



Risk Literacy for Everyone: a Meta-analysis on the Comprehension of Decision Aids for Health Risks

A Dissertation By

Francisco M. Gomes

Master Thesis in Modeling, Data Analysis
and Decision Support Systems

Supervised by

Prof. Dr. Maria Eduarda Silva

Faculdade de Economia

Universidade do Porto

2021

Acknowledgments

I would like to thank Prof. Dr. Maria Eduarda Silva for having supervised this Master Thesis and for showing keen interest since day one, allowing me to confidently explore a theme with which I personally connected. A special thank you to Vasco Brazão and Henry Nasser from CLOO for the particularly warm guidance and presence all throughout this 10-month journey, making me feel self-assured and valued for all the work I put in.

I extend my gratitude to all my MADSAD colleagues, especially to Paulo Azevedo for always supporting me and making my academic experience as appealing and engaging as possible.

A very special thank you to my Mom, Dad, Grandma, sisters, Angie and all my close friends for carrying me for all these years and hopefully those to come.

Lastly, thank you to Vi and Luís for reading and commenting on this Thesis, and also to all those who showed interest in knowing more about it or gave me any sort of feedback.

Abstract

Everyday people around the world are faced with decisions that directly influence their own health or that of someone akin. These decisions, which usually address a comparison between the risks and benefits of different treatments or medical procedures, may be conveyed through a myriad of possible communication formats, from a simple verbal statement to complex dynamic graphics. While these vehicles of communication intend to present information as clearly and effectively as possible, individuals exhibit different numeracy and graph literacy skills, which in turn affect their comprehension of risks. And despite visual decision aids being widely praised and used due to their general positive effect on comprehension, it is not clear if low numerates benefit from this type of communication format.

To explore this issue we executed a systematic review and meta-analysis focused on measuring the comprehension of visual and non-visual decision aids by individuals with low numeracy skills. Our analysis included 12 experiments from 10 studies in a universe of 1906 collected articles.

No clear advantage was found between visual or non-visual aids regarding low numerates' comprehension. Initial outcomes pointed towards a slightly better performance in visual aids [$g = -0.28$ (95% CI -0.64 to 0.08)], but said results were affected by large heterogeneity, which, when addressed, shifted the effect size to non-visual aids [$g = 0.06$ (95% CI -0.02 to 0.15)]. No moderating variables were found. High numerates displayed similar results between visual and non-visual aids, but were found to achieve significantly better comprehension scores than low numerates in both formats.

More research is of necessity, either in primary studies as in extensive systematic literature reviews, especially those focusing in the interaction between numeracy and graph literacy, and ideally using participant-level data.

Keywords: Numeracy, Graph Literacy, Health Risks, Risk Communication, Decision Aids

Resumo

Diariamente inúmeras pessoas são confrontadas com decisões que influenciam directamente a sua saúde ou a de alguém próximo. Estas decisões, que normalmente requerem a comparação entre riscos e benefícios de diferentes tratamentos ou intervenções médicas, podem ser comunicadas utilizando diversos formatos de comunicação, desde uma simples comunicação verbal até um complexo gráfico dinâmico. Ainda que o uso destes formatos tenha como objectivo a apresentação clara e eficaz da informação, diferentes pessoas têm capacidades de numeracia e literacia de gráficos distintas, o que por sua vez afecta a sua compreensão dos riscos; e apesar dos materiais visuais serem amplamente utilizados e elogiados devido ao seu efeito geralmente positivo sobre a compreensão, não é claro se pessoas com baixa numeracia beneficiam deste formato de comunicação.

Para explorar este tópico executámos uma revisão sistemática e meta-análise focada na compreensão de materiais de apoio à decisão visuais e não-visuais por pessoas com baixas capacidades numéricas. A nossa análise incluiu 12 experiências de 10 estudos num universo de 1906 artigos.

