Probablement, Likely, Wahrscheinlich? A Cross-Language Study of the Verbalization of Probabilities through Data Visualizations

by

Noëlle Rakotondravony

A Thesis

Submitted to the Faculty

of the

WORCESTER POLYTECHNIC INSTITUTE

In partial fulfillment of the requirements for the

Degree of Master of Science

in

Computer Science

April, 2022

APPROVED:

Professor Lane Harrison, Major Thesis Advisor

Professor Gillian Margaret Smith, Major Thesis Reader

Professor Craig E Wills, Head of Department

Abstract

Visualizations today are used across a wide range of languages and cultures. Yet the extent to which language impacts how we reason about data and visualizations remains unclear. In this paper, we explore the intersection of visualization and language through a cross-language study on estimative probability tasks with icon-array visualizations. Across Arabic, English, French, German, and Mandarin, $n = 50$ participants per language both chose probability expressions — *e.g. likely*, *probable* — to describe icon-array visualizations (Vis-to-Expression), and drew icon-array visualizations to match a given expression (Expression-to-Vis). Results suggest that there is no clear one-to-one mapping of probability expressions and associated visual ranges between languages. Several translated expressions fell significantly above or below the range of the corresponding English expressions. Compared to other languages, French and German respondents appear to exhibit high levels of consistency between the visualizations they drew and the words they chose. Participants across languages used similar words when describing scenarios above 80% chance, with more variance in expressions targeting mid-range and lower values. We discuss how these results suggest potential differences in the expressiveness of language as it relates to visualization interpretation and design goals, as well as practical implications for translation efforts and future studies at the intersection of languages, culture, and visualization. Experiment data, source code, and analysis scripts are available at the following repository: `https://osf.io/g5d4r/?view_only=859b329ad27847a69c8641e019ab76cf`

Acknowledgements

I would like to express my gratitude to my advisor, Prof. Lane Harrison, who believed that my idea for the “Language project” would make a good research to pursue, and for his patience and support throughout the process. Many thanks as well to my co-author, Yiren Ding. I also thank my reader, Prof. Gillian Smith, for her feedback and for trusting that this is a discussion worth spending her valuable time on. And finally, my endless gratitude to my friends and family, close and remote, who believe I always know what I’m doing. You’re funny and you keep me going!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

English remains the dominant language in the study and practice of visualization, but the landscape is changing. Creators from diverse languages and cultures are producing more visualizations, in part due to increasing access to visualization authoring tools and publishing ecosystems. In fact, international newsrooms are producing visualization-laden data journalism, and citizens on social media are sharing and discussing visualizations in their native languages. While the use of data visualizations in cross-national communication of global events enables broader access to essential information, new arising challenges highlight the importance of gaining a deeper understanding of the interplay of language, culture, and visualization. Climate change, pandemics, and misinformation—all will require a global collective engagement with data to navigate.

Efforts in Human-Computer Interaction (HCI) show how effects of language and culture might emerge in visualization. Several HCI studies and design guidelines that focus on WEIRD populations (Western, Educated, Industrialized, Rich, and Democratic) have failed to generalize when considering other languages and cultures [46]. Other studies have shown that HCI design processes can successfully integrate language and culture as an influence for interaction mechanisms and interface design [15, 51]. More broadly, some

HCI research agendas have included multicultural populations and their needs from the outset [41], by proposing experimental studies in languages outside English, and considering cultural aspects proper to the target population. These findings raise questions for the visualization community: *to what extent do studies centering on WEIRD populations generalize to the broader global population?*

There are comparatively few studies examining language and culture in visualization. A few studies have explored the associations of languages and data visualization, and its impact on data visualization practices on viewer's behaviour towards data visualization, both during design and use. A notable exception is Kim *et al.*'s study of color names across languages [26], which found that some languages have more distinct names within certain color ranges, potentially influencing visualization color palette and the verbal communication about visualized data to audiences speaking different language. Their study also highlighted that using traditional translation engines only does not always capture the nuances and subtle differences in the perception of color names across languages [26]. Related to visualization, studies have found that different languages and culture impact the use of color [21], and forms of how people represent time visually [17].

Besides the categorical aspects of the components of design and interaction, another essential characteristics of data visualization is its use in quantitative reasoning with data. Studies focusing on language and statistics offer a promising means for further exploring visualization across languages. In a widely replicated study, Kent surveyed intelligence analysts across his organization about their understanding of the *Words of Estimative Probability*, probability expressions such as *"likely"* or *"almost certain"* used in intelligence reports. The study showed that participants gave different numerical estimations the expressions, highlighting the variations betewen the intended and perceived meaning of the expressions [25] (see Figure 1.1). Later studies also examined probabil-

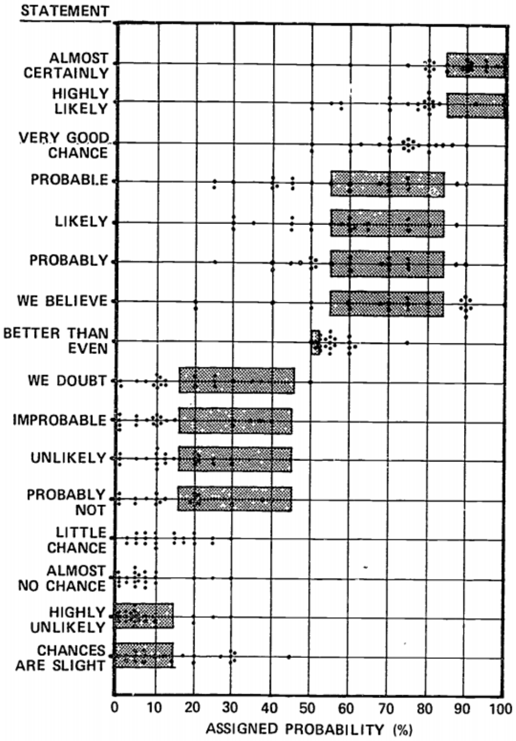


Figure 1.1: Results from Kent's survey of 23 intelligence officers. Each dot represent a probability assigned to an expression. The shaded areas indicate the scale range that Kent proposed for the verbal expressions [2]. For comparison, we superimpose the shaded areas on our results, see Figure 4.3.

ity expressions across languages. Renooij and Witteman elicited numerical values for several probability expressions with Dutch speakers [42], Willems *et al.* also surveyed the numerical interpretation of probability phrases in Dutch news articles [43] calling for careful use of the expressions, especially when communicating about scientific data. Doupnik *et al.* studied how German and English-speaking accountants interpret verbal probability expressions in International Accounting Standards, finding significant differences depending on the language [12]. Recently, visualization researchers Henkin and Turkay extend similar methodologies to study expressions related to correlation estimation, concluding with an explicit call to examine possible effects across languages [23]. Given the sustained focus in uncertainty communication for visualization, findings at the intersection of probabilistic reasoning and language may lay the groundwork for examining visualization in similar ways. Additionally, when using data visualizations across cultures, while the intuition can be that translation is the tool to go, the extent to which the reasoning through data visualizations varies across languages (especially when comparing from an English-centered lens) is unclear.

In this paper, we explore the intersection of languages, probability expressions, and visualization. Beginning with the expression to probability methodologies of Kent [25], Renooij and Witteman [42], and others, we adapt these for visualization by having participants specify values given an expression by drawing icon-array visualizations. We then invert this procedure by having participants choose expressions for a given icon-array visualization, for a two-part randomized within-subjects study. We collect expressions from prior studies, resolving issues like phrase asymmetry, ending with a list of $n = 18$ base expressions in English. To extend to other languages, specifically French, German, Arabic, and Mandarin, we recruit native speakers in each language for a collaborative translation activity, using inter-coder agreement measures to finalize a set of translations. Using the crowdsourcing platform Prolific, $n = 250$ participants ($n = 50$ native speakers for each language) completed both *VisVis-to-Expression* and *Expression-to-Vis* sections.

Results suggest that people vary in how they visualize a given probability expression, with differences both within and across languages. Across languages, participants appear to agree more (*i.e.* the response ranges are tighter) when given expressions that indicate higher and lower probability values, such as *very good chance* and *highly unlikely*, see Figure Figure 4.3-a. Exceptions exist between languages, however, with some expressions producing substantially different value ranges from corresponding expressions in other languages, see Figure Figure 4.3-b. People also vary in how they choose expressions when given an icon-array visualization. In Arabic for example, participants chose

15/18 possible expressions when given an icon-array depicting a 40% chance, compared to 7/18 expressions for Mandarin-speaking participants, see Figure 4.7 . Additional exploratory analyses between experiments reveal differences in elicitation method, where people across languages tended to draw values for a given expression that were more extreme, while less extreme values were common when expressions were chosen for a given icon-array, see Figure Figure 4.8.

Taken together, the experiments and results reveal substantial differences in the expressiveness of language as it relates to how people interpret visualizations. We discuss these findings, and how such differences may impact aspects of the visualization design process, particularly as it relates to communication or visualization translation efforts.

We make the following contributions:

- Evidence of no clear mapping between drawn visualizations and probability expressions across languages, suggesting cross-language differences, see Figure Figure 4.3.
- Results suggesting that different languages exhibit varying degrees of expressiveness for associated icon-array visualizations, see Figure Figure 4.7.
- Cross-language experiment materials in 5 languages and datasets reflecting judgments of $n = 250$ participants, including 4,500 *Expression-to-Vis* judgments, and 4,750 *Vis-to-Expression* judgments.

Chapter 2

Background

In crafting an experiment targeting estimative probability expressions spanning multiple languages, we draw on methodologies from studies on statistics and language, as well as considerations from cross-cultural studies in the HCI community. For design choices related to the icon-array visualizations, we refer to several visualization studies using icon-arrays in various contexts.

2.1 Probability Expression Interpretation

Numerical formats provide a measure that can facilitate probability comparison [50]. Yet because the concept of probability is not understood in the same way by everyone, studies have shown that numbers can provide illusory precision [8]. Other studies have shown that people may prefer handling uncertainty with verbal expressions in conversation [52]. Verbal expressions of probability, however (*e.g.* "*highly likely, probable*"), can be interpreted differently. In a survey of 23 intelligence officers, Kent found variation in the numerical values and ranges that participants assigned to probability expressions that were commonly used in intelligence reports [25]. Kent's work is an early study of the interpretation of probability expressions, and highlighted the uneven relationship

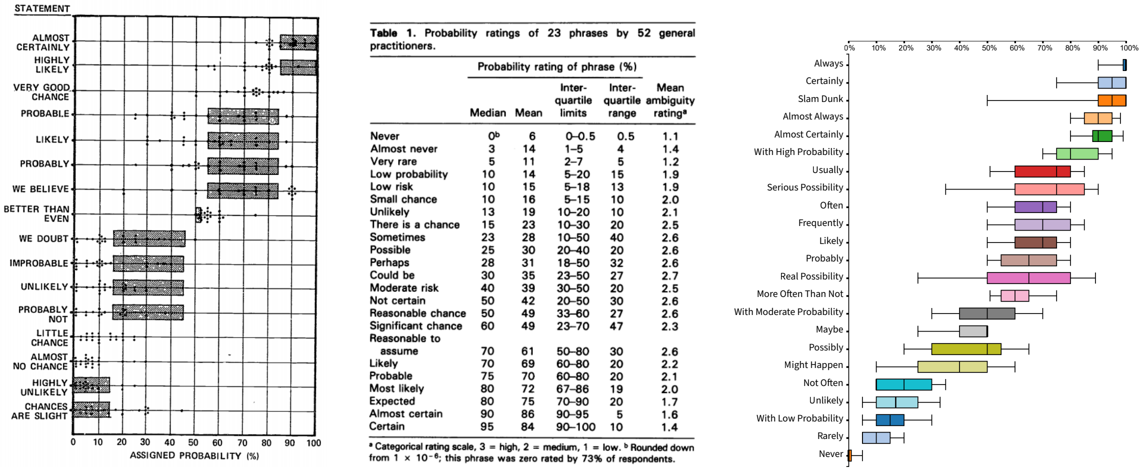


Table 1. Probability ratings of 23 phrases by 52 general practitioners.

	Probability rating of phrase (%)				
	Median	Mean	Inter-quartile limits	Inter-quartile range	Mean ambiguity rating[a]
Never	0[b]	6	0–0.5	0.5	1.1
Almost never	3	14	1–5	4	1.4
Very rare	5	11	2–7	5	1.2
Low probability	10	14	5–20	15	1.9
Low risk	10	15	5–18	13	1.9
Small chance	10	16	5–15	10	2.0
Unlikely	13	19	10–20	10	2.1
There is a chance	15	23	10–30	20	2.5
Sometimes	23	28	10–50	40	2.6
Possible	25	30	20–40	20	2.6
Perhaps	28	31	18–50	32	2.6
Could be	30	35	23–50	27	2.7
Moderate risk	40	39	30–50	20	2.5
Not certain	50	42	20–50	30	2.6
Reasonable chance	50	49	33–60	27	2.6
Significant chance	60	49	23–70	47	2.3
Reasonable to assume	70	61	50–80	30	2.6
Likely	70	69	60–80	20	2.2
Probable	75	70	60–80	20	2.1
Most likely	80	72	67–86	19	2.0
Expected	80	75	70–90	20	1.7
Almost certain	90	86	90–95	5	1.6
Certain	95	84	90–100	10	1.4

[a] Categorical rating scale, 3 = high, 2 = medium, 1 = low. [b] Rounded down from 1 × 10^{-6}; this phrase was zero rated by 73% of respondents.

