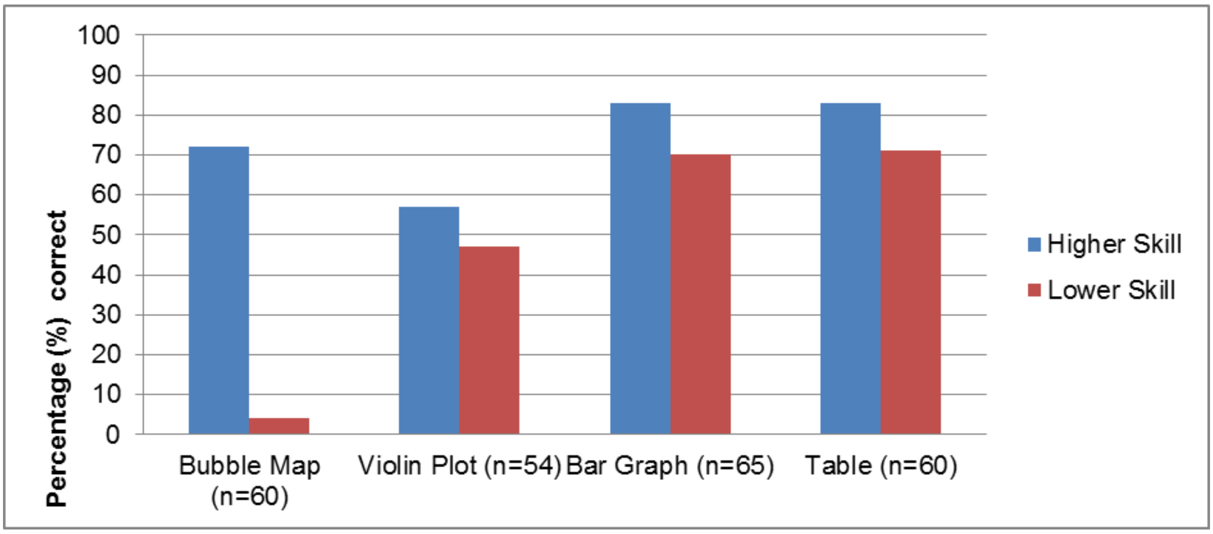
Participants' objective understanding of each prediction that they were presented with was measured by the same questions used in the Abridged Decision Lab. That is to say that for both the Higher Skill and Lower Skill versions of each type of communication strategy, they were asked to indicate a) the likelihood of warmer than average temperatures (upper tercile) according to the prediction; b) the likelihood of colder than average temperatures (lower tercile) according to the prediction; and c) the skill of the prediction. However, as this study used a within-participants rather than between-participants design, we are able to compare how the same people responded to each of the communication strategies for experienced (3.4.3.1) and novice users of statistical information (3.4.3.2). In addition to this, participants were asked visualisation specific questions about their understanding of the spatial elements of the Bubble Map (3.4.3.3), and forecast spread of the Violin Plot (3.4.3.4).

As the number of novice users of statistical information completed all questions for all communication strategies was comparatively low (*n*=7), we provide a descriptive analysis of objective understanding, but do not perform any multivariate tests as these are not supportable given the sample size. It should also noted that each ANOVA performed on the data controlled for whether participants were classified as users/potential users only, researchers, or climate service providers. However, as this was not found to affect objective understanding we do not report the coefficients here.

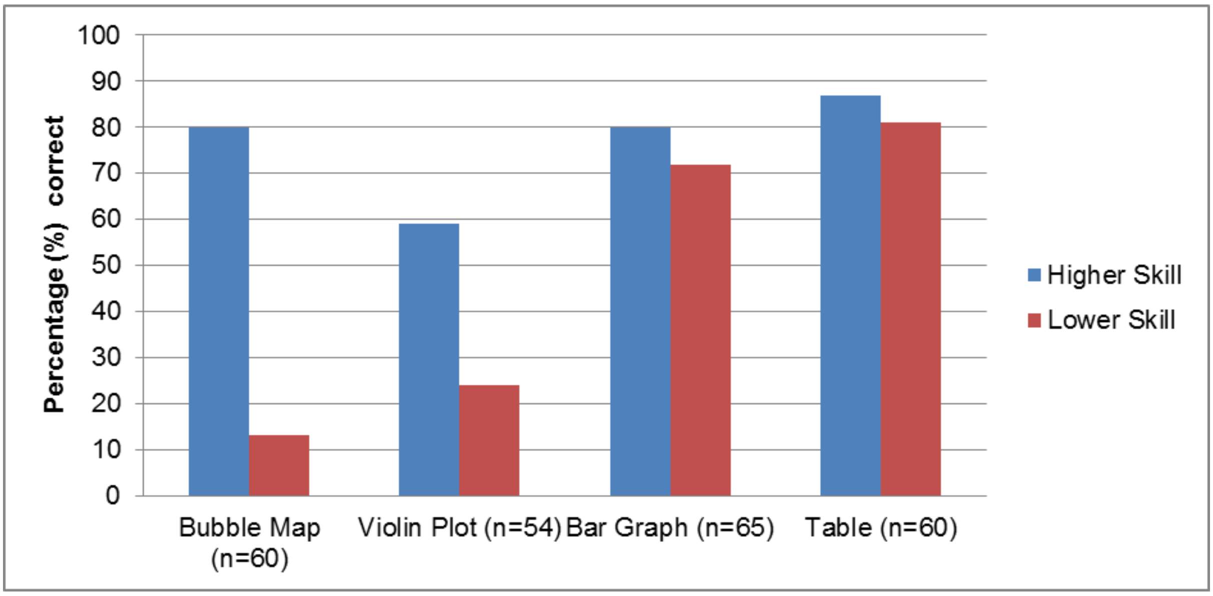
3.4.3.1 Full Decision Lab: Objective understanding of communication strategies for experienced users of statistical information

Figures 3.4.1a-c show the proportion of experienced users of statistical information who correctly answered the questions about predicted likelihood and skill for each of the communication strategies. Compared with the earlier Abridged Decision Lab, it is immediately clear that participants in the Full Lab tended to respond more accurately to all questions.

(a) Predicted likelihood of warmer than average temperatures (upper tercile)



(b) Predicted likelihood of colder than average temperatures (lower tercile)



(c) Skill

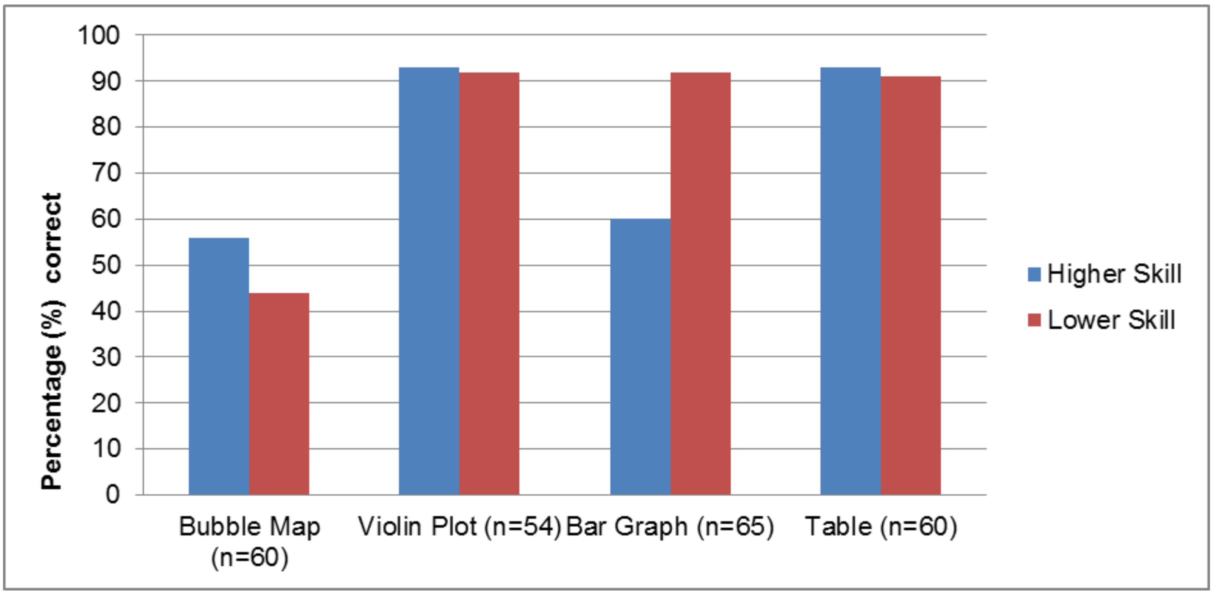


Figure 3.4.1 Proportion of experienced users of statistical information correctly answering questions about the (a) predicted likelihood of warmer than average temperatures (upper tercile), (b) predicted likelihood of colder than average temperatures (lower tercile), and (c) prediction skill, for the Higher Skill (Ethiopia) and Lower Skill (Iberian Peninsula) visualsations.