Não foi encontrada nenhuma vantagem entre materiais visuais ou não-visuais no que toca à sua compreensão por parte de participantes com baixa numeracia. Os resultados iniciais apontavam para uma maior eficácia de materiais visuais [$g = -0.28$ (95% IC -0.64 a 0.08)], mas sofriam de elevada heterogeneidade. Esta, quando removida, deslocou o efeito para os materiais não visuais [$g = 0.06$ (95% IC -0.02 a 0.15)]. Não foram encontradas quaisquer variáveis moderadoras. Participantes com elevadas capacidade numéricas tiveram resultados semelhantes entre materiais visuais e não visuais, ainda que em ambos os formatos tenham conseguido uma compreensão significativamente superior do que aqueles com menor numeracia.

Será necessária mais investigação, quer em estudos primários como em revisões de literatura sistemáticas, especialmente focadas na interacção entre numeracia e literacia de gráficos, idealmente utilizando dados a nível do participante.

Palavras-chave: Numeracia, Literacia de Gráficos, Riscos de Saúde, Comunicação de Risco, Materiais de Apoio à Decisão

Contents

1	Introduction, Motivation and Problem	1
1.1	Introduction	1
1.2	Motivation	4
1.3	Problem	5
2	Literature Review	7
2.1	Numeracy and Graph Literacy	7
2.2	The impact of Numeracy	9
2.3	Communication formats	10
2.4	Perception of Risk	13
2.5	Other aspects	15
2.6	General discussion	16
2.7	Conclusions	17
3	Meta-analysis Plan and Rationale	18
3.1	Meta-analysis	18
3.1.1	Steps	18
3.1.2	Rationale	19
3.1.3	Procedure	20
3.1.4	Research Questions	21
3.2	Search and Screening Strategy	22
3.3	Data and Software	24
3.4	A priori Limitations	26
4	Search, Screening and Data Extraction	27
4.1	Search Execution and Study Screening	27
4.2	Data Extraction and Handling	29
4.3	Limitations	34
5	Meta-analysis	37
5.1	Data Processing	37
5.2	Data Imputation	39
5.3	Results	42

5.3.1	Heterogeneity	43
5.3.2	Subgroup Analysis	45
5.3.3	Meta-regression	46
5.3.4	Publication Bias	48
5.3.5	High Numeracy	50
5.3.6	High vs. Low Numeracy	51
5.4	Discussion	52
6	Limitations and Future Work	54
6.1	Limitations	54
6.2	Future Work	55
Appendix		i
A	Data extraction with plot digitizer	i
B	Objective Graph Literacy Test	ii
C	Computer Adaptive Berlin Numeracy Test	iii
D	Interpretation of Probability Phrases	iv
E	PRISMA Flow Diagram	v
F	PRISMA-P Checklist	vi
G	Logic Grids	viii
H	Search Strings	xi
I	SD Estimation Conversion Factors	xiv
J	Influence Diagnostics	xv
K	Cumulative Meta-analysis	xvi
L	High Numeracy GOSH Analysis	xvii
M	Funnel Plot Variations	xviii

List of Figures

1.1	Datawrapper Daily Chart Visualizations	2
1.2	Interest in Risk Communication research since 2000	3
1.3	Interaction between Numeracy and Graph Literacy	5
2.1	Tabular and icon-array fact boxes	11
2.2	Examples of Visual Aids	12
4.1	Power Analysis for Effect Size and Heterogeneity	31
4.2	Power Analysis for Number of Studies and Effect Size	32
4.3	PRISMA Flow Diagram depicting the number of studies through each step . . .	32
5.1	Forest Plot	42
5.2	Baujat plot	43
5.3	GOSH Plot analysis of outliers	44
5.4	GOSH cluster analysis	44
5.5	Subgroup Analysis	46
5.6	Meta-regression	47
5.7	Meta-regression excluding outliers	47
5.8	Funnel Plot	49
5.9	Funnel Plot after outlier removal	49
5.10	Forest Plot on High Numerates	51
5.11	Forest Plots on High vs. Low Numeracy	51
A1	Interaction between Numeracy and Graph Literacy	i
B1	Objective Graph Literacy Test	ii
C1	Computer Adaptive Berlin Numeracy Test	iii
D1	Interpretation of Probability Phrases	iv
E1	PRISMA Flow Diagram	v
J1	Influence diagnostics plot	xv
K1	Cumulative meta-analysis Forest Plot by Year of Publication	xvi
K2	Plot of cumulative results	xvi
L1	GOSH analysis for High Numeracy (N05)	xvii
L2	GOSH cluster analysis for High Numeracy	xvii
M1	Funnel plot on Standard Error	xviii
M2	Funnel plot on Sampling Variance	xviii