(a) (b) (c)

Figure 2.1: Example studies of the quantitative perception of probability words in (a) decision-making [2], (b) general practice [34], and (c) from an online public survey [31].

between the meaning that a communicator intends and the meaning that the audience may perceive (see Figure 1.1). We aim to see if this effect can be found for visualizations of similar values, and whether different patterns are found in languages beyond English.

Studies targeting probability expressions have also been contextualized in concrete scenarios such as risk assessment, medical communications, and domains including both scientific and popular investigations. Empirical studies on the numerical estimation of probability expressions have used elicitation methods comprising word-to-number translations [8], number-to-word conversion [40], and rank ordering of expressions [42, 33]. In these studies, probability expressions are generally studied by giving people probabilistic outcomes for specific scenarios. Results have isolated several potential factors that may impact how people understand, assess, and communicate probabilistic data. For example, the combination of verbal and numerical formats like percentage, frequency or numerical range, have been shown to aid peoples' probabilistic reasoning [52, 7].

Other studies have explored factors related to culture and language. Doupnik and

Richter find that German accountants' interpreted probability expressions in international accounting standards as reflecting significantly lower values than that of their American counterparts [13]. Follow up studies have speculated that this may reflect differences in cultural values where German accountants express more conservatism and stronger risk-avoidance [4, 11].

2.2 Uncertainty Visualization and Icon-Arrays

Visual depictions of probability are widely considered to be effective means for communicating uncertainty, aiding audiences of different backgrounds in various scenarios to improve decisions, trust and judgment [38, 45]. Today, uncertainty visualization is widely studied and applied in both scientific domains [6] and in communication with general audiences [24]. One of the most common approaches in uncertainty visualizations implements *frequency framing*, in which the probabilistic information are displayed in frequency or ratio format [44]. In a frequency-based representation, the chance of occurrence of an event is shown as a part-to-whole proportion, considered to align better with how people naturally think of probability [22, 48]. Studies have found that visualizations depicting uncertainty in frequency format tend to be effective in communicating risks, especially for people with low numeracy [54, 38, 18, 19].

The icon-array is a common visualization type that implements frequency framing. Icon-arrays typically include one shape (or icon) repeated a number of times, with some of the shapes colored or otherwise marked to represent a proportion (*e.g.* 35/100). Figure 2.2 shows examples of icon array arrangements used in the literature, in this study, and in public-facing dashboards. Several studies have shown that icon-arrays are an effective method for communicating risk, such as simple ratio-based probability values [37]. This part-to-whole representation of proportion reflects the frequency of events

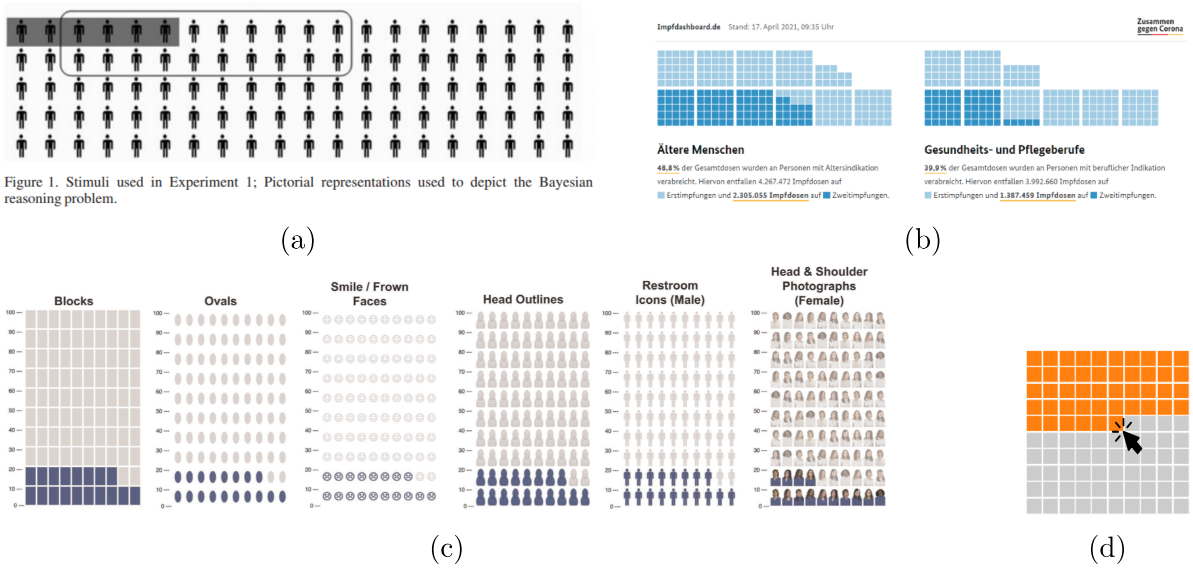


Figure 2.2: Example of icon array used in used in (a) a Bayesian reasoning study [5], (b) public-facing visualization of Covid19 vaccination in Germany [9], (c) a study of the impact of the icons on risk perceptions and recall of information [4], (d) the type of icon array used in the experiments reported in this study.

and chances, providing a visual affordance for audiences with different statistical and visualization experience to grasp. Studies have identified potential additional benefits of icon arrays including increased accuracy risk estimation tasks [35], reduced denominator neglect [19], and better understanding of medical risk severity [18].

In the present study, we identify a direct connection between probability expressions from studies that focus on statistics and language, and icon-array visualizations. Building on icon-array designs used in prior work, we use $10x10$ icon arrays and similar encoding methods, following studies by Kreuzmair *et al.* [28], Bancilhon *et al.* [1], Garcia-Retamero *et al.* [19], and Galesic *et al.* [18].

2.3 Visualizations, Text, and Language

While there is little prior work in visualization across languages and culture, topics of visualizations and text or visualizations and natural language provide perspectives that inform the current work. Verbal, text-based answers visualization tasks are a common methodology, from Cleveland and McGill's graphical perception experiment where people specify a text-based answer [10], to open-ended conversational methodologies such as Peck *et al.*'s study of visualization perceptions in a local farmer's market [39]. Additionally, visualization is also considered in context with the text that surrounds it. Different combinations of text and visualization can affect the statistical reasoning fluency of users [36, 18]. The composition and the framing of text and titles can induce bias and influence the attitude of readers who might perceive opposite messages from the same visualization [27].

The alignment between the perception of data through visualization and the language used to talk about the data has been sought by researchers who investigate the expressiveness of data visualizations. In their study of the "Words of Estimative Correlation", Henkin and Turkay analyzed utterances and verbal descriptions from experiment participants to find how people reason and talk about different levels of correlation seen in scatter-plots [23]. Their study highlights variations between how people use correlation terms to describe a visualizations and how they actually choose to visualize the terms or phrases. Drawing on this and other prior work, we focus on language in the sense of probability expressions, translating them across multiple languages, and determining how people associate these expressions with icon-array visualizations (and vice versa).

2.4 Studies Across Languages and Culture in HCI

Studies in human-computer interaction have investigated the impact of languages and culture on interface design norms and user behavior. In an online experiment, Baughan *et al.* found differences in how Japanese and American participants approach website navigation and information search, leading to concrete guidelines in which information might be presented across cultures [3]. Similarly, Evers showed that peoples' understanding of a graphical interface can be influenced by their cultural experience and language, with implications for interface metaphor design and interpretation [15]. Examining the transferability of primarily Western models of design in African contexts, Winschiers and Bidwell conduct information design activities with indigenous populations in South Africa and Namibia. Their findings surface Afro-centric paradigms which can shape interface design, with themes including cultural values such as interconnectedness, spirituality, and language used more through oral and performed communication [51]. More closely related to visualization, Gibson *et al.* analyzed the World Color Survey of 110 languages and show that the number of color names are related to how often colors are used within a given culture. They also noted an effect of industrialization, where color becomes an essential part of the identification of objects, impacting how well people identify and name certain colors [21].

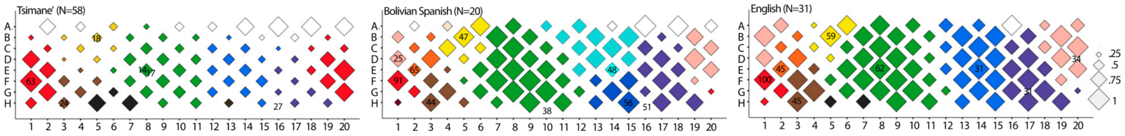


Fig. 1. The Amazonian Tsimane' people show large individual differences in color naming, but at the population level, similar color categories to those observed among Bolivian-Spanish and English speakers.

Figure 2.3: Example of cross-language and cross-culture studies in HCI showing the variability in color-term use (diamond sizes) among Tsimane', Bolivian Spanish, and English population in Gibson *et al.*'s study [21]

While it can be argued that computer interfaces are more widely distributed throughout languages and cultures than data visualizations, visualization appears to be on the rise as well. These findings in human-computer interaction suggest that the visualization community could be doing more to question its assumptions about the universality of approaches and guidelines, particularly as data becomes more global. We aim to take another step towards this goal by designing a study— similar in spirit to the internationalized HCI-focused studies of LabintheWild [41]— to examine probability expressions in relation to icon-array visualizations, across multiple languages.

Table 2.1: Beginning with English phrases from prior work, we conduct a translation activity to produce translations across 5 languages.

	English	**French**	**German**	**Mandarin**	**Arabic**
1	plausible	plausible	plausibel	貌似可信	معقول
2	almost certain	presque certain	ziemlich sicher	几乎确定	شبه مؤكد
3	highly likely	fort probable	sehr wahrscheinlich	极有可能	من المرجّح جدّا
4	very good chance	de très grandes chances	sehr gute Chance	很有可能	احتمال كبير
5	probable	probable	wahrscheinlich	可能	محتمل
6	likely	possible	möglich	或许	مرجح
7	probably	probablement	vermutlich	也许	من المحتمل
8	chances better than even	plus d'une chance sur deux	überdurchschnittliche Chancen	超过一半概率	أكثر من متساوي
9	chances about even	chances à peu près égales	ungefähr gleiche Chancen	大约一半	شبه متساوي
10	chances less than even	moins d'une chances sur deux	unterdurchschnittliche Chancen	不到一半	أقل من متساوي
11	probably not	probablement pas	wahrscheinlich nicht	可能不会	من غير المحتمل
12	improbable	improbable	unwahrscheinlich	不太可能	
13	unlikely	invraisemblable		未必	من غير المرجح
14	little chance	peu de chance	kleine Chance	没什么几率	فرصة ضئيلة
15	almost no chance	presque aucune chance	fast chancenlos	几乎没概率	تقريبا لا توجد فرصة
16	highly unlikely	très peu probable	sehr unwahrscheinlich	极不可能	من المستبعد جدا
17	chances are slight	les chances sont faibles	die Chancen sind gering	机会渺茫	فرص ضعيفة
18	implausible	peu plausible	nicht plausibel	难以置信	غير المعقول

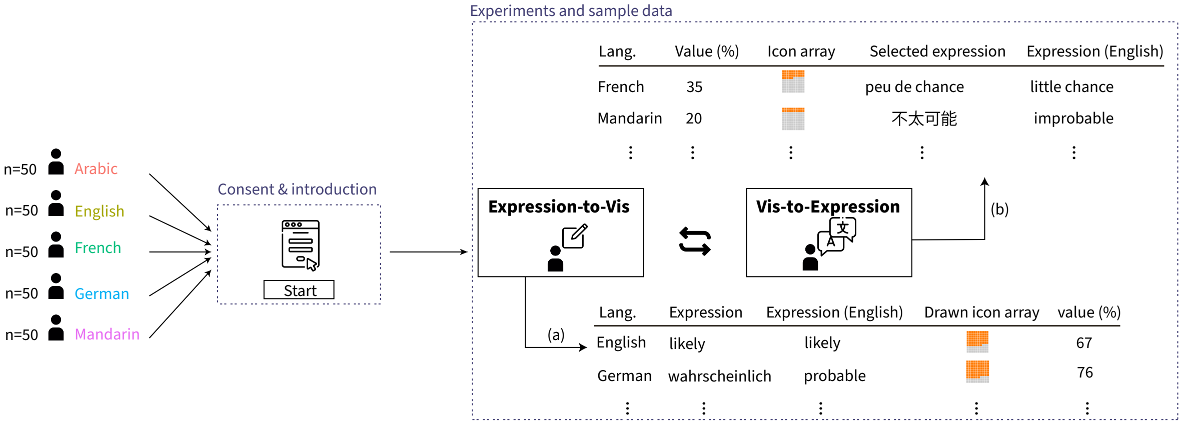
Chapter 3

Methodology

Designing a cross-language requires addressing several challenges, primarily centered around translation, but also typical concerns such as participant scenarios/prompts, visual encoding design, and interaction. We begin with a baseline methodology, extended from probability expression studies including Kent [25], Renooij and Witteman [42], and Henkin and Turkay's study in the visualization community [23]. These inform two experiments described here that investigate (1) how people visually represent a given expression through icon-arrays, across multiple languages and (2) how people across languages choose expressions to describe a particular icon-array. An overview of our experiment methodology can be found in Figure 3.1.