Likelihood

A glance at Figures 3.4.1a-b shows that, as was the case in the Abridge Decision Lab, participants responded to the questions about tercile likelihood more accurately when they were presented with the Bar Graph or Table than with the Violin Plot. We also find that while those presented with the Bubble Map tended to answer correctly when shown the Higher Skill (Horn of Africa) prediction, but incorrectly when shown the Lower Skill (Iberian Peninsula) prediction. Indeed, we see that for all four communication strategies participants responded more accurately to the questions about predicted likelihood when shown the Higher Skill forecast.

A pair of 2x4 repeated measures ANOVAS confirmed that these differences were statistically significant. An overall effect of Higher versus Lower Skill was found for both the “likelihood of warmer than average temperatures (upper tercile)” (*F(1, 35)*=24.6 , *p*<.001) and “likelihood of lower than average temperatures (lower tercile)” questions (*F(1, 36)*=29.9, *p*<.001). A main effect of Format was also found (Likelihood of warmer than average temperatures: *F(3,105)*=12.2, *p*<.001); Likelihood of colder than average temperatures: *F(3,108)*=19.4, *p*<.001) along with an interaction between Format and Skill Level due to responses to the Bubble Map differing particularly strongly between the Higher Skill and Lower Skill predictions (Likelihood of warmer than average temperatures: *F(3,105)*=11.3, *p*<.001); Likelihood of colder than average temperatures: *F(3,108)*=10.1, *p*<.001). Pairwise comparisons (Bonferroni adjusted) showed that participants responded more accurately to the Bar Graph and Table than the Bubble Map and Violin Plot.

In Summary
When it came to correctly interpreting information about the predicted likelihood of different terciles we find that the Bar Graph and Table elicited the highest proportion of correct responses overall. We also find that judgements were more accurate for the Higher Skill visualisation than the Lower Skill visualisation. This was especially pronounced when it came to the Bubble Map, where correct responses were high for the Higher Skill visualisation and very low for the Lower Skill visualisation. Once again, this suggests that many participants did not correctly interpret what the presence of 'white space' on the Bubble Map meant (i.e. that all terciles should be considered equally likely).

Skill

When it came to correctly interpreting information about prediction skill, we see that accurate responses were highest for participants who were shown the Table and Violin Plot and lowest for those shown the Bubble Map. Unlike the Abridged Decision Lab, where participants tended to perform worse on this measure when shown the Lower Skill predictions, we did not find this in the Full Decision Lab.

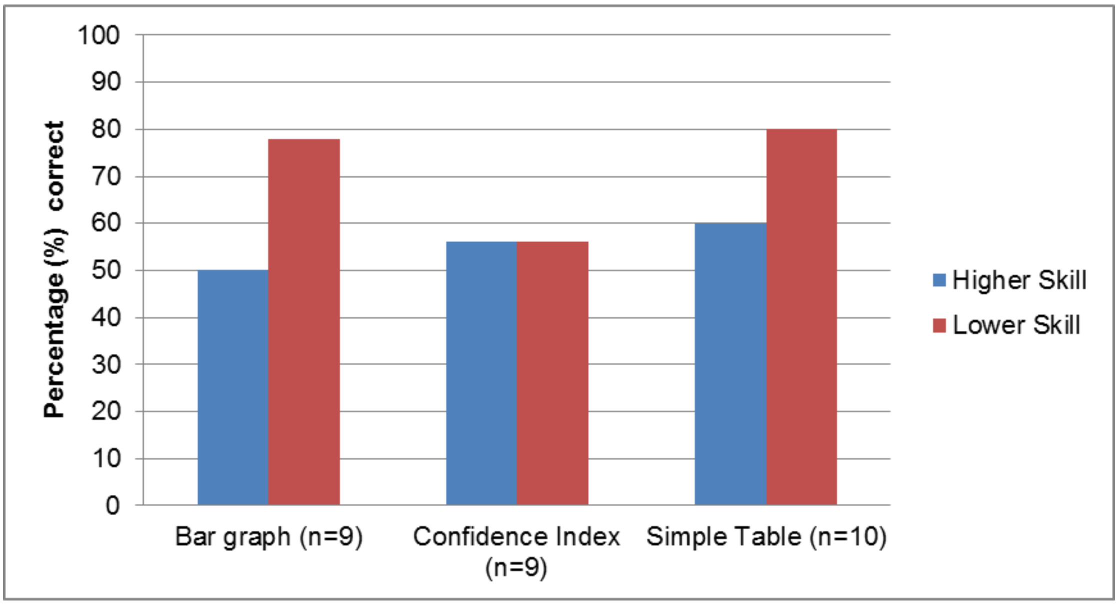
A 2x4 repeated measures ANOVA confirmed that Format had a significant effect on the accuracy of responses to the questions about skill (*F(3, 117)*=27.8, *p*<.001). It also indicated the existence of an interaction between Format and Skill Level (*F(3, 117)*=4.8, *p*<.01); a consequence of the Lower Skill Bar Graph eliciting a greater proportion of correct responses than the Higher Skill Bar Graph. Pairwise comparisons (Bonferroni adjusted) indicated that, when it came to questions about skill, participants responded less accurately to the Bubble Map than the Table (*p*<.001), Violin Plot (*p*<.001) and Bar Graph (*p*≤.01). Overall responses were however less accurate for the Bar Graph than the Violin Plot (*p* ≤ .05) or the Table (*p* ≤ .05).

In Summary
We find that participants in the Full Decision Lab interpreted information about skill most accurately when it was presented as a single RPSS score for each forecast (i.e. as was the case in the Table and Violin Plot). When it came to the Bar Graph, participants were considerably more accurate in the assessment of Skill when it came to the Lower Skill visualisation (where skill was negative for all three terciles) than the Higher Skill visualisation (where skill varied from "some" to "good" between terciles). This suggests that the presence of multiple skill scores can make appropriate interpretation more difficult.

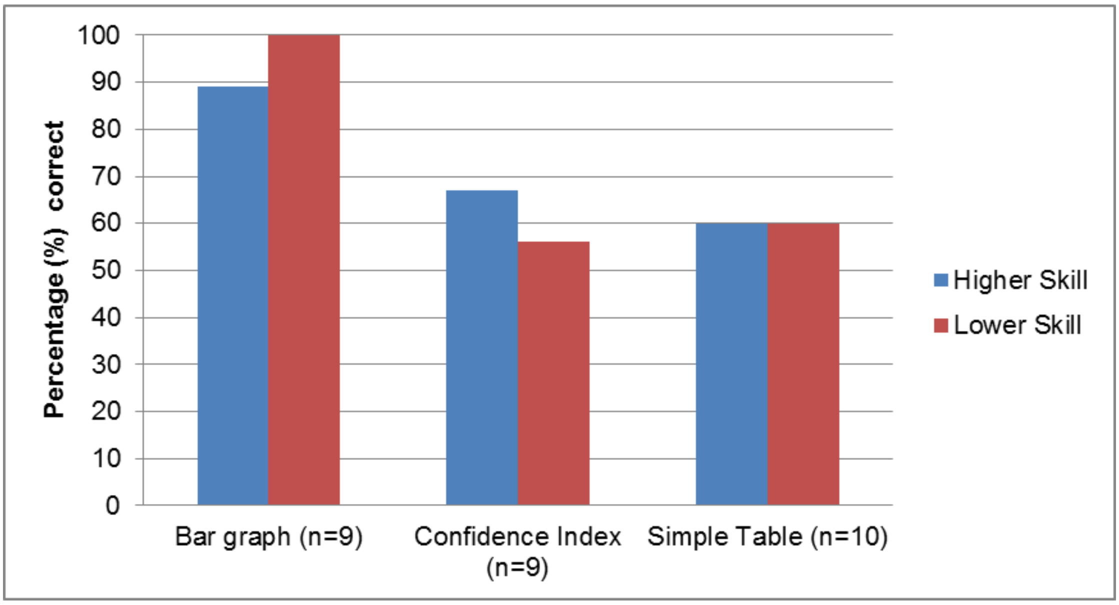
3.4.3.2 Full Decision Lab: Objective understanding of communication strategies for novice users of statistical information

For novice users of statistical information, Figures 3.4.2a-c show the proportion who answered each of the questions about likelihood and skill correctly. Unlike the Abridged Decision Lab, we do not see a clear pattern in participants' objective understanding of the communication strategies. This is likely to be due to the fact that only a small proportion of participants could be classified as novice users of statistical information.