List of Tables

2.1	Example of questions in Numeracy and Graph Literacy tests	8
3.1	Inclusion and Exclusion Criteria	23
3.2	Search and Screening methods	24
3.3	Data Extraction	25
4.1	Decision matrix resulting from initial Title and Abstract screening	28
4.2	Number of studies per parameter of interest and data availability	29
4.3	Simplified table with characteristics of collected studies	33
4.4	Simplified table with characteristics of collected experiments	34
5.1	Transformation of formats into Visual and Non-Visual categories	37
5.2	Predicted events and non-events for Study N01	41
5.3	Reported and Predicted descriptive statistics for Study N01	41
5.4	Main outcomes of meta-analysis (total vs. influential removed)	45
5.5	Regression results	48
5.6	Regression limit estimates	50
F1	PRISMA-P Checklist	vi
G1	Logic Grid 1	viii
G2	Logic Grid 2	viii
G3	Logic Grid 3	ix
G4	Logic Grid 4	x
I1	Conversion factors for estimating SD for a sample size N	xiv

Chapter 1

Introduction, Motivation and Problem

This Thesis sprouted from a challenge put up by CLOO - Behavioral Insights Unit, initially directed at Data/Uncertainty Visualization and Communication. From an extensive search through these fields, aimed at procuring an engaging and feasible topic to explore, several possible paths arose, such as analysing the relationship of uncertainty with motivated reasoning, measuring the effects of uncertainty visualizations on decision-making and creating guidelines for tailored risk communication. Among these potential research topics, the one that follows was the one that, garnering support from all those involved, seemed to be the most impactful and aligned with my personal goals but also with those of the Masters.

1.1 Introduction

With all the changes it triggered upon the world, the Covid-19 pandemic brought an overwhelming flurry of charts alongside it (line graphs and maps at its forefront). Datawrapper, a tool for data visualization used by major news outlets across the globe, reported a surge like no other in daily chart views (Rost, 2020), as can be seen in Figure 1.1 below, with Tableau Public, another major data visualization tool, also reporting a substantial growth in impressions (Schutt, 2020). As the world gasped at the daily number of new cases and deaths, often at different granularities (city, region, country, continent, world) which could then be compared with others to track events of success or failure, diverse visualizations options arose: linear scales soon became unable to convey the pace of the pandemic, that seemed to be evolving exponentially with no end in sight, giving place to logarithmic scales, which smoothed out the steepness of the plots.

Be that as it may, the choice of one scale over the other is not without its consequences. In fact, Romano, Sotis, Dominioni, and Guidi (2020) unveiled that using a logarithmic scale in the context of the new coronavirus pandemic had a negative impact on understanding and levels of worry, and shifted policy preferences while potentially being confusing for the general public. These findings are aligned with previous research by Heckler, Mikula, and Rosenblatt (2013) and Menge et al. (2018), all pointing to the fact that a seemingly innocent choice of scale can have numerous effects in the interpretation, transformation and use of information; a reasoning that will be continually presented along this Thesis for several different aspects of Data and Risk Communication.