3.1 Participant Prompts/Context

In general, studies targeting probabilistic reasoning and uncertainty give participants a specific context that defines the nature of the task. Such framings are known to impact peoples' behavior [49]. Visualization studies have explored a range of scenarios, from the relatively neutral "when is my bus coming?" [24] to the charged "what is the chance that someone has cancer?" [36]. Because studies have found that language and culture



Lang.	Value (%)	Icon array	Selected expression	Expression (English)
French	35	▨	peu de chance	little chance
Mandarin	20	▨	不太可能	improbable
⋮	⋮	⋮	⋮	⋮

Lang.	Expression	Expression (English)	Drawn icon array	value (%)
English	likely	likely	▨	67
German	wahrscheinlich	probable	▨	76
⋮	⋮	⋮	⋮	⋮

Figure 3.1: The methodology that we follow in our studies. After signing the consent form and viewing an introduction, participants start the study with either Expression-to-Vis or Vis-to-Expression. **(a)** and **(b)** point respectively to sample data sets collected from the two experiments.

can impact how people perceive risk [12], we adapt a neutral context of a game which we introduce at the beginning of each experiment as follows:

> You are participating in a game which consists of drawing a tile from a set at random. Some of the tiles are orange, and some are gray. The game has two possible outcomes:
>
> - You draw an orange tile and you win a prize
> - You draw a gray tile and you do not win a prize

3.2 Selecting and Translating Probability Expressions

In studies targeting language and statistics, various ranges of probability expressions have been used. For example, Kent began with five and expanded to 16 [25], while Renooij and Witteman use the seven expression that were most suggested by the study participants [42]. Methods for collecting expressions include scanning prior literature, eliciting expressions from study participants [42], and borrowing from specific documents

[13, 7]. We began with Kent's original list of 16 expressions, as they have been adapted across several studies (*e.g.* [40, 4, 33, 13]). Through pilot studies, we identified two ambiguities that impacted the symmetry of the list. We add "chances less than even" a complement to "chances better than even" and "implausible" to match "plausible". As a result, for this study, we use a list of 18 expressions, and provide options for participants to specify their own if none fit what they would prefer to choose during the *Vis-to-Expression* experiment.

Our goal was to conduct the study in English, French, German, Mandarin, and Arabic. For each language, we recruited three independent translators who were native speakers, but also fluent in English. Before proceeding to translation, translators were reminded to consider the study scenario and asked to provide the closest translation of the probability expressions in their language.

To measure agreement, we adopt inter-rater reliability metrics (*e.g.* [32]). We calculate the Fleiss' Kappa values [16] for the translations. Results show that $\kappa_{French} = 0.55$, and $\kappa_{German} = 0.585$ are similar, while $\kappa_{Mandarin} = 0.187$ and $\kappa_{Arabic} = 0.341$ are lower in agreement. In terms of counts, the number of expressions for which all 3 translators disagree include $disag_{french} = 3$, $disag_{german} = 1$, $disag_{mandarin} = 8$, and $disag_{arabic} = 4$. The numbers of expressions for which all translators agreed are $ag_{french} = 8$, $ag_{german} = 8$, $ag_{mandarin} = 1$, and $ag_{arabic} = 3$.

Given high levels of disagreement in Mandarin (*i.e.* only 9/18 expressions had two people agreeing), we engaged native speakers for possible explanations. One potential reason that arose from this discussion is that a group of expressions in a source language (English) can map to a group of expressions in a target language (Mandarin) in an interchangeable way. For instance, {"probable", "likely"} and {"improbable", "unlikely" , "probably not" } were translated to Mandarin as {可能, 很可能, 大概, 也许} and {不太可能, 可能不会, 未必}. In such cases, personal preferences might play a role in word/phrase selection.

Next we resolve disagreements in the translation. In cases where two agree, we take the majority as the final expression. In cases where all translators disagree, we use a mediation procedure until agreements are reached about the expressions [32]. However, providing a single translation to each English expression was not always feasible. For example, “unlikely”and “improbable”were both repeatedly translated as “unwahrscheinlich” in German, and “probably not” and “improbable” were both translated as من غير المحتمل in Arabic. In these cases, we reduce the number of expressions in the target language, and mark them accordingly in results. Overall, we expect that some of these differences and similarities in languages will be reflected in the experiment results. Table 2.1 shows the final list of probability expressions used in this study with their translations in French, German, Arabic, and Mandarin.

3.3 Visual Encodings for the Icon-Arrays

Visualization studies have suggested that the type and arrangement of an icon-array can impact reader perception and engagement with the underlying data [54, 44]. To align with the scenario described in our study, we use a 10 x 10 grid of square icons, a typical ratio in the literature for problems with a population of 100 items (used in *e.g.* [36] and [53]) . Color is used to denote icons representing different outcomes in the event of interest: *Drawing an orange square and winning a prize*, and icons were arranged consecutively (see Figure 2.2d).

3.4 Participants and Procedure

Study participants were recruited using the online platform Prolific. Participants were required to have the target language of the experiment defined as their primary language in their Prolific profile. Further, Prolific users who were registered with two native languages can only take one version of the study. For each version of the study, Vis-to-Expression and Expression-to-Vis were assigned in random order, where half of the participants see Experiment 1 first while the other half see Experiment 2 first. 50 participants were recruited for each version of the study, making a total of 250 participants for all languages.

Chapter 4

Cross-language study of the verbalization of probabilities in icon arrays

Our study of probability languages and icon array visualizations consists of two experiments through which we collect data pertaining to (1) how people visually represent an expression through icon arrays and (2) how people choose expressions to describe an icon-array. Figure 4.1 highlights excerpts of results which we detail in this Section.

4.1 Experiment 1: Expression-to-Vis, from probability expressions to icon-arrays

In this experiment, we aim to understand how people visualize probability expressions through icon arrays, and how that varies across English, French, German, Mandarin, and Arabic. Unlike existing studies about numerical estimation of probability expressions, we ask participants to represent their estimations graphically.

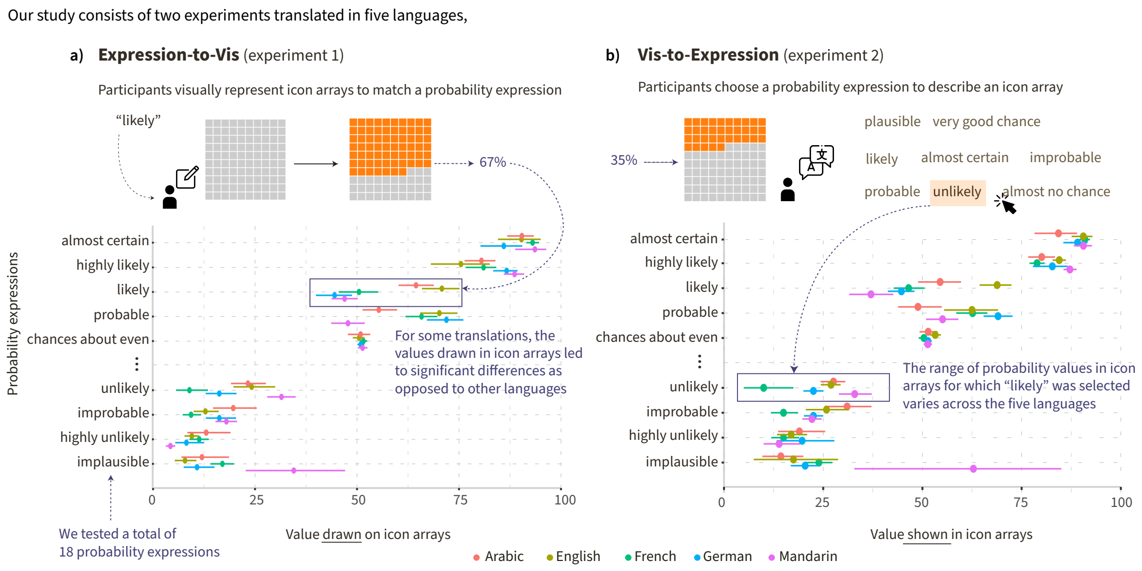


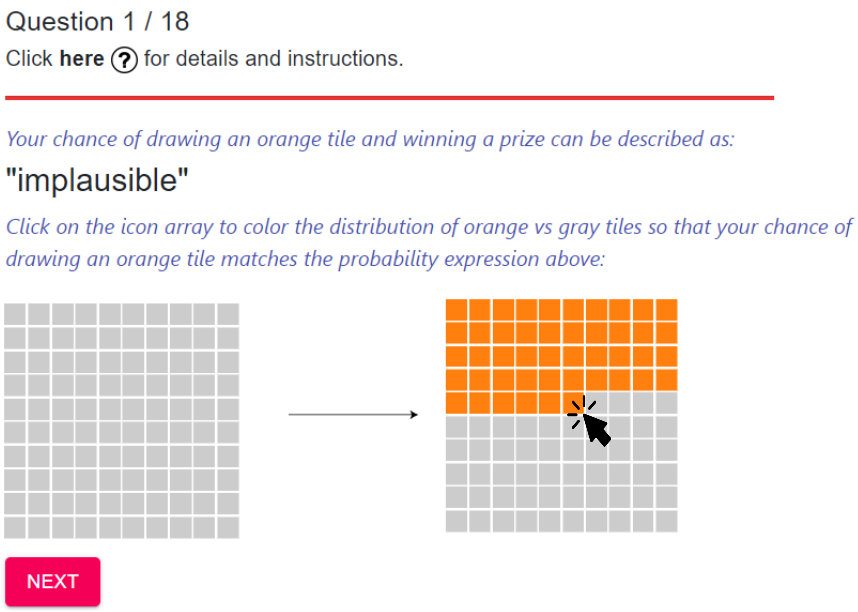
Figure 4.1: An overview of of the two experiments and their respective results in our cross-language study.

4.1.1 Procedure

Participants see an initial icon array with only gray icons, along with a probability expression that describes their chance of *picking an orange tile and winning a prize.* They are asked to click or click-and-drag on the icon array to show the proportion of orange icons that achieves the proposed verbal probability expressions. Figure 4.2 shows an example question in experiment 1. The Arabic, and German versions have 17 questions, whereas the English, French, and Mandarin versions have 18 questions. For each question, we collect the data format highlighted in the sample data in Figure 3.1(a).

4.1.2 Results

In total, we collected 4,400 answers from participants across English, French, German, Mandarin, and Arabic. Because there were two instances where translators agreed about a 2-to-1 mapping from an English expression to the target languages, we duplicate these confidence intervals and perform statistical comparisons separately for each. These



Question 1 / 18

Click **here** ? for details and instructions.

Your chance of drawing an orange tile and winning a prize can be described as:

"implausible"

Click on the icon array to color the distribution of orange vs gray tiles so that your chance of drawing an orange tile matches the probability expression above:

Figure 4.2: Instructions for experiment 1 *Expression-to-Vis* in English. The orange-colored icon array indicates a sample answer by the participant, which we evaluate numerically as 46%. Participants can access the instructions at anytime during the experiment.

include the entries for unwahrscheinlich for the English expressions “unlikely” and “improbable”, and من غير المحتمل for the English expressions “probably not” and “improbable”.

Figure 4.3 shows 95% confidence intervals of means for all expressions across the five languages. There appears to be general alignment with Kent’s suggested ranges, though some such as “highly likely” and “improbable” deviate somewhat. This may be due to sample differences, *i.e.* Kent studied 23 intelligence analysts in the 1960s. More generally, we notice several differences for a given expression across languages. For example, visualizations drawn for expressions aligning with “likely”, “probable” and “probably” in French, German, Mandarin and Arabic deviate lower than English, in some cases below the 50% mark.

While we generally align our analyses with recommendations in the VIS and HCI communities to move beyond dichotomous statistics [14], we provide statistical com-

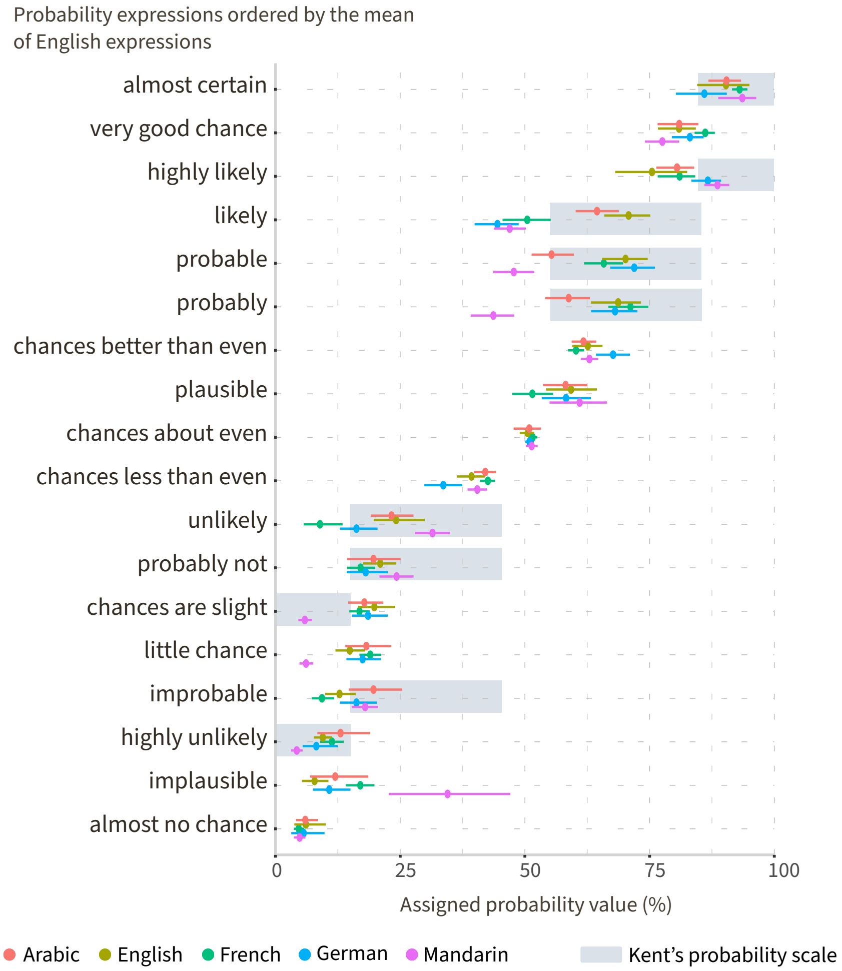


Figure 4.3: 95% confidence intervals of mean responses in *Expression-to-Vis*. The range of responses across the five languages are tighter for expressions indicating high and low values. Shaded areas indicate the scale range of probabilities proposed by Kent [25] for the corresponding expression.

parisons here to go along with analysis shown in Figure 4.4. Our aim is to identify expressions that are substantially above or below the associated English translations. While comparisons between other languages are possible, we focus on English since the expressions were originally translated from English.