(a) Predicted likelihood of warmer than average temperatures (upper tercile)



(b) Predicted likelihood of colder than average temperatures (Lower tercile)



(c) Skill

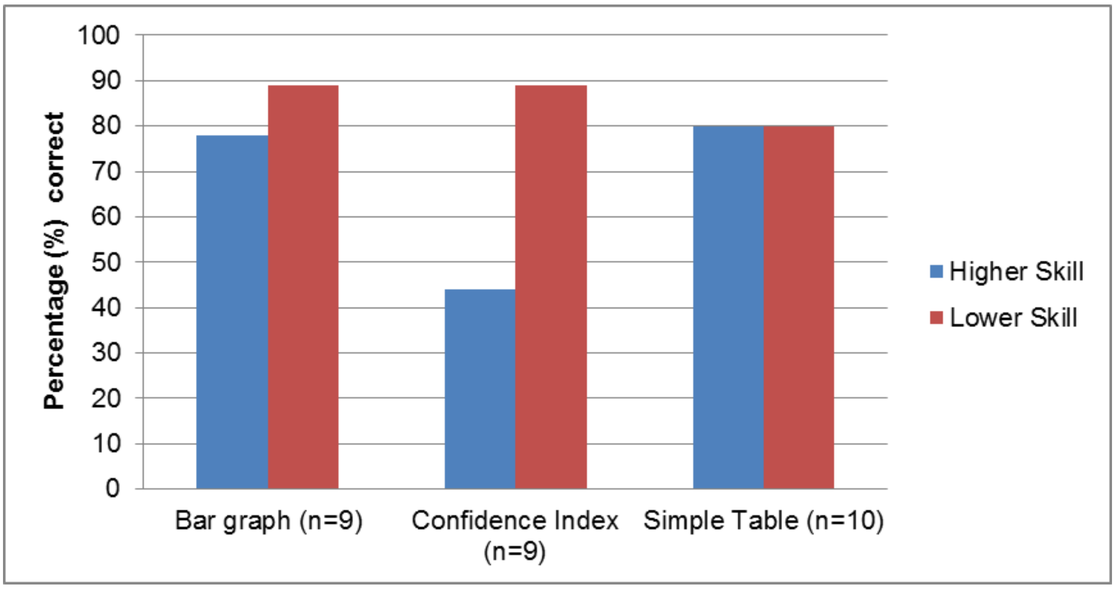


Figure 3.4.2 Proportion of novice users of statistical information correctly answering questions about the (a) predicted likelihood of warmer than average temperatures (upper tercile), (b) predicted likelihood of colder than average temperatures (lower tercile), and (c) prediction skill, for the Higher Skill (Ethiopia) and Lower Skill (Iberian Peninsula) visualsations.

3.4.3.3 Objective Understanding: Bubble Map

Spatial Elements

The proportion of respondents who correctly answered questions about the spatial information provided on the map was 52% for the Higher Skill map (Horn of Africa) and 65% for the Lower Skill Map (Iberian Peninsula). This was slightly higher than that observed in the Abridged Decision Lab. An evaluation of incorrect responses suggested that, in the former case, errors were likely due to overgeneralization in the case of the Higher Skill prediction (i.e. suggesting that warmer than average temperatures were predicted across all areas, when there were a few areas where average temperatures were predicted); and difficulty determining the colour of faintly shaded regions in the Lower Skill prediction.

General Comments

Once again, participants were found to be better at extrapolating information about tercile likelihood for the Higher Skill prediction than the Lower Skill prediction. That is to say that very few participants recognised that white space indicated that all terciles should be considered equally likely (due to either the absence of skill of the fact that the prediction does not favour a particular tercile). This indicates that the meaning of white space requires greater explanation, and that a key detailing what different sizes of Bubbles represent may aid interpretation.

3.4.3.4 Objective Understanding: Violin Plot

Range

When it came to the questions about the forecast range depicted in the Violin Plots the rate of correct response was higher in the present study than the Abridged Lab (Higher Skill prediction = 65%; Lower Skill prediction = 63%).

General Considerations

While performance on the likelihood questions was better in the Full Lab than the Abridged Lab, it was still lower than that for other communication strategies. Once again this indicates that despite the ensemble members being colour coded, this format is less useful for communicating precise tercile probabilities than other communication strategies, and is more suited for instances where information about forecast spread is required. It should however be noted that most participants who were shown this visualisation, were able to correctly interpret and categorise the skill score (i.e. as no skill, some skill, good skill.)

3.4.3.5 Objective Understanding: Bar Graph

Like other communication strategies, the information presented in the bar graph appears to have been more accurately interpreted by the participants in this study than in the Abridged Decision Lab, especially with respect to information about likelihood. However, it should be noted that an inspection of incorrect responses strongly suggested that some participants were mistaking the skill values underneath each column for tercile probabilities. Additionally, the presence of three different scores (ROCSS for each tercile) appears to have led to some misinterpretations where levels of skill differ between terciles. Taken together, this suggests that skill scores on this visualisation should be repositioned, and evaluative categories (e.g. No Skill, Some Skill, and Good Skill) made clearer.

3.4.3.6 Objective Understanding: Table

Overall, this was the communication strategy that experienced users of statistical information interpreted the most accurately with respect to likelihood and skill. In contrast to the Abridged Decision Lab, an overwhelming majority of participants in this study correctly identified that a negative skill score indicates that there is no skill.

3.4.3.7 Objective Understanding: Confidence Index

As the present sample contained few novice users of statistical information, we are unable to provide a full assessment of objective understanding. However, the pattern of responses obtained suggested that some participants made inferences about the ‘likelihood of colder than average temperatures’ based on the stated likelihood of warmer than average temperatures, rather than conclude that this information was not provided. As was the case in the Abridged Lab, this suggests that while this communication strategy may suit situations where a binary split exists using it to represent the likelihood of particular terciles may lead to misinterpretation.

3.4.3.8 Objective Understanding: Simple Table

Once again, the fact that the present sample contained relatively few novice users of statistical information means that a full assessment of this communication strategy is not possible in the present study. However, responses did suggest that – as was the case in the Abridged Decision Lab – it was generally well understood.

3.4.4 Full Decision Lab: Subjective Perception

As was the case in the Abridged Decision Lab, participants were asked to rate a) how likely they thought that warmer than average temperatures were for a time period, given the information about likelihood and skill provided in the forecast; and b) how confident they were in this judgement. Once again, the absence of skill in the Lower Skill predictions should mean that, from a normative standpoint, judgements about the likelihood of warmer than average temperatures should not substantially differ between the different Lower Skill predictions. We see that, compared to participants in the Abridged Decision Lab, those participants classified as experienced users of statistical information demonstrated less variability in their judgements of likelihood (Table 3.4.2). There are two possible reasons for this difference between the two studies. The first is the fact that these participants are more engaged with climate information, which means that they are better able to calibrate their subjective interpretation with objective information. The second is that, as participants in this study were presented with all four communication strategies, they may have been aware of a need for consistency when it came to their interpretation of the ‘No Skill’ forecasts.