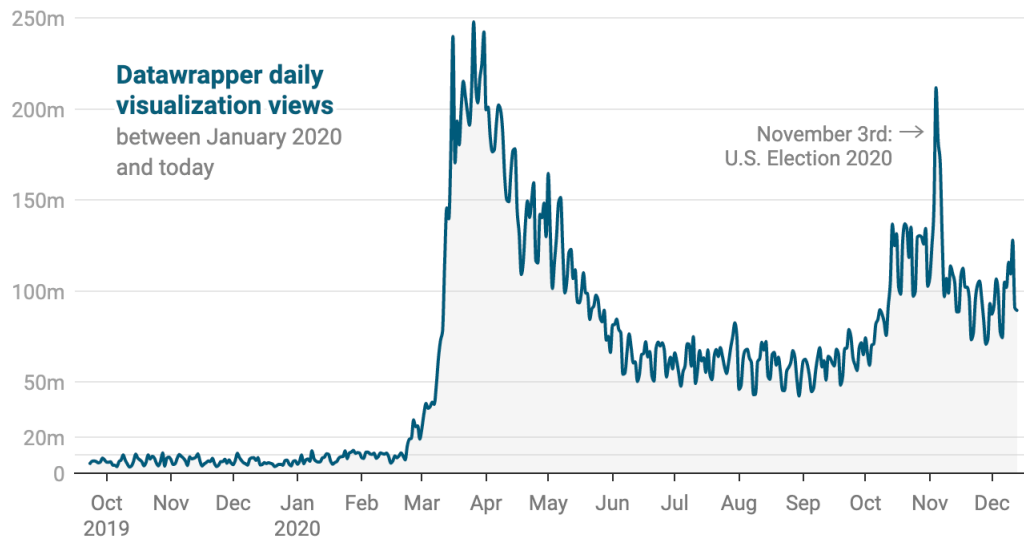


Figure 1.1: Datawrapper Daily Chart Visualizations from October 2019 to December 2020. Retrieved from Rost (2020).

Regardless of the novelty of the pandemic, the use of numerical information, data or visualizations has long been part of everyday lives in a multitude of contexts: weather forecasts, election polls, unemployment figures and productivity estimates, environmental and climate change projections, even advertisements; but one of those contexts in which comprehension of statistical information and risks is essential for informed decision making is that of Health (Gaissmaier et al., 2012; Garcia-Retamero & Cokely, 2013). Not only patients often have to weigh benefits and harms in order to compare treatments (Garcia-Retamero & Cokely, 2017), as personal health information is nowadays more accessed and accessible by end-users in their own devices (*Turchioe et al., 2020). However, in a context which is already often enveloped in uncertainty, non-transparent information and insufficient levels of risk literacy put informed and accurate decisions by both doctors and patients in jeopardy (Garcia-Retamero, Okan, & Cokely, 2012).

Lack of awareness on the varying effects and biases of different formats leads to non-effective communication and possibly unintended effects (Garcia-Retamero et al., 2012). In fact, aspects including individual features (Garcia-Retamero & Cokely, 2013), context (Garcia-Retamero et al., 2012), format type (Schapira, Nattinger, & McAuliffe, 2006; *Tait, Voepel-Lewis, Zikmund-Fisher, & Fagerlin, 2010a) or risk information characteristics (Slovic, Peters, Finucane, & Macgregor, 2005), all have a role in the perception and comprehension of risk information and ultimately influence the decision-making process (Slovic et al., 2005).

Several experimental studies, articles and reviews have embraced the challenge of defining guidelines and formats to ensure comprehension, recall and other targets across different types of risk communications, with a special focus in health-based contexts (Vischers, Meertens, Passchier, & de Vries, 2009). The field, which spans through areas such as Psychology, Statistics, Data Visualization, Numeracy and Uncertainty Communication, shares some agreement upon several choices that should be avoided, the qualities that render a communication transparent and the

formats that ought to be employed - some regardless of the context (e.g. always include denominators), others dependent on the information (e.g. use either frequencies, percentages or both) or the target audience (e.g. use visual aids instead of numbers) (Hawley et al., 2008). Despite the aforementioned conclusions and the constant growth in research in the last two decades (see Figure 1.2), several questions remain open and others have concurrently risen.