Analyzing the between-language variance of participant-drawn visualizations with a one-way ANOVA, we find that only five expressions do not show at least one significant difference across the five languages. These stable expressions include "plausible" (and its

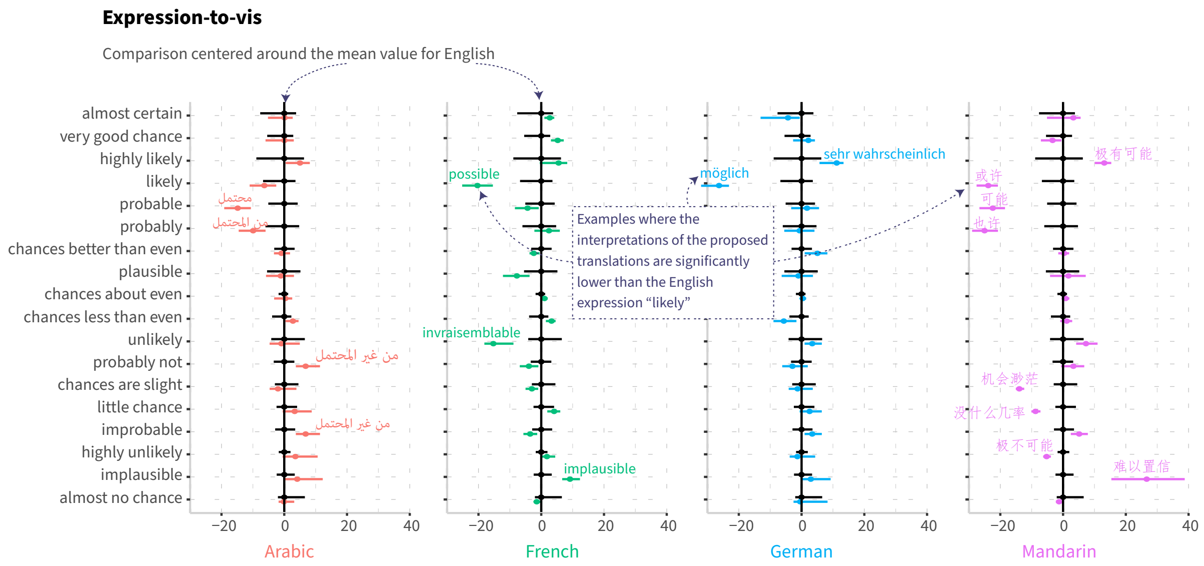


Figure 4.4: 95% confidence intervals for differences in mean of English expressions and their translations. Values are centered on the mean value in English. Intervals on the left indicate that the mean value for an expression is lower than its associate in English. We label translations that are significantly misaligned with the English expressions. In German, entries for “unlikely” and “improbable” are duplicated (they repeatedly translated as the same German unwahrscheinlich); similarly in Arabic, entries for “improbable” and “probably not” are duplicated (they are translated in the same Arabic expression من غير المحتمل).

translations: “plausible”, “plausibel”, “معقول”, “貌似可信”), “almost certain” (presque certain, ziemlich sicher,شبه مؤكد, 几乎确定), “chances about even” (chances à peu près égales, ungefähr gleiche Chancen, شبه متساوي, 大约一半), “probably not” (probablement pas, wahrscheinlich nicht, من غير المحتمل, 可能不会), and “almost no chance” (presque aucune chance, fast Chancenlos, تقريبا لا توجد فرصة, 几乎没概率).

To further analyze differences between languages, we use Tukey posthoc tests to identify pairs where the expressions significantly differ from English.

In French, we find two deviating expressions:

possible (mean: -20.32, 95% CI: [-29.23, -11.40], p.adj = 1.69E-8, English: likely), and

invraisemblable (mean: -15.26, 95% CI: [-23.80, -6.75], p.adj = 1.65E-5, English: unlikely).

Two expressions in Arabic also deviate from the English counterpart:

محتمل (mean: 14.84, 95% CI: [6.24, 23.44], p.adj = 3.52E-5, English: probable), and

من المحتمل (mean: 9.92, 95% CI: [0.83, 19], p.adj = 0.0246, English: probably).

For German, we find three deviating expressions:

sehr wahrscheinlich (mean: 11.2, 95% CI: [2.2, 20.19], p.adj = 0.0064, English: highly likely),

möglich (mean: -23.6, 95% CI: [-35.22,-17.38], p.adj = 3.03E-13, English: likely), and

unterdurchschnittliche Chancen (mean: -5.66, 95% CI: [-10.88, -0.44], p.adj = 0.026, English: chances less than even)

Finally, the Mandarin set of expressions has the highest number (seven) expressions that differ from English:

极有可能 (mean: 13.14, 95% CI: [4.15, 22.13], p.adj = 0.00075, English: highly likely),

可能 (mean: -22.4, 95% CI: [-31, -13.8], p.adj = 9.5E-11, English: probable),

或许 (mean: -23.84, 95% CI: [-32.76, -15], p.adj = 3.02E-11, English: likely),

也许 (mean: -25.02, 95% CI: [-34.10, -15.93], p.adj = 7.76E-12, English: probably),

没什么几率 (mean: -8.78, 95% CI: [-15.13, -2.42], p.adj = 0.0017, English: little chance),

机会渺茫 (mean: -13.94, 95% CI: [-19.9, -7.98], p.adj = 6.44E-9, English: chances are slight), and

难以置信 (mean: 26.66, 95% CI: [13.9, 39.42], p.adj = 2.76E-7, English: implausible).

These results show multiple instances where participants in a particular language consistently draw icon-arrays that align with different probability ranges than the associated English expression, both above and below. We will discuss possible reasons behind these differences, including implications for visualization, in chapter 5. Interestingly, translations for the duplicated entries both in German and Arabic did not significantly differ from the original English expression. This suggests that the proposed translation in German and Arabic do align with both expressions in English.

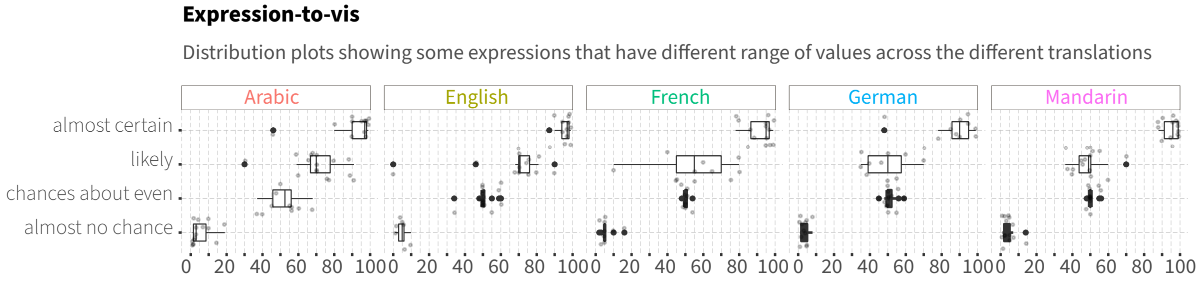


Figure 4.5: Range plots showing several example expressions and values of the drawn icon-arrays. There appears to be rounding behavior around multiples of 5 and 0. Value ranges for some expressions also seem to vary across translations (*e.g.* "likely" differs between French and other languages).

Other patterns in the analysis reflect possible drawing affordances in the icon-array that persist across languages. Most responses end in 0 or 5, similar to rounding behavior in graphical perception studies [47]. These include 51% of total answers for English, 54% for French, 52% for German, 45% for Arabic, 48% for Mandarin. This pattern also aligns with previous results where people tend provide numerical estimations that are multiples of 10 for verbal probabilities [29, 43] (see Figure 4.5)

We can also view each base expression from the perspective of its range across languages, see Figure 4.5. For example, although there is no particular pattern across the five languages for the expressions, the Mandarin translation of "implausible" shows the largest IQR with 91%. Expressions with narrow ranges suggest that, across associated expressions in other languages, participants will draw similar ranges in icon arrays. Another pattern is that expressions near extreme low/high and center values appear to have smaller interquartile ranges. Expressions at the extremes are consistently evaluated, while mid-range expressions (but not central) may convey less precise estimates of probability.

4.2 Experiment 2: Vis-to-Expression, From icon arrays to probability expressions

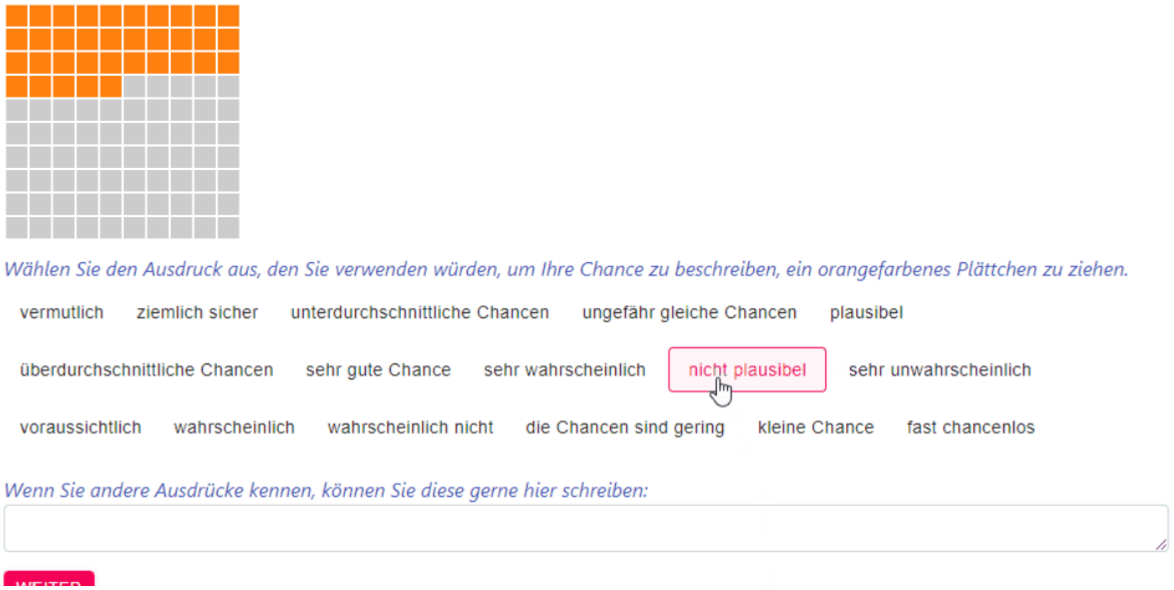
This experiment is essentially an inversion of Experiment 1. Here, we aim to understand how people choose probability expressions TO describe a given icon array visualization.

4.2.1 Procedure

Using the same neutral scenario, participants are given an icon-array of a specific value, along with a list of probability expressions (see Figure 4.6). Participants are asked to select the expression that they believe best describes the icon-array shown. Expression lists consist of 18 expressions in English, French, and Mandarin, 17 expressions in German, and Arabic. Participants are also encourage to provide their own answer, if desired. To cover the probability space, we encode 19 values between 5%to 95% with a step of 5%. For each trial, we collect the pair {icon-array value, selected probability expression}.

4.2.2 Results

Across 50 participants for each of the 5 languages, we obtain 4,750 icon-array to expression pairs. Each icon array across steps of 5% is described fifty times, a visualization of these results is provided in Figure 4.7. The upper bar graph in Figure 4.7 shows the count of unique expressions for each icon-array value. Across languages, there appears to be consistency in that higher values (90%, and 95%) are described using fewer unique expressions. This may reflect a more consistent expressiveness of languages of probability expressions in higher ranges. However, given the average counts across all value possibilities, it is clear that few, if any perfect matches of probability expressions and visualization exists.



Wählen Sie den Ausdruck aus, den Sie verwenden würden, um Ihre Chance zu beschreiben, ein orangefarbenes Plättchen zu ziehen.