Table 3.4.2 Mean (SD) subjectively perceived likelihood of temperatures being warmer than average for the Higher Skill and Lower Skill predictions amongst experienced users of statistical information

	Bubble Map	Violin Plot	Bar Graph	Table
Higher Skill (Ethiopia)	7.8 (1.2)	7.1 (1.5)	8.2 (1.7)	8.0 (1.3)
Lower Skill (Iberian Peninsula)	2.5 (1.6)	3.6 (1.6)	3.0 (1.7)	3.0 (1.4)

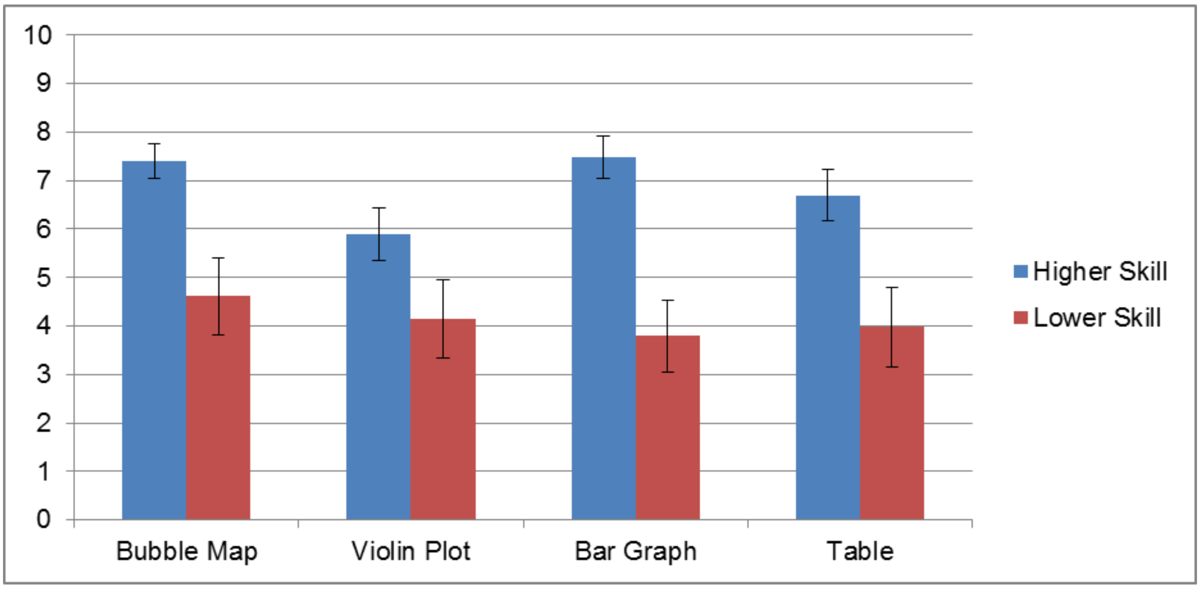
Amongst novice users of statistical information however, we once again find that when it comes to the Lower skill predictions, subjective judgements of likelihood are higher for the Simple Table than the other formats; despite there being no skill for any of these predictions (Table 3.4.3). This suggests that participants were unduly influenced by the predicted probabilities displayed in the Lower Skill predictions.

Table 3.4.3 Mean (SD) subjectively perceived likelihood of temperatures being warmer than average for the Higher Skill and Lower Skill predictions amongst novice users of statistical information

	Bar graph	Confidence Index	Simple Table
Higher Skill (Ethiopia)	7.1 (2.2)	6.9 (1.5)	6.8 (2.3)
Lower skill (Iberian Peninsula)	2.8 (1.8)	3.3 (1.6)	5.2 (2.2)

When asked about their confidence in these subjective judgements, we find that, as one might expect, confidence was greater for the Higher Skill than the Lower Skill predictions (see Figures 3.4.3a-b). This was confirmed by a pair of 2x4 repeated measures ANOVAS, which found that this difference was statistically significant for both experienced (*F(1,41)*=66.3, *p*<.001) and *novice* (*F(1,5)*=6.4, *p*=.05) users of statistical information*.* For experienced users, it was also found that confidence was higher for those presented with the Bar Graph (*p*=.001) and Bubble Map (*p*=.004) than the Violin Plot.

(a) Mean confidence in subjective judgements of likelihood amongst experienced users of statistical information



(b) Mean confidence in subjective judgements of likelihood amongst novice users of statistical information

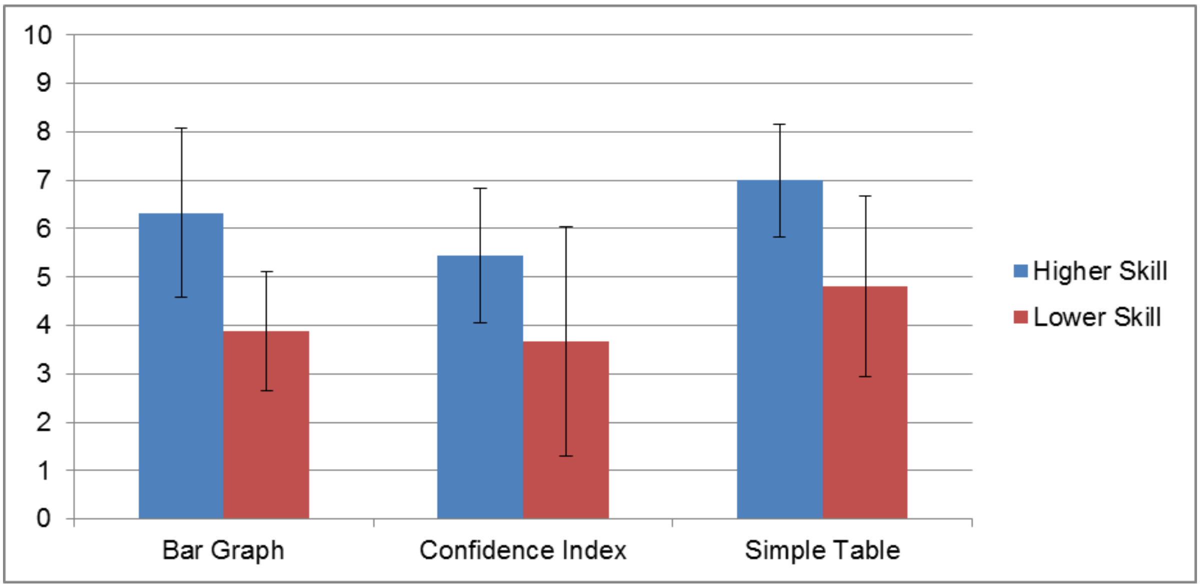


Figure 3.4.3 Mean confidence in subjective judgement of likelihood (1=Not confident at all, 10=Very confident) amongst those with (a) experienced and (b) novice users of statistical information. Error bars represent 95% confidence interval around the mean.

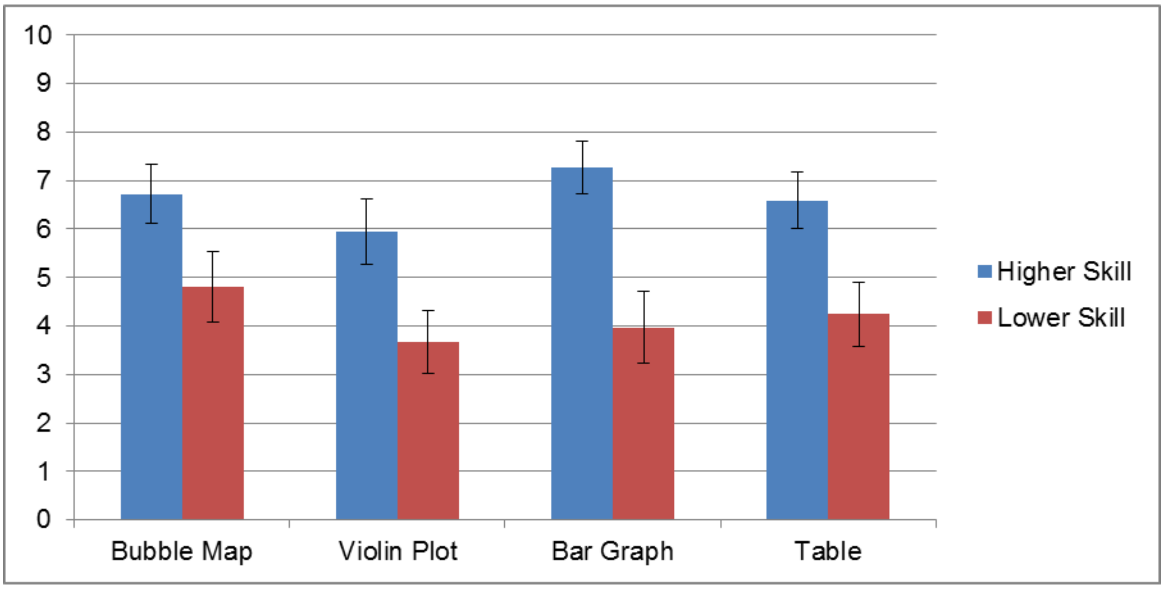
In Summary

We find that in contrast to the Abridged Decision Lab, experienced users of statistical information were largely consistent in their subjective judgements of likelihood across communication strategies. On the other hand, the likelihood judgements of novice users of statistical information failed to discount the predicted probabilities contained within the 'no skill' predictions. However, in all cases reported confidence in subjective judgements of likelihood were substantially lower for Lower Skill predictions than for Higher Skill predictions.