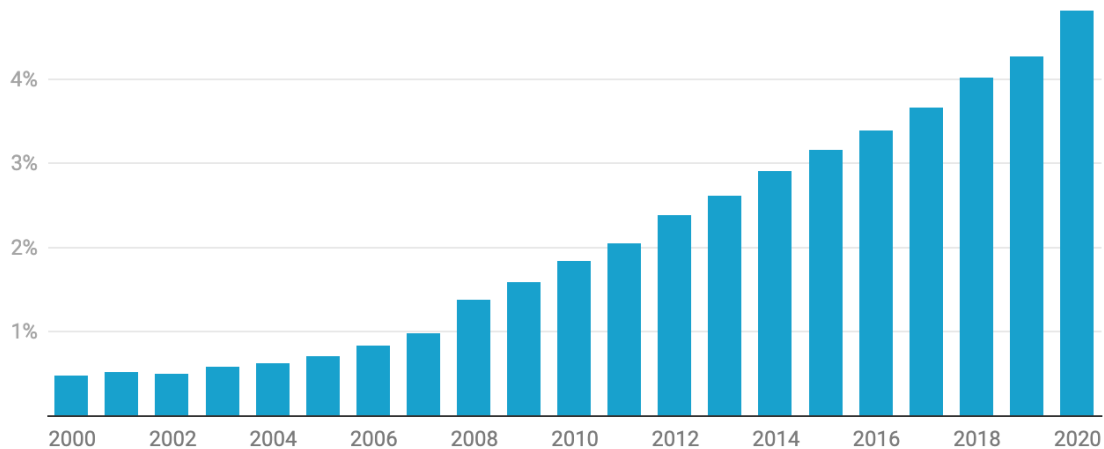


Figure 1.2: Interest in Risk Communication research in proportion to the total number of publications in Europe PMC database (2000-2020)¹.

Numeracy, regarded as a factor with significant weight in the comprehension of numerical information and thus important for the selection of risk communication formats (Zikmund-Fisher, Smith, Ubel, & Fagerlin, 2007), has been identified in very disparate degrees between different individuals, some of which with below-basic levels (Peters et al., 2006). To illustrate this, a study by Galesic and Garcia-Retamero (2010) concluded that at least 25% of the adult population in both Germany and the United States suffer from acute numeracy deficits. In this study, among questions adapted from the widely used numeracy test developed by Schwartz, Woloshin, Black, and Welch (1997) and Lipkus, Samsa, and Rimer (2001), participants had to select the largest risk among three options: 1 in 10, 1 in 100 and 1 in 1000; 28% of Germans and 25% of Americans could not answer correctly. Moreover, only 46% and 23.5%, respectively, could accurately transform a frequency of 1 in 1000 to a percentage of 0.1%. In addition, even highly-educated individuals have been found to have low numeracy skills (Lipkus et al., 2001) and a study on medical staff by Anderson and Schulkin (2014) concluded that 25% to 47% of the involved physicians could not answer elementary numeracy questions. These profound deficits have therefore been found present in large and heterogeneous proportions of individuals, and attest for the widely spread lack of competencies to deal with numerical risks.

Additionally, the interaction between numeracy skills and health indicators has long been researched, and the absence of the former is considered a predictor to a patient's total body mass

¹Data retrieved using the *europemc* package (Jahn, 2020).

index, use of emergency services, overconsumption of prescribed medication, issues in adhering to complex dosing programs, risk of comorbidity, among others (Garcia-Retamero & Cokely, 2017). This relationship will be further described in Chapter 2.