Wenn Sie andere Ausdrücke kennen, können Sie diese gerne hier schreiben:

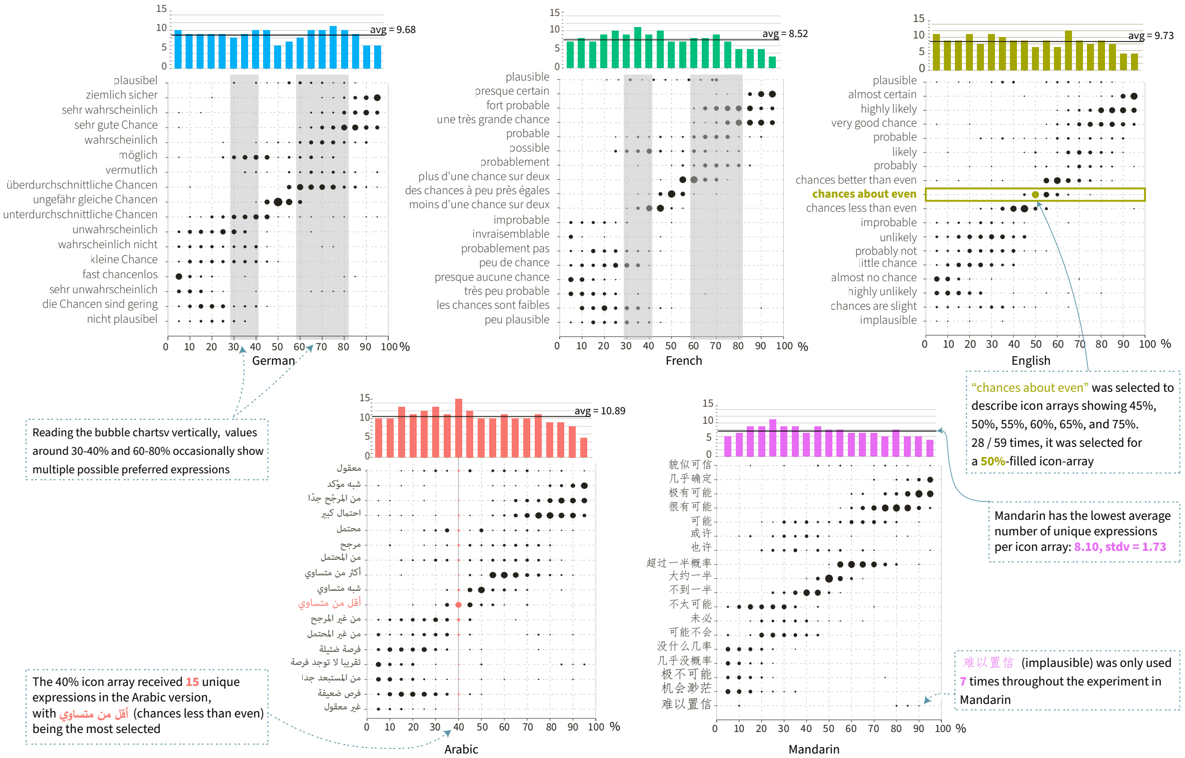
Figure 4.6: Instructions for experiment 2 *Vis-to-Expression* in German. Participants were also given the option to input their own expression to describe the icon array.

Arabic has the highest average number of unique expressions. On average, 10.89 expressions (stdv = 2) were used to describe each value of probability. For example, for icon-arrays of 40%, participants in Arabic chose 15 out of the 17 possible probability expressions. 10 of these expressions were selected at least twice. Looking back at the Expression-to-Vis results in Figure 4.3, there appears to be a gap above 25% and below 45%, although أقل من متساوي or "chances less than even" did end up being the most selected expression for this value (19 times).

In contrast, Mandarin has the lowest average count of unique expressions per icon array 8.10 (stdv = 1.73) than the other versions of the experiment. Implausible is a notable exception, which was rarely used and had a wide range associated with it. To a lesser degree, German, French, and English show the central values around 50% with a low number of unique expressions (below their averages), while mid-range values above and below 50% show more variance.

Figure 4.7 also shows the intersection of icon-array values and expression counts. The

Expression-to-vis



Reading the bubble chartsv vertically, values around 30-40% and 60-80% occasionally show multiple possible preferred expressions

"chances about even" was selected to describe icon arrays showing 45%, 50%, 55%, 60%, 65%, and 75%. 28 / 59 times, it was selected for a 50%-filled icon-array

Mandarin has the lowest average number of unique expressions per icon array: 8.10, stdv = 1.73

The 40% icon array received 15 unique expressions in the Arabic version, with أقل من متساوي (chances less than even) being the most selected

难以置信 (implausible) was only used 7 times throughout the experiment in Mandarin

Figure 4.7: Results from experiment 2 *Expression-to-Vis*. Barcharts represent the number of unique expressions that participants selected for a given icon-array. Bubbles indicate the count of each probability expression and value pair. Results show several similarities and differences across languages.

size of the circles indicate the ratio at which an expressions has been used to describe a given value.

Looking vertically, there appears to be variance in how icon-arrays depicting particular values are described, with some expressions preferred over other. For example, values around 30-40% and 60-80% occasionally show multiple possible preferred expressions. Across all values and expressions, larger circles appear to exist for central and high probability values, reflecting some of the consistency seen in the first experiment. People across languages seem to have clear preferences for translations of "chances about even" for an icon-array at 50%. Other patterns show variance and disagreement. The translations for "probable", "likely" appear to be used to specify a large range of probability values in a high range. Notably, however, similar patterns do not appear to exist across languages for wide ranges of low probability values. For instance, while "implausible" translations potentially fits this low-and-wide range role is German, French, and Arabic, it is scarcely chosen at all in English and Mandarin.

During the experiment, participants had the option to suggest additional probability expressions whenever needed. Participants provided their own expressions 55 times for Arabic, 99 for English, 145 for French, 65 for German, and 14 for Mandarin. Example trends in these include people using the listed expressions in a full sentences (*e.g.* Il est *improbable* de gagner ("it is improbable to win")), or with varying qualifiers (*e.g.*فرصة عالية, فرصة عالية جدا ("high chance, very high chance")). Among the new expressions and phrases that were suggested, we notice some referring directly to the proportion shown in the icon array arrangement (*e.g.* Etwa jeder Vierte gewinnt for a 25% icon array). We provide these data in the project repository[1], as an extended analysis of written answers by participants may give an opportunity to explore additional language or cultural factors that people refer to when making judgments about icon-array visualizations.

4.3 Exploratory Analysis: Comparing Experiments

Given the within-subjects design of both experiments, it is possible to make comparisons across them. In experiment 1 *Expression-to-Vis*, people were given each expression and drew a specific icon-array design (stored as a percentage value). Similarly, In experiment 2 *Vis-to-Expression*, icon-arrays of particular values were given and participants chose expressions.

Shown in Figure 4.8, one observation between the two experiments is that people tend to draw extreme values for given expressions, but when people are given icon-arrays

[1] https://osf.io/g5d4r/?view_only=859b329ad27847a69c8641e019ab76cf

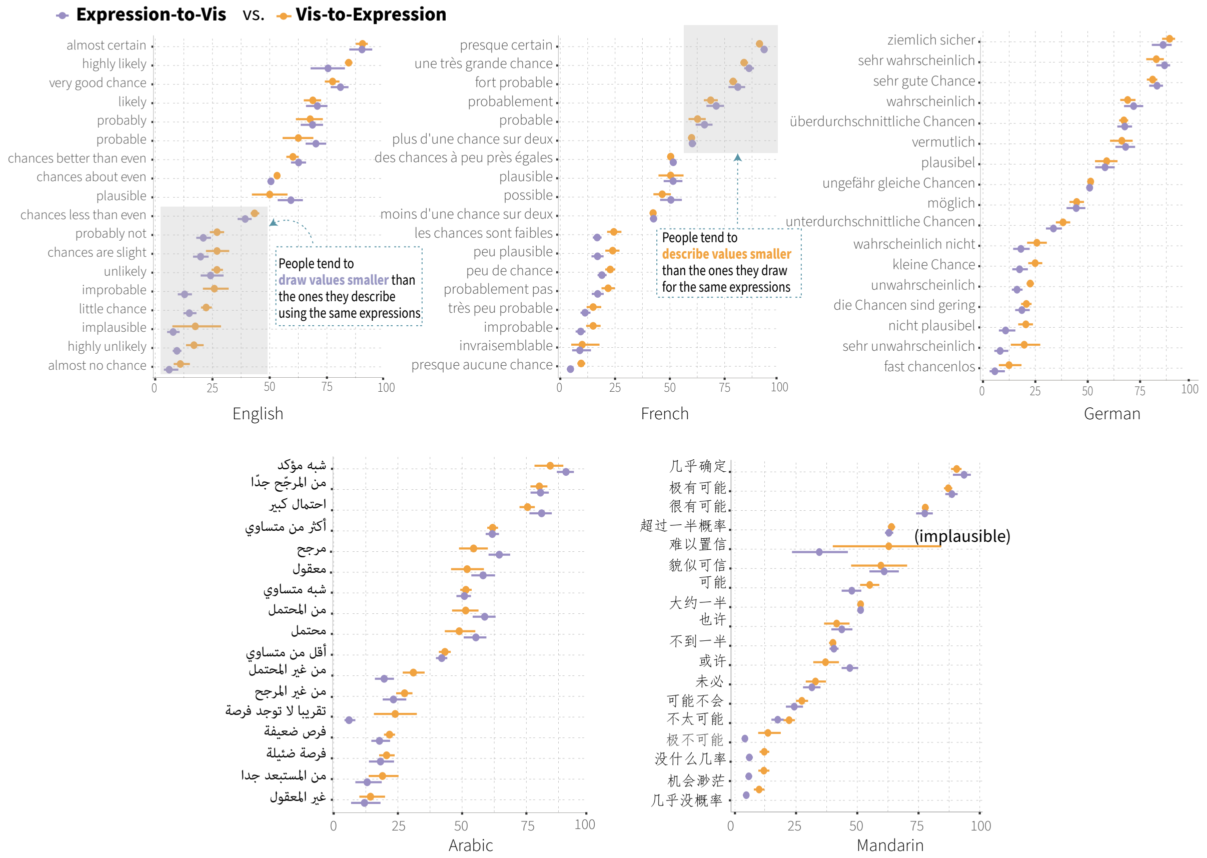


Figure 4.8: Comparing the two experiments, results show differences in elicitation methodology that hold across languages. People appear to draw visualizations with more extreme values when given an expression. But when given a visualization of similar value, people tend to choose different expressions.

depicting similar values, they choose different expressions. A concrete example is that when someone receives an expression *"almost no chance"*, they typically draw an icon array with very few colored squares. However, when showed with an icon-array with more colored squares than are typically drawn for this expression, people still describe this icon-array as *"almost no chance"*.

With these values for each participant, we can also explore how consistent people are in assigning expressions to visualizations versus visualizations to expressions. While several distance metrics are possible, in an exploratory analysis we define a straightforward distance measure for each probability expression as

$$Distance\ D = exp2vis\_val\ \text{-}\ mean(vis2exp\_val).$$

Where $exp2vis\_val$ is the value that a participant drew for an expression, and $mean(vis2exp\_val)$ is the average of all values for which participants assigned this expression. A large distance indicates that a participant is less consistent their mapping of language to visualization, while a low distance suggests more consistency. We find several instances where participants exhibit consistency across the two experiments, and other instances where participants differ between themselves widely. Example visualizations and data for these are included in the Appendix B.1. These also motivate potential individual-level modeling efforts for future work.

Chapter 5

Discussion

The results of the *Expression-to-Vis* and *Vis-To-Expression* experiments suggest that language plays an important role in the specification and interpretation of icon-array visualizations. Differences are found within languages, for example in Mandarin there was no overlap in Kent's suggested scale range of 55%-85% for the expressions *likely, probable*, and *probably* and the visualizations participants drew when given translations of these expressions.

There are other differences across languages. Using English as a source translation, we identified instances in every tested language that deviated significantly above or below the ranges for the associated English expression (Figure 4.3). These and other reported findings raise questions about the interplay between language expressiveness, visualization, and translation.

5.1 Implications for Visualization Translation

Translation is one aspect of the methodology that may play a role in the observed differences. Although multiple native speakers were consulted and mediation processes were used to reach agreement, it is possible that other speakers would have generated slightly different expressions. Even with other translators, however, there is no guarantee that other expressions would have matched the ranges specified in the source translation. This would also not guarantee that the resulting expression list would cover equal spans of the probability value ranges in the target language. Automatic translation such as Google Translate is another possibility, but will likely suffer from similar limitations.

These translation challenges raise one possible application of the methodology and the results here. It may be possible to *align* expressions based on the resulting participant-driven probability ranges. For example, while the expression plausible in French differs significantly from the English plausible, another expression in French, plus d'une chance sur deux (originally translating *chances better than even*), does align better with this range in the observed data. Computational methods could be designed to construct these translations for the tested languages and others. As an initial exploration, we iteratively paired up expressions across two languages until we found pairs that do not significantly differ following a Tukey test comparison. The results of this approach showing multiple possible "aligned" translations are shown in the Appendix A.2.

Importantly, computational approaches to language-expression alignment could address challenges in cross-language statistical reporting scenarios. The IPCC (Intergovernmental Panel on Climate Change) report, for example, specifies guidelines for its writers to use certain probability expressions for certain ranges [7]. As medical tests and associated symptom displays also rely on an intersection of icon-arrays and language (*e.g.* [36, 19, 20]), these efforts may also aid medical or pandemic risk communication. Our results provide a possible path towards further refining these standards, helping ensure the intended meaning of statistics and charts is communicated faithfully across languages.