3.4.5 Full Decision Lab: Perceived usefulness

Participants were asked to rate how useful they felt that each type of predictions they were presented with would be for decision making within their organisation. Figures 3.4.4a-b below show mean ratings perceived usefulness for each of the Higher Skill and Lower Skill predictions. Unsurprisingly, participants rated the Higher Skill visualisations as more useful than the Lower Skill visualisations; with this difference reaching statistical significance for experienced (*F(1,41)*=59.5, *p*<.001) but not novice users of statistical information (*F(1,6)*=3.8, *p*=.10); although failure to find a significant effect in the latter case is likely due to the low number of novice users in the sample.

(a) Mean perceived usefulness of visualisations amongst experienced users of statistical information



(b) **Mean perceived usefulness of visualisations amongst novice users of statistical information**

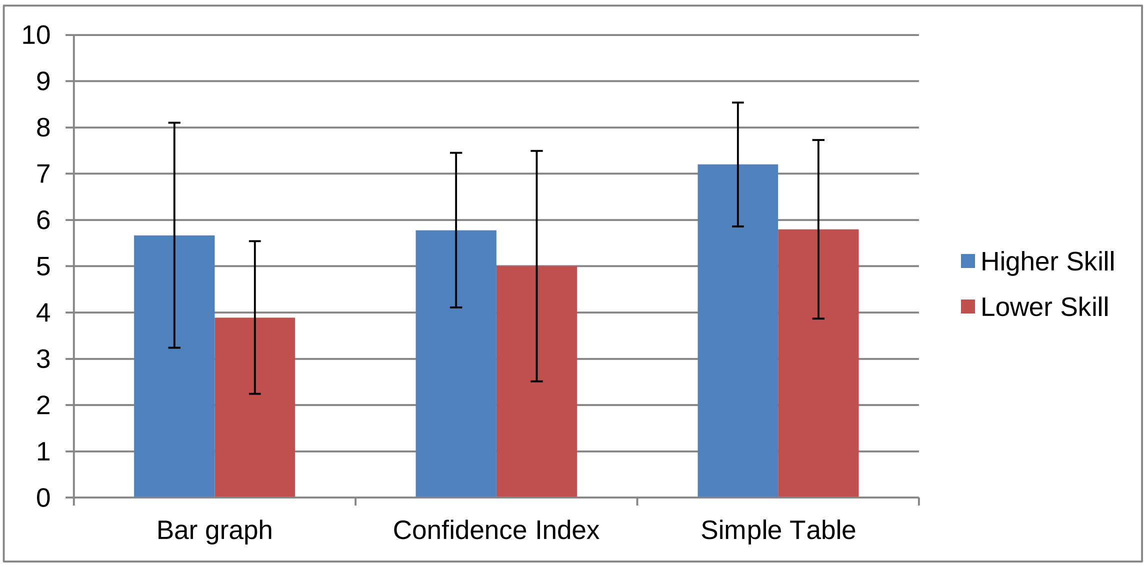


Figure 3.4.4 Mean perceived Usefulness of each of the communication strategy for decision making in one's organisation for (a) experienced and (b) novice users of statistical information. Error bars represent 95% confidence interval around the mean.

In Summary
In Summary, participants tended to perceive Higher Skill predictions as more useful than Lower Skill predictions. This was not affected by the communication strategy used.

3.4.6 Full Decision Lab: Preference and Familiarity

As was the case in the Abridged Decision Lab, a Preference score was calculated by taking the mean of the five Subjective Preference items described in Section 3.4.2, while familiarity was measured by level of agreement with the statement: "I already use this type of [FORMAT] in my work". Mean Preference and Familiarity scores are reported for experienced and novice users of statistical information in Table 3.4.4 and Table 3.3.5 respectively.

Experienced users of statistical information

For experienced users of statistical information mean Preference and Familiarity ratings were slightly higher for the Bar Graph and Table than the Bubble Map and Violin Plot. Repeated measures ANOVAs indicated that there was no significant effect of Format on Preference (*F(3,111)*=1.9, *p*=.14). However, a small effect of Format on Familiarity was found (*F(3, 111)*=1.9, *p*=.01), with post-hoc Bonferroni tests showing that the Bar Graph was rated as significantly more 'Familiar' than the Bubble Plot (*p*=.02).

Table 3.4.4 Mean (standard deviation) ratings of Preference and Familiarity amongst experienced users of statistical information

	Bubble Map Mean (SD)	Violin Plot Mean (SD)	Bar Graph Mean (SD)	Table Mean (SD)
Preference	3.3 (1.2)	3.2 (1.1)	3.5 (0.9)	3.6 (0.9)
Familiarity	1.8 (1.1)	1.8 (1.3)	2.3 (1.4)	2.0 (1.1)

In Summary

Ratings of preference did not significantly differ between the four different types of communication strategy, although the Bar Graph was rated as being more familiar than the Bubble Map.

Novice users of statistical information

Amongst novice users of statistical information the Simple Table received the higher Preference Rating, while the Bar Graph was received a slightly higher Familiarity rating than the Confidence Index. Repeated measures ANOVAs indicated that neither difference reach statistical significance (Preference: (*F(2,14)*=1.7, *p*=.21); Familiarity: (*F(2, 14)*=2.9*, p*=.11). Once again however we should be cautious about drawing firm inferences from this due to the small number of novice users in the present sample. It is also worth keeping in mind that the pattern of response is similar to that observed in the Abridged Decision Lab.

Table 3.4.5 Mean (standard deviation) ratings of Preference and Familiarity amongst novice users of statistical information

	Bar Graph Mean (SD)	Confidence Index Mean (SD)	Simple Table Mean (SD)
Preference	2.9 (1.1)	2.6 (1.1)	3.4 (0.5)
Familiarity	2.0 (1.5)	1.3 (1.1)	1.5 (0.8)

In Summary

Amongst novice users of statistical information differences in ratings of preference and familiarity between the three types of communication strategy did not reach statistical significance. However, as our sample for this Decision Lab contained few novice users we may have simply had too few participants in the category to detect an actual difference. In keeping with the Abridged Decision Lab however we do see a trend for participants to prefer the Simple Table and perceive the Bar Graph as more familiar.

3.4.6 Full Decision Lab: Does Preference Correspond with greater Familiarity and Objective Understanding?

To examine whether participants preferred those Formats that were a) more familiar; and b) best understood, a set of multiple linear regression analyses were performed in which Preference was entered as a criterion (dependent) variable, and Familiarity and Objective understanding (sum of questions answered correctly) as predictors. Owing to the low number of novice users of statistical information in the sample, these analyses were performed for the formats for experienced users of statistical information only. As can be seen from Table 3.4.6, we find that while a significant or marginally significant positive association between Familiarity and Preference existed for three out of the four communication strategies, no link between Preference and Objective Understanding was in evidence.

Table 3.4.6 Linear regression analyses examining the extent to which Objective Understanding and Familiarity predict Preference (Unstandardized B and Standardised β coefficients reported)

	Bubble Map		Violin Plot		Bar Graph		Table	
	B (SE)	*β*	*B (SE)*	*β*	*B (SE)*	*β*	*B (SE)*	*β*
Familiarity	.29 (.14)	.28*	.24 (.13)	.29 †	.22 (.08)	.36**	.06 (.10)	.07
Objective Understanding	-.02 (.11)	-.02	.23 (.14)	.26	.01 (.09)	.01	.08 (.08)	.12
ANOVA	2.3		2.27		3.68**		.45	
R^2	.08		.10		.13		.02	

†Marginal significant at *p*≤.10 *Significant at *p*≤.05 **Significant at *p*≤.01

In Summary

Amongst the highly engaged sample of participants who took part in the Full Decision Lab, familiarity with particular formats was less strongly associated with preference than was the case in the Abridged Decision Lab. However, we did not find that preference corresponded with objective understanding.