5.2 Towards Better Elicitation for Cross-Language Visualization

Methodology was a key challenge in this study. While we aimed to carefully adapt and extend prior studies to begin exploring the intersection of visualization and languages, new possibilities emerged through the design and resulting analysis. In the *Vis-to-Expression* experiment, participants offered 86 (out of 4750) additional expressions. While some of these overlap with existing expressions in the study, it is likely that there are other expressions or phrases that each of the studied languages and associated cultures use in talking about probabilistic events. Finding ways to elicit these could be a challenging but rewarding effort for visualization.

For instance, a participant in the Mandarin version commented:

> 貌似可信”在汉语中不是一个好的表达，我作为母语者都不能完全理解你们用这个词想表达什么 (It seems “plausible” is not a good expression in Chinese. As a native speaker, I can’t fully understand what you are trying to express with this word)

In English, the word plausible is common, but might there be other translations or similar expressions in other languages that fill a similar role?

One possibility is to move beyond English as a source, and instead develop in-language elicitation methodologies. These might be graphically based, using interaction and visualization with input capability to allow participants to specify ranges and expressions. Alternately, they may be large crowdsourced studies, following similar scenario and trial-based methods. In either case, the goal would be to elicit a wider range of expressions and visualization descriptions from participants. Such efforts could reveal additional ranges and expressive capabilities within languages, beyond those studied here.

Beyond the results presented here, there are other useful starting points for exploring possibilities in cross-language elicitation for visualization. One source would be to consider the history of large-scale color elicitation studies such as the World Color Survey [30]. Other language and statistics studies such as Renooij and Witteman[42], and the NLP-driven analysis of Henkin and Turkay [23] might inform approaches to scale.

5.3 Limitations and Future Work

One limitation of the current work is the restriction to 5 languages. While these were chosen as an initial step in the space, and intended to cover several major language families, there are thousands of languages in the world and various language groups that could be explored. Especially in the context of a global pandemic, for example, it is important to support effective data-focused communication as broadly as possible. The neutral scenario / context given to participants and sole use of icon-array visualizations are other practical limitations to explore in future work. Cultural differences such as risk avoidance (*e.g.* [4, 12]) may be less pronounced in neutral scenarios, but become more pronounced with carefully designed contexts. We might refer to studies targeting medical reasoning (*e.g.* Ottley *et al.* [36]) or natural disaster risk (*e.g.* Padilla *et al.* [38]) for promising scenarios and visualizations to explore in future work.

Chapter 6

Conclusion

With the changing global landscape of data and visualization practice, it is important that the research community explores the intersection of language and visualization. We present two experiments with the goal of understanding how people across five languages draw icon-array visualizations given probability expressions and assign probability expressions given icon-array visualizations. Results of these experiments show several differences both across and within languages, with no clear mapping across languages, and several instances of possible “gaps” between expressions in a given language. We discuss implications of these results for ongoing efforts such as data and visualization translation, targeting areas such as climate change and pandemic communication. Taken together, these studies and results are intended to offer a limited yet useful step in broadening the focus of the visualization community beyond traditional WEIRD populations.

Appendix A

A.1 Experiment 1: Between-language one-way ANOVA

Experiment 1: One-way ANOVA analysis of the between-language variance of participant-drawn visualizations. Five expressions which do not show at least one significant difference across the five languages are highlighted in the table. Significance code *** [0, 0.001], ** (0.001, 0.01], * (0.01, 0.05], . (0.05, 0.1], (0.1, 1]

Expressions	**mean**					**F_value**	**p_value**	**sig**
	Arabic	**English**	**French**	**German**	**Mandarin**			
plausible	58.16	59.26	51.5	58.26	60.98	2.0232	0.091787	.
almost certain	90.42	90.32	93.04	85.98	93.6	2.1307	0.077602	.
highly likely	80.5	75.48	81.02	86.68	88.62	5.1533	0.000529	***
very good chance	80.94	80.9	86.16	83.08	77.56	3.2259	0.013255	*
probable	55.34	70.18	65.82	71.92	47.78	21.7627	2.21E-15	***
likely	64.46	70.8	50.48	44.5	46.96	25.3281	1.43E-17	***
probably	58.78	68.7	71.16	68.08	43.68	23.4020	2.12E-16	***
chances better than even	61.74	62.62	60.24	67.68	62.94	4.1575	0.002818	**
chances about even	50.88	50.54	51.56	50.92	51.36	0.2456	0.912138	
chances less than even	42.06	39.3	42.58	33.64	40.46	7.0990	2.01E-05	***
probably not	19.66	21	17.06	18.1	24.28	2.0015	0.094935	.
improbable	19.66	12.84	9.32	16.24	17.98	5.1166	0.000563	***
unlikely	23.26	24.18	8.92	16.24	31.48	15.1844	4.18E-11	***
little chance	18.22	14.9	19	17.46	6.12	10.4008	8.51E-08	***
almost no chance	5.98	6.1	4.66	5.66	4.84	0.2582	0.904477	
highly unlikely	13.04	9.5	11.32	8.18	4.26	4.1514	0.002847	**
chances are slight	17.82	19.82	16.84	18.56	5.88	13.5673	5.26E-10	***
implausible	11.98	7.84	17	10.78	34.5	10.4886	7.37E-08	***

A.2 Experiment 1: Post-hoc test and acceptable translations

For some English expressions, for which the post-hoc test showed that the initially proposed translation significantly differ in the value assigned in experiment 1, as an initial exploration, we iteratively paired up expressions across two languages until we found pairs that do not significantly differ following a Tukey test comparison. The results of this approach showing multiple possible "aligned" translations are presented below (see section 5.1). Empty rows indicate the cases where no translation could be found from the list of expressions in the target language.

English	**French**	**p_val**	**t0**	**95% CI**	
				meanDiff.L	**meanDiff.H**
likely	probablement	0.911505	-0.11143	-6.77	6.05
	probable	0.118598	1.574963	-1.3	11.26
unlikely	peu de chance	0.075487	1.806088	-0.55	10.91

Table A.1: New suggested translations in French

English	**Mandarin**	**p_val**	**t0**	**95% CI**	
				meanDiff.L	**meanDiff.H**
probably	貌似可信	0.056795	1.927948	-0.23	15.67
little chance	不太可能	0.141134	-1.48373	-7.2	1.04
chances are slight	可能不会	0.08825	-1.72187	-9.6	0.68
	不太可能	0.428595	0.795172	-2.76	6.44
implausible	几乎没概率	0.064321	1.881696	-0.18	6.18
	没什么几率	0.297549	1.049379	-1.55	4.99
	机会渺茫	0.232074	1.205551	-1.28	5.2
highly likely	很有可能	0.625069192	-0.490913461	-10.53	6.37
probable					
likely					

Table A.2: New suggested translations in Mandarin

English	German	p__val	t0	95% CI	
				meanDiff.L	meanDiff.H
highly likely	wahrscheinlich	0.4290	0.7948	-5.35	12.47
	vermutlich	0.1049	1.6398	-1.58	16.38
	überdurchschnittliche Chancen	0.072186	1.8256	-0.72	16.32
likely	wahrscheinlich	0.7367	-0.3370	-7.71	5.47
	vermutlich	0.4213	0.8075	-3.96	9.4
	überdurchschnittliche Chancen	0.3072	1.0267	-2.92	9.16
chances less than even	möglich	0.0677	-1.8501	-10.79	0.39

Table A.3: New suggested translations in German

English	Arabic	p__val	t0	95% CI	
				meanDiff.L	meanDiff.H
probably	مرجح	0.229102916	-1.210412272	-11.19	2.71
probable		0.083849961	-1.746564262	-12.22	0.78

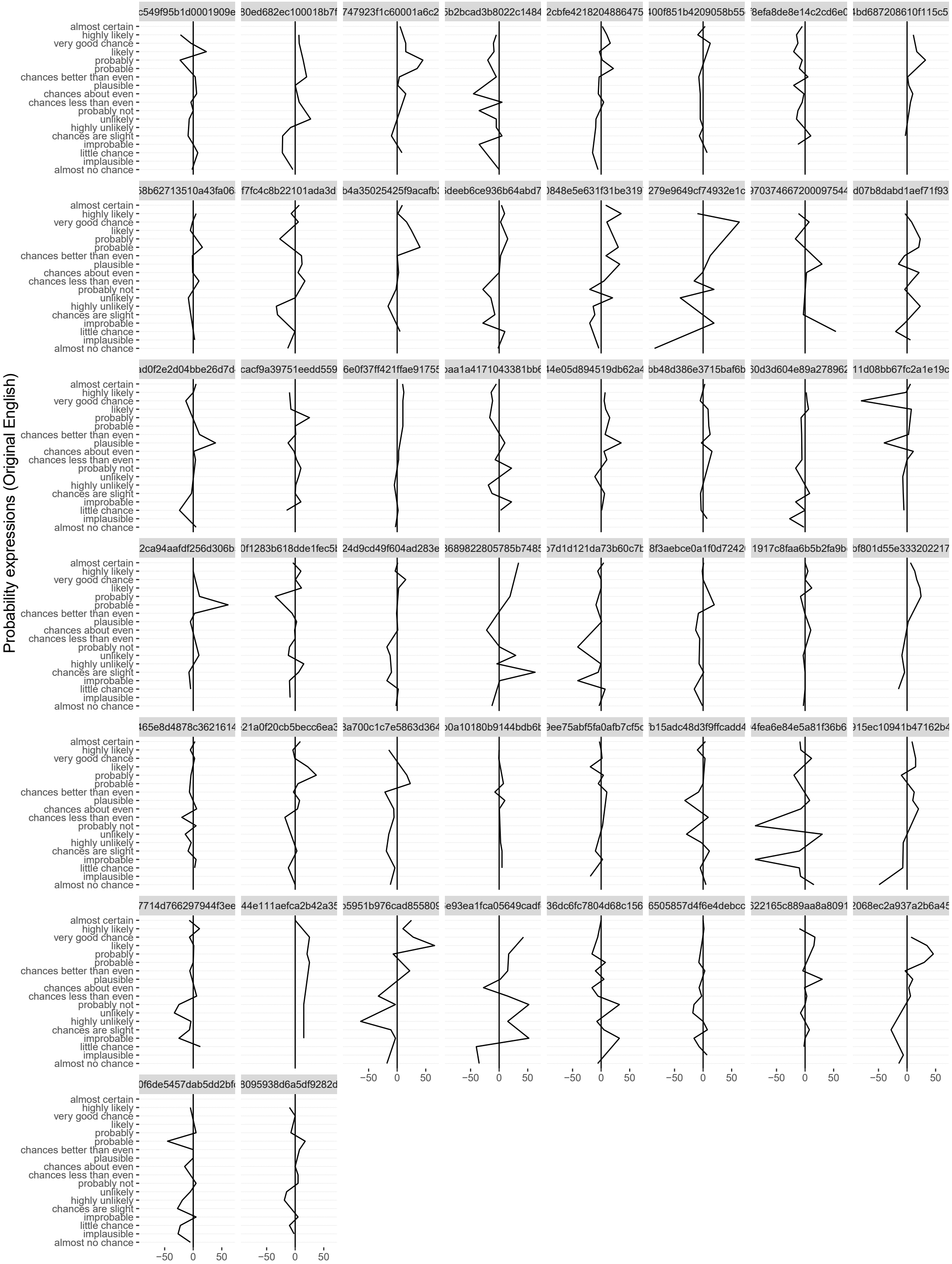
Table A.4: New suggested translations in Arabic

Appendix B

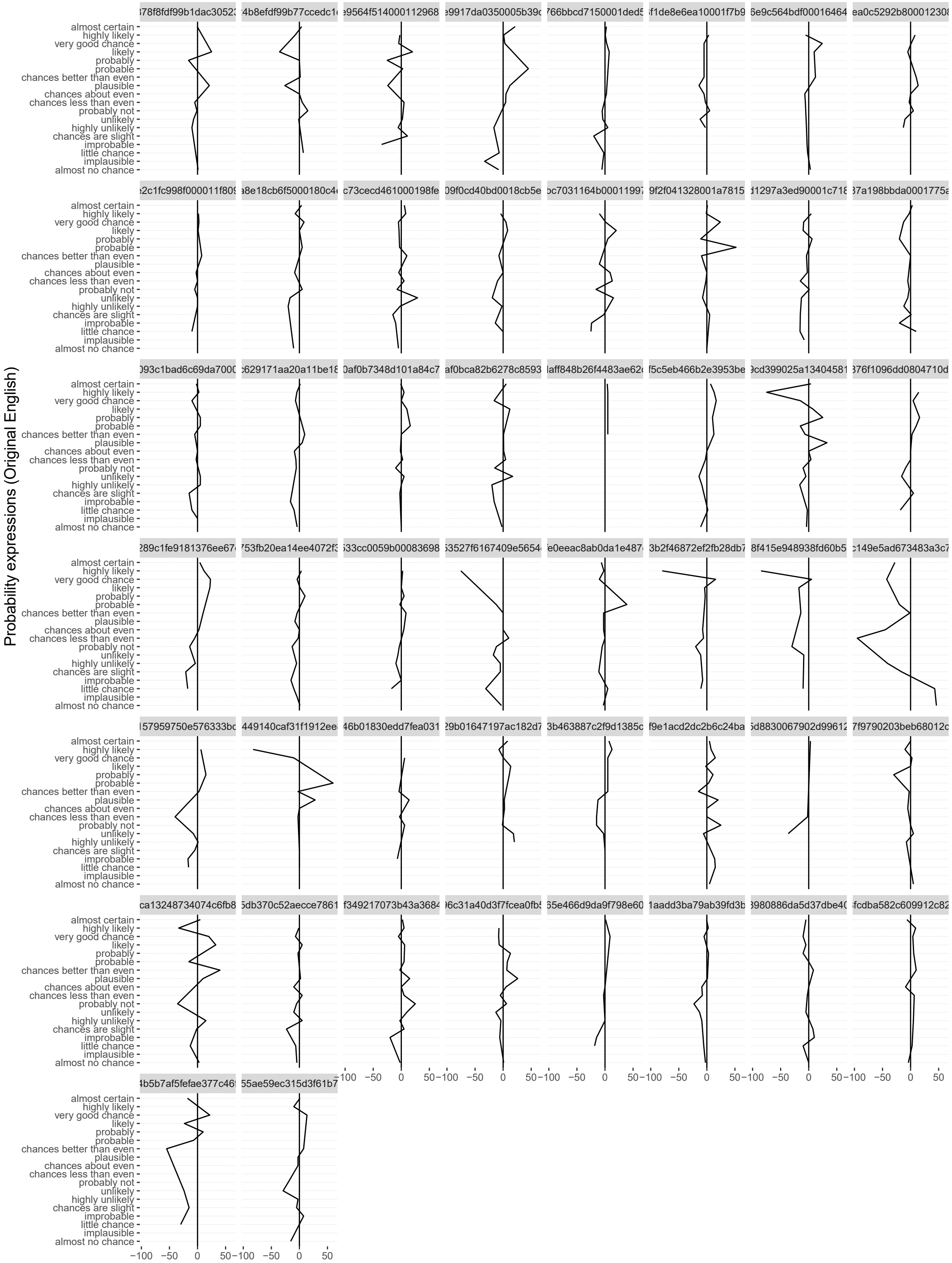
B.1 Exploratory Analysis: Comparing Experiments

This section presents a within subject exploratory analysis showing a distance measure between the two experiments for each probability expressions (see section 4.3)

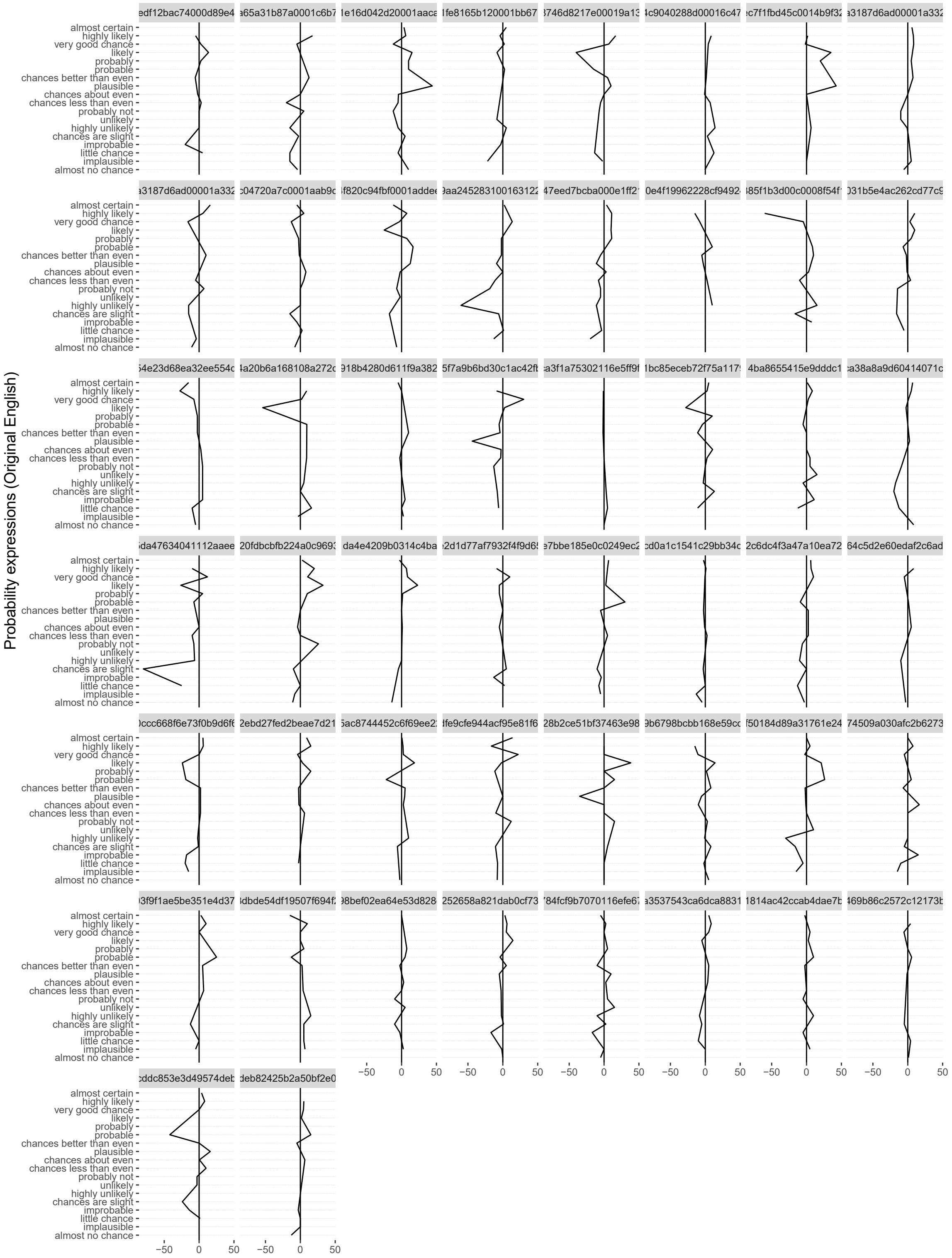
Distance between the values used in the two experiments by each participants in the **Arabic** version. Values on the negative sides indicate that the Expression-to-Vis experiment shows lower values drawn on the icon-arrays.



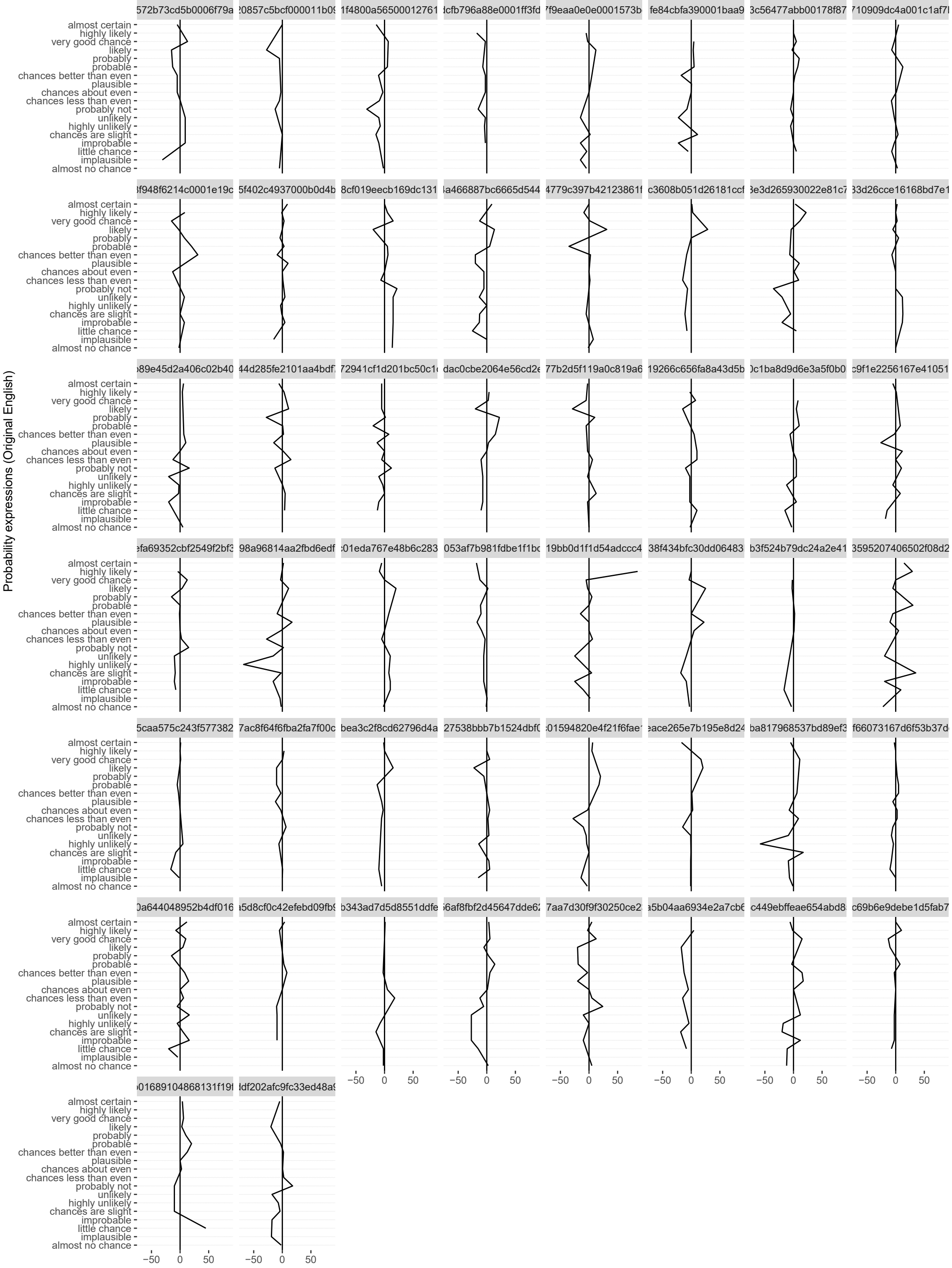
Distance between the values used in the two experiments by each participants in the English version. Values on the negative sides indicate that the Expression-to-Vis experiment shows lower values drawn on the icon-arrays.



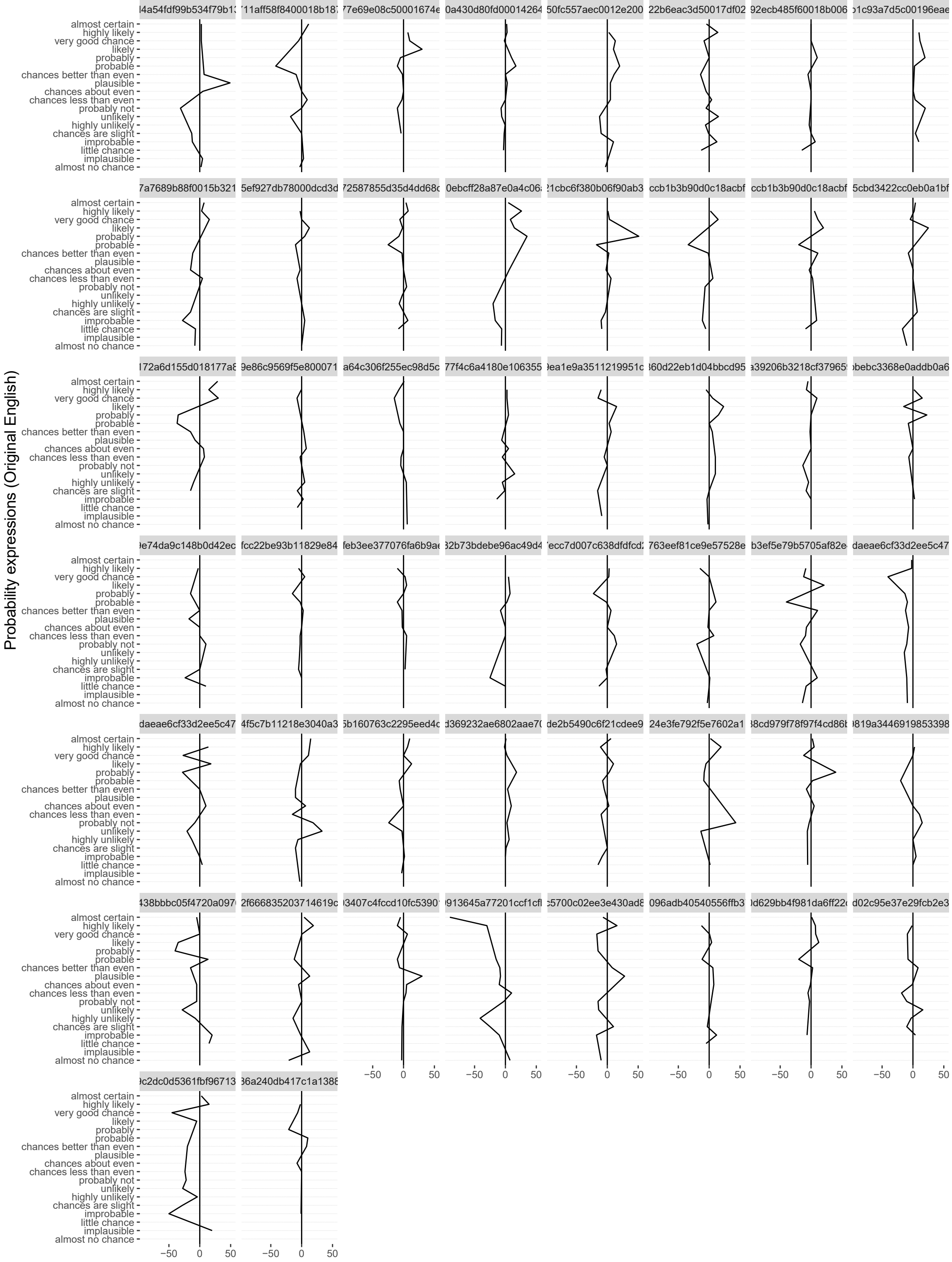
Distance between the values used in the two experiments by each participants in the **French** version. Values on the negative sides indicate that the Expression-to-Vis experiment shows lower values drawn on the icon-arrays.