3.5 Responses to open ended questions

For each communication strategy shown, participants in the full decision lab were presented with three open ended questions asking them to detail 1) what, if anything, they liked about the formats (optional); 2) what, if anything, they disliked about the formats (optional); and 3) how, if at all, they would use the formats in their decision making. Here, we discuss responses to these opened-ended questions for the Bubble Map, Violin Plot, Bar Graph and Table. The Confidence Index and Simple

Table are not included here as very few participants responded to the open-ended questions for these communication strategies.

Responses were coded based on common characteristics. Interestingly, for each communication strategy there was greater consensus amongst respondents about the characteristics they ‘liked’ about the formats than the characteristics they ‘disliked’. It was also evident from responses to questions about potential use in decision making that, where formats were thought to be potentially useful, many participants would use them in conjunction with other formats and types of information.

While many responses were specific to the individual communication strategies, it should be noted that some participants made general statements about 1) only being able to use forecasts where skill was sufficiently high; 2) not being interested in the variable (temperature) used in the examples presented; and 3) not being interested in the regions depicted. As the formats presented to participants were chosen on the basis that they could be used to represent different variables and metrics for different regions, the latter two points are not necessarily a barrier to use. The former point however highlights the importance of making skill salient. With respect to skill, it was also found that participants differed in their preferences for receiving this information: with some indicating a preference for formats where a single score was presented separately for the whole forecast, and others a preference for separate scores for each tercile.

3.5.1 Bubble Map

Aspects liked

Of the 59 participants who completed the questions for this visualisation, 39 participants chose to respond to the optional question “What, if anything, do you like about this map”. The most frequently mentioned characteristic was the **spatial element of the map** (*n*=16) (e.g. *“You get spatial information about the forecasted values”*), followed by **general statements about the information content of the visualisation** (*n*=14) (e.g. “*It obliges me to think about interpreting different layers of information”),* specific statements about the **combined presentation of likelihood and skill** (*n*=6) (e.g. *“Combination of forecast and skill in the same map is very useful”*), **ease of understanding** (*n*=5) (e.g. *”It is relatively simple to understand with simply three different outcomes (colder, normal, warmer)”)* and comments on the **graphical characteristics of the visualisation** (*n*=5) (e.g. “*Colours can help in understanding information at first sight*”).

Aspects disliked

There was less consensus amongst participants about which elements they disliked. However, some common themes did emerge. While some participants found the Bubble Map easy to understand, and appreciated the way that skill and likelihood were combined, others reported finding it **difficult to interpret**, or expressed concerns about it being too complex (*n*=12):

> *"Too confusing! Joining, spatial, likelihood and skill variables in the same forecast makes it incredibly cryptic and hard to derive information from."*

With respect to more specific characteristics, 5 participants mentioned that they found it **difficult to ascertain what the size of the bubbles indicated in terms of precise probabilities** (e.g. *"…diameter info is only described in text, would be better to have it displayed in graphic as well"*), while 4 mentioned the presence of **'white space'** on the Lower Skill map (e.g. *"Blank/empty maps can be a bit confusing"*). Some participants also indicated that they found the map **difficult to read when skill was low**:

> *"Le bleu pale est difficile à distinguer du gris pale".* Translation: *"The pale blue is difficult to distinguish from the pale grey"*.

> *"Difficult to read if there are small dots or not."*

Other participants indicated a preference to see all three terciles, and receive numeric information about skill.

Potential use in decision making

When asked to describe its potential uses in decision making 22 indicated that they would not use it in their decision making. Of those who felt that the visualisation was potentially useful 7 stated that they would use it in **conjunction with other information**, with some suggesting that it could be used to provide a **general 'qualitative' overview**, to be used in conjunction with more quantitative information:

> *"Qualitative support to decision making based on more quantitative results (perhaps extracted from this map)"*

> *"It would be included in interpretation of general conditions. It could extend the current forecast range."*

Other uses included **communication with others**, and **planning and scheduling** (e.g. *"Energy consumption forecasting"* Sector: Energy). However, only 3 participants mentioned that they would use it to make specific decisions.

In Summary

- The most widely valued characteristic of this visualisation was the fact that it provides spatial information.
- The integration of likelihood and skill was appreciated by some participants, but disliked by others, who indicated a preference for this information to be presented separately.
- The presence of a legend indicating what different sizes of bubbles represent with respect to likelihood would aid interpretation.
- Some participants found it difficult to extract information from the 'Lower Skill' map due to the paleness and small size of the bubbles.
- With respect to decision making, responses suggest that this format could be used to supplement numeric information.

3.5.2 Violin plot

Aspects liked

When participants were asked about what they liked about the Violin Plot, the most common response was the **comprehensiveness of the information** provided in the visualisation (*n* = 16). More specific elements mentioned included the fact that it showed the **distribution of the forecast** (*n* = 8) (e.g. *"It gives a good impression of the spread of the forecasts"*), as well as **individual ensemble members** (*n* = 4) (e.g. *"ensemble members are shown"*), and **climatology** (*n* = 2) (*"You get explicit information about the climatology…"*)

Aspects disliked

While the comprehensiveness of the information provided in this visualisation was cited as its most well liked feature, the amount of information presented meant that participants frequently found it to be **difficult to understand and overly complex** (n=17). Some participants commented on the fact that that it was **difficult to extract information about precise likelihoods from the plot** (e.g. *"Dificulta la estimación de la probabilidad"* translation: *"Difficult to estimate the probabilities"*), with the size of the dots representing each ensemble member making them difficult to count (*"counting small circles is needed to determine probabilities, which is difficult to achieve"*). One participant noted that this would become more difficult with a greater number of ensemble members (*"It is difficult to read, particularly with more ensemble members than 15"*)

Potential use in decision making

Of the 46 participants who completed the questions for this visualisation 23 stated that they would not use this visualisation in their decision making. Of the remainder, potential uses varied, with the most commonly cited uses being **combination with other forms of information** (*n*=4) and **deciding on specific actions to take** (*n*=3). It is worth noting that in the latter case all three participants referenced decisions related to hydrology and hydro-electricity. Other potential uses included: **planning and scheduling** (*n*=2), **illustrating confidence/variance** (*n*=2), **risk analysis** (*n*=1) and **deriving warnings** (*n*=1).

While most participants who stated that they would not use the visualisation did not elaborate this point, some did provide reasons. These included: the visualisation not displaying metrics of interest, the level of skill, and the fact that the visualisation does not include spatial information.

In Summary

- While some participants liked the amount of detailed information provided, many disliked the complexity of the visualisation.
- The fact that this visualisation presented information about the forecast range rather than just the probability of terciles was valued by some participants.
- Participants' comments on the difficulty of extracting tercile probabilities from this visualisation suggest that while this style of visualisation may be useful for showing forecast spread, it is less useful when precise probabilities are required.
- While relatively few participants mentioned that they would use this style of visualisation in making decisions about specific actions, those who did detailed decisions related to hydrology and hydro-electricity.

3.5.3 Bar Graph

Aspects liked

The most commonly cited positive aspect of the Bar Graph was **ease of understanding** (*n*=19) (e.g. *"It is simple , you have plenty information in a single graph"*), followed by the inclusion of **information about skill** (*n*=9) (e.g. *"skill value and clear skill value explanation"*, the inclusion of **information about tercile likelihood** (*n*=5) (e.g. *"It summarises different information in only one graph"*), **perceived comprehensiveness of the information provided** (*n*=5), and **the visual (rather than numeric) nature of the format** (*n*=5) (e.g. *"visually it is easier to understand than a table"*). Two participants also expressed a **preference for tercile formats** (e.g. "*terciles are simpler than many categories"*), while one commented favourably on the inclusion of the climatology line (e.g*. "La línea horizontal hace fácil la comparación y permite valorar el valor añadido que aporta la predicción"* translation: *"The horizontal line makes comparisons easy and enables the assessment of the added value of the prediction"*).

Aspects disliked

The aspects of this visualisation disliked by participants varied. Some found the visualisation **difficult to interpret** (*n*=9), while others felt that it was **overly simple** *(n*=3). Multiple participants commented found the graph **visually unappealing** (*n*=4) (e.g. *"I'd like to see more colour"*), **did not like the way that information about skill was presented** (*n*=4) (e.g. *"three skill scores rather than one aggregate"),* or **disliked the absence of spatial information** (*n*=4). Others expressed a **dislike of the tercile format** in general (*n*=2), and commented that **precise probabilities were difficult to discern** (*n*=3) (e.g.*" The actual probabilities are harder to detect. I preferred a table in which I am given the percentage of likelihood better…"*).

One participant commented on the potential for skill scores to be confused with probabilities, something that was supported by our analysis of participants' objective understanding of this visualisation:

> *"The probability and skill score numbers are likely to get mixed up and confused by most users. Not very easy to explain to most users of climate information"*

Potential use in decision making

Of the 66 participants who answered questions on this format, 22 indicated that they would not use it in decision making. While most did not give reasons for this, one participant commented that they preferred the precise numeric values offered by the table *("I would be less likely to use than a table. If I were to share the data with my organisation, I would prefer the actual numbers."*), while another indicated that they felt that other formats were more useful for communication *("I do not think I would use this presentation, since I consider the others more informative and/or easier to understand").*

Of those remaining uses included **planning** (*n*=7), **communication with others** (*n*=7), **combination with other information** (*n*=6), **deciding on specific actions to take** (*n*=4), **anticipating consumer demand** (*n*=1) and **marketing** (*n*=1). Three participants stated that they **already used this type of format**.

In Summary

- While a large proportion of participants reported finding this visualisation easy to understand (and in some cases overly simple), others found it difficult to interpret. This further underscores the fact that different formats are required by users with different levels of statistical expertise.
- Concerns about the potential conflation of likelihood and skill raised by some participants are supported by our assessment of objective understanding.

3.5.4 Table

Aspects liked

The most frequently mentioned positive characteristic of the table was **ease of understanding** (*n*=22). Favourable comments were also made about the way that **skill** (*n*=9) and **likelihood** (*n*=6) were represented (e.g. *"The direct linkage of the forecast with its skill (for this season"*), along with the **spatial component of the table** (*n*=2) (e.g. *"Good, if information for a special location is needed"*).

Aspects disliked

The most common type reason for disliking this format was a perceived **lack of visual appeal** (*n*=6), with some participants indicating that they would prefer a graphic or map, or that the table should be presented alongside one of these (e.g. *"They should be used with graphical element…"*). Others commented that the **spatial element was too limited** (n=6) or too specific (*n*=1) for their needs. A minority of participants also indicated that they found the table **difficult to understand,** or felt that less experienced users would find it so (*n*= 4). It was also remarked that while this format contained information about tercile likelihood and skill, it did not provide other information that would be of interest to users: (e.g. *"…it could not be useful for other purposes that do not require probabilities and skill. It is only for a certain city, not for a region."*)

It should also be noted that while some participants responded favourably to the way that skill was represented, others expressed concern that **skill scores were not salient enough to prevent over-interpretation of probabilistic information** (e.g. *"Seeing 0% or 100% in the table gives a very strong message while skill score is not that high. Leaves room for over-interpretation."*). This concern is supported by our analysis of participants' subjective interpretation of the Lower Skill forecasts, indicating a need to increase the salience of skill in this type of format. One participant indicated that they would **prefer that skill be represented using evaluative categories** (*"The skill can be presented in an easier way (by adding no skill, some/very little skill, high/very high skill)"*).

Potential use in decision making

Of the 56 participants who completed the questions for the Table 20 said that they would not use this format. As with other communication strategies, potential uses for the table included **communication with others** (n=9), **combination with other information** (n=4), planning (n=3) **deciding on specific actions to take** (*n*=3), and **marketing** (*n*=1). One participant in the climate services meteorology sector indicated that they would consider presenting the table alongside a map (*"It might be added to seasonal forecast maps"*).

In Summary

- This format is widely perceived to be easy to use by those with Greater Statistical Experience.
- While information about probability and skill was considered to be clear by many participants, others note that the table does not provide any information beyond this, and may be of limited value in isolation.
- Some participants disliked the numeric format, and it was suggested that the table should be presented with a map or graph.
- Concerns were expressed by some participants about the potential over-interpretation of likelihood information. These are substantiated by our analysis of participants' subjective interpretation of the information, and thus point to a need to increase the salience of skill.

3.6 General Discussion

The goal of the studies reported here was to examine objective understanding of and subjective preference for six types of communication strategy developed to communicate levels of confidence in S2D predictions. To do this we conducted two Decision Labs: an Abridged Lab with a broad sample of European decision makers in relevant sectors, and an extended Full Lab one with a smaller, more specialised sample of participants with a high level of engagement with climate information. In this section we discuss general findings regarding participants' objective understanding (3.6.1), subjective interpretation (3.6.2) and preferences (3.6.3), before focussing on key findings regarding each of the formats (3.6.4)

3.6.1 Objective Understanding

Looking at participant responses in the two studies we find that those in the Full Decision Lab tended to demonstrate better understanding than those in the Abridged Decision Lab. In addition to possible 'practice effects' arising from the fact that those in the former group saw all communication strategies appropriate for their level of statistical experience, it is very likely that this was due to their greater existing engagement with climate information. This highlights the fact that, even where recipients of predictions are comfortable with using statistics, those who have previously been less engaged with climate prediction (and indeed climate information more generally) may find it considerably more difficult to accurately interpret this information. Overall it appears that, even when allowances are made for the fact that some communication strategies don't present likelihoods as precisely as others (e.g. by permitting participants a larger range of correct responses for the Bubble Map, Violin Plot and Bar Graph than the Tables), assessments of predicted likelihood were still more accurate when they were numerically represented rather than graphically represented.

When it came to understanding of skill, we find that both experienced and novice users of statistical information struggled to accurately interpret information about skill when different skill scores were given for different terciles. That is to say, that when ROCSS showed ‘good’ skill for some terciles and ‘some’ skill for others, a large proportion of participants failed to recognise this. Hence, it appears that there is a trade-off between providing the additional insight that comes from having separate skill scores for each quantile (e.g. in situations where the prediction performs far better for some than for others), and providing a single score which offers less nuance but may be easier for some users to interpret. Other issues highlighted by participants’ responses to the questions about forecast skill indicated that a) skill scores can be mistaken for likelihoods; and b) where skill is negative participants may not always recognise that this means that there is no skill (i.e. failure to correctly interpret what the presence of a minus sign means. To address the former problem, it is to be recommended that skill scores not be placed in a position on graphs, where such confusion is likely to happen. In the present case, it appears that putting skill scores directly under bars on the Bar Graph led to this misinterpretation. To address the latter problem, it is suggested that negative skill scores should either be replaced or placed directly next to a ‘No Skill’ warning.

3.6.2 Subjective Interpretation

For each communication strategy participants were asked to report how likely they thought that warmer than average (upper tercile) temperatures would be based on information about likelihood and skill provided in the communication strategies, how confident they were in this judgement, and how useful they thought that forecasts of this nature would be for decision making within their organisations. As one might expect, Higher Skill predictions were judged to be more useful than Lower Skill predictions. However, we find that while participants had less confidence in their judgements of likelihood when presented with predictions with no skill than predictions with high skill, the information about the predicted likelihood presented in the communication strategies still affected their subjective judgements of how likely warmer than average temperatures were. Hence, where no skill exists for a particular region or time period providers of climate predictions may wish to consider a) providing climatology only; or b) explicitly statement that the absence of skill means that predicted likelihoods should be disregarded to those recipients who are not domain experts.

3.6.3 Preference

Amongst experienced users of statistical information, no clear subjective preference for any one type of communication strategy emerged in the two Decision Labs, suggesting that preference depends on one’s own occupational context. Amongst novice users of statistical information however, a preference for the Simple Table was in evidence in the Abridged Decision Lab. When it came to the relationship between preference and other measures we found, as was the case in Task 33.1 (Taylor and Dessai, 2014; Taylor et al., 2015a), that participants tended to prefer those formats that they rated as being most familiar to them. No direct relationship between preference and objective understanding was found. However, in the Abridged Decision Lab it was found a) that greater reported familiarity corresponded with lower objective understanding; and b) that when familiarity was controlled for, objective understanding was positively related to preference. This could suggest that while people tend to prefer familiar formats this familiarity can lead to misinterpretation when the similarities are superficial, or when different visual elements are used in different contexts.

3.6.4 Comments on Specific communication strategies

Bubble Map

We found that participants' understanding of the Bubble Map was substantially better when they were shown the Higher Skill version of the map. When shown the Lower Skill version (where just a few areas showed skill above 0) responses were much less accurate. It seems that many participants in both the Abridged and Full Decision Labs that 'white space' on the map indicated either that the prediction indicated that all terciles were equally likely or that there was no skill for this season. In either case, white space on the map means that all terciles should be considered equally likely. Hence, this should be made clearer on any future iterations of this map. Including an additional legend to show what different sizes of bubble indicate may also go towards addressing the difficulty that some participants had in interpreting forecast likelihood on this map. Indeed, in response to the open-ended questions about the communication strategies presented in the Full Decision Lab, several participants indicated that found it difficult to interpret what the size of the bubbles meant, and would like such a legend to be included.