Distance between the values used in the two experiments by each participants in the German version. Values on the negative sides indicate that the Expression-to-Vis experiment shows lower values drawn on the icon-arrays.



Distance between the values used in the two experiments by each participants in the **Mandarin** version. Values on the negative sides indicate that the Expression-to-Vis experiment shows lower values drawn on the icon-arrays.



Bibliography

[1] M. Bancilhon, Z. Liu, and A. Ottley. Let's Gamble: Uncovering the Impact of Visualization on Risk Perception and Decision-Making. *Computing Research Repository (CoRR)*, abs/1910.09725, 2019.

[2] S. Barclay et al. *Handbook for Decision Analysis.* 1977.

[3] A. Baughan, N. Oliveira, T. August, N. Yamashita, and K. Reinecke. Do Cross-Cultural Differences in Visual Attention Patterns Affect Search Efficiency on Websites? In *CHI '21: CHI Conference on Human Factors in Computing Systems, Virtual Event / Yokohama, Japan, May 8-13, 2021*, pp. 362:1–362:12. ACM, 2021. doi: 10.1145/3411764.3445519

[4] F. Bocklisch, A. Georg, S. Bocklisch, and J. Krems. Do You Mean What You Say? The Effect of Uncertainty Avoidance on the Interpretation of Probability Expressions-A Comparative Study between Spanish and German. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 35. cognitivesciencesociety.org, 2013.

[5] G. L. Brase. Pictorial Representations in Statistical Reasoning. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 23(3):369–381, 2009.

[6] P. G. Brust-Renck, C. E. Royer, and V. F. Reyna. Communicating Numerical Risk: Human Factors That Aid Understanding in Health Care. *Reviews of Human Factors and Ergonomics*, 8(1):235–276, 2013. doi: 10.1177/1557234X13492980

[7] D. V. Budescu, H. H. Por, S. B. Broomell, and M. Smithson. The Interpretation of IPCC Probabilistic Statements around the World. *Nature Climate Change*, 4(6):508–512, 2014. doi: 10.1038/nclimate2194

[8] D. V. Budescu and T. S. Wallsten. Consistency in Interpretation of Probabilistic Phrases. *Organizational behavior and human decision processes*, 36(3):391–405, 1985.

[9] Bundesministerium für Gesundheit. Wie ist der Fortschritt der COVID-19-Impfung? Aktueller Impfstatus. `https://impfdashboard.de/`, 2021. Accessed on Apr. 18, 2021.

[10] W. S. Cleveland and R. McGill. Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American statistical association*, 79(387):531–554, 1984.

[11] T. S. Doupnik and E. L. Riccio. The Influence of Conservatism and Secrecy on the Interpretation of Verbal Probability Expressions in the Anglo and Latin Cultural Areas. *The International Journal of Accounting*, 41(3):237–261, 2006.

[12] T. S. Doupnik and M. Richter. Interpretation of Uncertainty Expressions: a Cross-national Study. *Accounting, Organizations and Society*, 28(1):15–35, 2003. doi: 10.1016/S0361-3682(02)00010-7

[13] T. S. Doupnik and M. Richter. The Impact of Culture on the Interpretation of "In Context" Verbal Probability Expressions. *Journal of International Accounting Research*, 3(1):1–20, 2004. doi: 10.2308/jiar.2004.3.1.1

[14] P. Dragicevic. Fair Statistical Communication in HCI. In *Modern statistical methods for HCI*, pp. 291–330. Springer, 2016.

[15] V. Evers. Cross-Cultural Understanding of Metaphors in Interface Design. *Proceedings CATAC'98, Cultural Attitudes towards Technology and Communication*, pp. 1–11, 1998.

[16] J. L. Fleiss. Measuring Nominal Scale Agreement among Many Raters. *Psychological bulletin*, 76(5):378, 1971.

[17] O. Fuhrman and L. Boroditsky. Cross-Cultural Differences in Mental Representations of Time: Evidence from an Implicit Nonlinguistic Task. *Cognitive science*, 34(8):1430–1451, 2010.

[18] M. Galesic, R. Garcia-Retamero, and G. Gigerenzer. Using Icon Arrays to Communicate Medical Risks: Overcoming Low Numeracy. *Health Psychology*, 28(2):210–216, 2009. doi: 10.1037/a0014474

[19] R. Garcia-Retamero and M. Galesic. Communicating Treatment Risk Reduction to People with Low Numeracy Skills: A Cross-Cultural Comparison. *American Journal of Public Health*, 99(12):2196–2202, 2009. doi: 10.2105/AJPH.2009.160234

[20] R. Garcia-Retamero and M. Galesic. Who Profits from Visual Aids: Overcoming Challenges in People's Understanding of Risks. *Social Science and Medicine*, 70(7):1019–1025, 2010. doi: 10.1016/j.socscimed.2009.11.031

[21] E. Gibson, R. Futrell, J. Jara-Ettinger, K. Mahowald, L. Bergen, S. Ratnasingam, M. Gibson, S. T. Piantadosi, and B. R. Conway. Color Naming Across Languages Reflects Color Use. *Proceedings of the National Academy of Sciences*, 114(40):10785–10790, 2017.

[22] G. Gigerenzer and U. Hoffrage. How to Improve Bayesian Reasoning without Instruction: Frequency Formats. *Psychological Review*, 102(4):684–704, 1995. doi: 10.1037/0033-295X.102.4.684

[23] R. Henkin and C. Turkay. Words of Estimative Correlation: Studying Verbalizations of Scatterplots. *CoRR*, abs/1911.12793, 2019.

[24] M. Kay, T. Kola, J. R. Hullman, and S. A. Munson. When (ish) is my Bus? User-Centered Visualizations of Uncertainty in Everyday, Mobile Predictive Systems. *Conference on Human Factors in Computing Systems - Proceedings*, pp. 5092–5103, 2016. doi: 10.1145/2858036.2858558

[25] S. Kent. Words of Estimative Probability. *Journal of the American Intelligence Professional*, 8(4):49–65, 1964.

[26] Y. Kim, K. Thayer, G. S. Gorsky, and J. Heer. Color Names Across Languages: Salient Colors and Term Translation in Multilingual Color Naming Models. In *EuroVis (Short Papers)*, pp. 31–35, 2019.

[27] H.-K. Kong, Z. Liu, and K. Karahalios. Frames and Slants in Titles of Visualizations on Controversial Topics. In *Proceedings of the 2018 CHI conference on human factors in computing systems*, pp. 1–12, 2018.

[28] C. Kreuzmair, M. Siegrist, and C. Keller. High Numerates Count Icons and Low Numerates Process Large Areas in Pictographs: Results of an Eye-Tracking Study. *Risk Analysis*, 36(8):1599–1614, 2016. doi: 10.1111/risa.12531

[29] D. Leclercq. J'en Suis Aussi Sûr que Vous, Mais pas avec le Même Pourcentage de Chances, que ce Soit Hors Contexte ou En Contexte. *Journal international de Recherche en Education et Formation Evaluer. Journal international de Recherche en Education et Formation*, 2(21):89–125, 2016.

[30] D. T. Lindsey and A. M. Brown. World Color Survey Color Naming Reveals Universal Motifs and their Within-Language Diversity. *Proceedings of the National Academy of Sciences of the United States of America*, 106(47):19785–19790, 2009. doi: 10.1073/pnas.0910981106

[31] A. Mauboussin and M. J. Mauboussin. If You Say Something Is "Likely," How Likely Do People Think It Is? `https://hbr.org/2018/07/if-you-say-something-is-likely-how-likely-do-people-think-it-is`, 2018. Accessed on 2022-04-20.

[32] N. McDonald, S. Schoenebeck, and A. Forte. Reliability and Inter-Rater Reliability in Qualitative Research: Norms and Guidelines for CSCW and HCI Practice. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), 2019. doi: 10.1145/3359174

[33] F. Mosteller and C. Youtz. Quantifying Probabilistic Expressions. *Statistical Science*, 5(1):2–12, 1990. doi: 10.1214/ss/1177012242

[34] B. J. O'Brien. Words or Numbers? The Evaluation of Probability Expressions in General Practice. *Journal of the Royal College of General Practitioners*, 39(320):98–100, 1989.

[35] Y. Okan, R. Garcia-Retamero, E. T. Cokely, and A. Maldonado. Improving Risk Understanding across Ability Levels: Encouraging Active Processing with Dynamic Icon Arrays. *Journal of Experimental Psychology: Applied*, 21(2):178–194, 2015. doi: 10.1037/xap0000045

[36] A. Ottley, E. M. Peck, L. T. Harrison, D. Afergan, C. Ziemkiewicz, H. A. Taylor, P. K. Han, and R. Chang. Improving Bayesian Reasoning: The Effects of Phrasing, Visualization, and Spatial Ability. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):529–538, 2016. doi: 10.1109/TVCG.2015.2467758

[37] L. Padilla, S. C. Castrob, and H. Hosseinpoura. A Review of Uncertainty Visualization Errors: Working Memory as an Explanatory Theory. *The Psychology of Learning and Motivation*, p. 275, 2021.

[38] L. Padilla, M. Kay, and J. Hullman. Uncertainty Visualization. *Handbook of Computational Statistics and Data Science*, 2020.

[39] E. M. Peck, S. E. Ayuso, and O. El-Etr. Data is Personal: Attitudes and Perceptions of Data Visualization in Rural Pennsylvania. *arXiv*, pp. 1–12, 2019.

[40] R. T. Reagan, F. Mosteller, and C. Youtz. Quantitative Meanings of Verbal Probability Expressions. *Journal of Applied Psychology*, 74(3):433–442, 1989. doi: 10.1037/0021-9010.74.3.433

[41] K. Reinecke and K. Z. Gajos. LabintheWild: Conducting Large-Scale Online Experiments With Uncompensated Samples. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW 2015, Vancouver, BC, Canada, March 14 - 18, 2015*, pp. 1364–1378. ACM, 2015. doi: 10.1145/2675133.2675246

[42] S. Renooij and C. Witteman. Talking probabilities: Communicating probabilistic information with words and numbers. *International Journal of Approximate Reasoning*, 22(3):169–194, 1999. doi: 10.1016/S0888-613X(99)00027-4

[43] C. A. Sanne Willems and I. S. Verbal. Variability in the Interpretation of Probability Phrases Used in Dutch News Articles —a Risk for Miscommunication. *Orphanet Journal of Rare Diseases*, 21(1):1–9, 2020.

[44] M. M. Schapira, A. B. Nattinger, and C. A. McHorney. Frequency or Probability? A Qualitative Study of Risk Communication Formats Used in Health Care. *Medical Decision Making*, 21(6):459–467, 2001.

[45] D. Spiegelhalter, M. Pearson, and I. Short. Visualizing Uncertainty about the Future. *Science*, 333(6048):1393–1400, 2011. doi: 10.1126/science.1191181

[46] C. Sturm, A. Oh, S. Linxen, J. Abdelnour Nocera, S. Dray, and K. Reinecke. How WEIRD is HCI? Extending HCI Principles to Other Countries and Cultures. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, vol. 18, pp. 2425–2428, 2015. doi: 10.1145/2702613.2702656

[47] J. Talbot, V. Setlur, and A. Anand. Four Experiments on the Perception of Bar Charts. *IEEE transactions on visualization and computer graphics*, 20(12):2152–2160, 2014.

[48] C. Till. Fostering Risk Literacy in Elementary School. *International Electronic Journal of Mathematics Education*, 9(1-2):85–98, 2014.

[49] V. H. Visschers, R. M. Meertens, W. W. Passchier, and N. N. De Vries. Probability Information in Risk Communication: A Review of the Research Literature. *Risk Analysis*, 29(2):267–287, 2009. doi: 10.1111/j.1539-6924.2008.01137.x

[50] T. S. Wallsten, D. V. Budescu, A. Rapoport, R. Zwick, and B. Forsyth. Measuring the Vague Meanings of Probability Terms. *Journal of Experimental Psychology: General*, 115(4):348–365, 1986. doi: 10.1037/0096-3445.115.4.348

[51] H. Winschiers-Theophilus and N. J. Bidwell. Toward an Afro-Centric Indigenous HCI Paradigm. *International Journal of Human-Computer Interaction*, 29(4):243–255, 2013.

[52] B. C. Wintle, H. Fraser, B. C. Wills, A. E. Nicholson, and F. Fidler. Verbal Probabilities: Very Likely to be Somewhat More Confusing than Numbers. *PLoS ONE*, 14(4):1–18, 2019. doi: 10.1371/journal.pone.0213522

[53] C. Xiong, A. Sarvghad, Ç. Demiralp, J. M. Hofman, and D. G. Goldstein. Investigating Perceptual Biases in Icon Arrays. 2022.

[54] B. J. Zikmund-Fisher, H. O. Witteman, M. Dickson, A. Fuhrel-Forbis, V. C. Kahn, N. L. Exe, M. Valerio, L. G. Holtzman, L. D. Scherer, and A. Fagerlin. Blocks, Ovals, or People? Icon Type affects Risk perceptions and Recall of Pictographs. *Medical Decision Making*, 34(4):443–453, 2014. doi: 10.1177/0272989X13511706