Violin Plot

When it came to determining the predicted likelihood of warmer than average and colder than average terciles this communication strategy was the least well understood, although assessments of what the skill score meant tended to be more accurate. Hence, while this style of visualisation may be useful for showing forecast ranges, it seems to be less useful for conveying precise probabilities. Additionally, participants responses to open-ended questions suggest that while some valued the 'completeness' of the information presented, many found the multiple layers of information made it overly complex and difficult to interpret. This suggests that usability may be increased by giving users the option to control which layers of information are viewed at any given time (e.g. PDF, climatology, individual ensemble members).

Table and Simple Table

Overall, the Table and Simple Table were the communication strategies best understood by participants when it came to making assessments of predicted likelihood and prediction skill. Hence, in cases where users are primarily interested in tercile likelihood and skill these may prove to be especially effective. However, the amount of information that these formats can provide beyond this is limited (e.g. less potentially useful when users are concerned with spread or require spatial information beyond a list of specific regions).

Bar Graph

This style of graph appears to have worked relatively well in conveying information about likelihoods. However, the way that skill was represented appears to have been difficult for many participants to correctly interpret; with some mistaking skill score for likelihood and others struggling to integrate three different ROCSS. There is also some evidence that the line representing climatology on the graph was in some instances mistaken for an indicator of likelihood. Hence, it is to be recommended that a) information about skill be saliently labelled or (where users are not concerned with the precise numeric value of the skill score) replaced by verbal categories; and b) climatology not be represented by a line across the graph.

Confidence Index

In the Abridged Decision Lab this communication strategy fell between the Simple Table and Bar Graph with respect to objective understanding and preference. However, responses to the questions about objective suggested that participants made inferences about the 'likelihood of colder than average temperatures' based on the stated 'likelihood of warmer than average temperatures', rather than conclude that this information was not provided. This suggests that, while this format may suit situations where a binary split exists, using it to represent the likelihood of particular terciles (or other quantiles) may lead to misinterpretation.

3.7 Key Conclusions

- Our findings suggest that numeric representations of likelihood may be easiest to interpret when the information of interest is the likelihood of a particular quantile.
- Even when users recognise that a prediction has no skill it may still influence their judgement of how likely a certain event is.
- Tables work well for conveying information about ‘likelihood of tercile’ and overall level of skill. However, they may not be the best choice for other types of information.
- Where ‘white space’ on maps indicates no skill, or that no prevailing tercile exists, this should be explicitly pointed out as it may otherwise lead users to judge the probability of all terciles to be very low.
- Evaluative categories can be useful for communicating skill to users with Lower Statistical Experience.
- Preference does not have a direct relationship with objective understanding.
- People tend to prefer communication strategies that they perceive as more familiar, but greater perceived familiarity can lead to misinterpretation where elements differ.
- It is vital that climate service providers give clear, explicit guidance as to how the communications that they provide should (and should not) be interpreted, taking into account the misconceptions that may arise.

3.7 References

Bray, J. H., & Maxwell, S. E. (1985). *Multivariate analysis of variance*. Sage.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*. 16, 297-334.

Ellsberg, D. (1990). 4. Risk, ambiguity, and the Savage axioms. *Rationality in Action: Contemporary Approaches*, 89.

Jupp, T. E., Lowe, R., Coelho, C. A., & Stephenson, D. B. (2012). On the visualization, verification and recalibration of ternary probabilistic forecasts. *Phil. Trans. R. Soc. A, 370*(1962), 1100-1120.

Kaye, N., Hartley, A., & Hemming, D. (2012). Mapping the climate: guidance on appropriate techniques to map climate variables and their uncertainty. *Geoscientific Model Development, 5*(1), 245-256.

Lowe, R., Bailey, T. C., Stephenson, D. B., Jupp, T. E., Graham, R. J., Barcellos, C., & Carvalho, M. S. (2013). The development of an early warning system for climate-sensitive disease risk with a focus on dengue epidemics in Southeast Brazil. *Statistics in medicine, 32*(5), 864-883.

Lorenz, S., Dessai, S., Forster, P., & Paavola, J. (2015). Tailoring the visual communication of climate projections for local adaptation practitioners in Germany and the UK. *Phil. Trans. R. Soc. A,* 373(2055) 20140457,

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable: An empirical study. Journal of Educational Measurement, 263-269.

McCown, R. (2012). A cognitive systems framework to inform delivery of analytic support for farmers' intuitive management under seasonal climatic variability. *Agricultural Systems, 105*(1), 7-20.

McCown, R., Carberry, P., Dalgliesh, N., Foale, M., & Hochman, Z. (2012). Farmers use intuition to reinvent analytic decision support for managing seasonal climatic variability. *Agricultural Systems, 106*(1), 33-45.

Nunnaly, J. (1978). *Psychometric theory*. New York: McGraw-Hill.

Peters, E. (2008). Numeracy and the perception and communication of risk. *Annals of the New York Academy of Sciences, 1128*(1), 1-7.

Peters, E., Dieckmann, N. F., Västfjäll, D., Mertz, C., Slovic, P., & Hibbard, J. H. (2009). Bringing meaning to numbers: the impact of evaluative categories on decisions. *Journal of experimental psychology: applied, 15*(3), 213.

Scheffe, H. (1999). *The analysis of variance*. John Wiley & Sons.

Slingsby, A., Lowe, R., Dykes, J., Stephenson, D., Wood, J., & Jupp, T. (2009). *A pilot study for the collaborative development of new ways of visualising seasonal climate forecasts.* Paper presented at the Proc. 17th Annu. Conf. of GIS Research UK, Durham, UK, April 2009.

Taylor, A. L. & Dessai, S. (2014) Deliverable 33.1 Report on survey of end-user needs for improved uncertainty and confidence level information, EUPORIAS http://euporias.eu/sites/default/files/deliverables/D33.1_Final.pdf

Taylor, A. L., Dessai, S, Buontempo, C., & Dubois, G. (2014) Deliverable 33.2 Report summarising review of existing approaches for communicating confidence and uncertainty, EUPORIAS http://euporias.eu/sites/default/files/deliverables/D33.2_Final.pdf

Taylor, A. L., Dessai, S., & Bruine de Bruin, W. (2015a). Communicating uncertainty in seasonal and interannual climate forecasts in Europe. *Phil. Trans. R. Soc. A,* 373(2055), 20140454.

Taylor, A. L. et al. (2015b) Report describing formulation of strategies for communicating confidence levels for S2D forecasts. http://euporias.eu/system/files/D33.3.pdf

Weedon, G. P., Balsamo, G., Bellouin, N., Gomes, S., Best, M. J., & Viterbo, P. (2014). The WFDEI meteorological forcing data set: WATCH Forcing Data methodology applied to ERA-Interim reanalysis data. *Water Resources Research*, 50(9), 7505-7514.

3.8 Deliverable delay

An extension of two months was given for this deliverable in order to maximise participation in the Decision Lab. This has not adversely affected any other tasks or work package deliverables, and has allowed us to obtain a greater number of highly engaged participants than would have otherwise been the case.

3.9 Planned publications

At least one peer reviewed publication reporting the findings of the work reported here is planned.

4. Links Built

- Several of the visualisations presented in this task were created with the help of Work Package 32, who made a substantial contribution to preceding work on development of strategies for communicating.
- Following on from prior collaboration with Work Package 12, in designing the Decision Labs we received feedback from colleagues working on Work Package 41.
- We have discussed emerging findings from the Decision Labs with partners working on the Clinton Devon Estates prototype.

5. Acknowledgements

In addition to the listed Work Package 33 Contributors, we would like to thank Marta Bruno Soares (University of Leeds), Susanne Lorenz (University of Leeds), Pete Falloon (Met Office), Maurice Skelton (ETH) , and Christoph Spirig (MeteoSwiss) for their feedback on the Decision Lab and help with its translation and dissemination. We also thank Maria Dolores Frias (University of Canabria), Jesus Fernandez (University of Cantabria) and Aiden Slingsby (City University) for their help in developing the visualisations, the Climate UK Network and Climate Service Partnership for their help in participant recruitment, and all of our participants for taking the time to contribute to this research.