Resuming the exposition just above, risk information is not bound to words or numbers and other formats have taken the stage in the last decades (Turchioe et al., 2019; *Turchioe et al., 2020). The use of tabular and visual decision aids has been praised and backed by several researchers and experts, as they considerably improve both engagement, comprehension, accuracy and other performance metrics in less numerate individuals when compared to the previously mentioned formats (Brick, McDowell, & Freeman, 2020; McDowell, Gigerenzer, Wegwarth, & Rebitschek, 2019). While the choice of graphic highly depends on the information to convey (Garcia-Retamero & Cokely, 2017), when dealing with risks and benefits a pictograph is often the preferred medium, as it simultaneously expresses the numerator and denominator and portrays risk as a frequency instead of a probability (Fagerlin, Zikmund-Fisher, & Ubel, 2011). A thorough analysis of several risk communication formats will be carried out in Chapter 2.

Following the referred advent in visual formats, some studies have recently brought to light that not all individuals share the same abilities to read and extract knowledge from graphs, leading to the identification of a particularly vulnerable subgroup of the population in regards to visual aids and risk communication as a whole: those with low numeracy as well as low graph literacy (Garcia-Retamero & Cokely, 2017). In point of fact, this distinct group has been found not to benefit as much as others from visual decision aids (Garcia-Retamero & Galesic, 2010b), lacking the necessary skills to its correct interpretation while potentially being puzzled by these formats (Ruiz et al., 2013), and may benefit more from numerical information accompanied by labels (Okan, Garcia-Retamero, Cokely, & Maldonado, 2015; Ruiz et al., 2013).

In conclusion, it is clear that a lack of consensus exists in respect to which Decision Aid is the most effective - if any - when dealing with this particular subgroup.

1.2 Motivation

After reviewing quite an extensive literature on the topics of Uncertainty Communication, Data Visualization and Risk Communication I came across a remarkable Systematic Review by Garcia-Retamero and Cokely (2017), two of the most prominent researchers in the latter field. Encapsulating the state of the art of research in Visual Decision Aids, the authors depict the vast benefits of graphics (especially static ones) in individual comprehension, engagement, recall and several other performance metrics, in the context of health-based risk information. Unfortunately, as claimed by the same authors, the scope of these benefits is not universal. In fact, according to several studies, for those with fewer numeracy and graph literacy skills the effects of visual aids could be negligible or even negative. Withal, the review states that a vast majority of the included studies (83%) concluded that subjects with low numeracy never achieved as high a level of comprehension when compared to those who were more numerate, which further evidentiates the gap between these groups.

Afterwards, a depiction of the effects of low numeracy unfolds - as mentioned above in Section 1.1 -, revealing the grim reverberations of this cognitive trait in individual and public

health indicators, as well as its impact on accurate decision-making. Not only there is a lack of consensus to which Decision Aid could be the most beneficial - and to what extent - to the referred subgroup of the population, as the individuals that compose it seem fated to never reach acceptable levels of risk literacy, which in turn can have palpable harmful consequences on their state of health.

This setting, along with 1) the fact that, to best of our knowledge, no Meta-analysis has been produced in the direction of individuals with both low numeracy and low graph literacy, 2) the potential that meta-analyses have by aggregating knowledge of a particular field, pointing the way to future research and possibly directing public policies, and 3) my personal interest in the fields of Communication and Data Visualization, establish the motivations behind this Thesis.

1.3 Problem

This Thesis intends to undertake an exploration on the statement that visual aids do not have that positive an impact on individuals with insufficient levels of numeracy and graph literacy when compared to those that are more skilled. As expressed in Sections 1.1 and 1.2 above, a systematic review by Garcia-Retamero and Cokely (2017) backs this assertion with a considerable amount of research and expertise. To illustrate this polarity, the researchers included a survey by Garcia-Retamero and Galesic (2010b) in which participants from Germany and the United States were presented with information in numerical and graphical formats - icon array or bar plots, with or without numerator -, and then had to extract simple facts regarding the number of patients affected by a stroke if they took 1) a placebo, or 2) a prescribed drug. Results are illustrated in Figure 1.3 below.

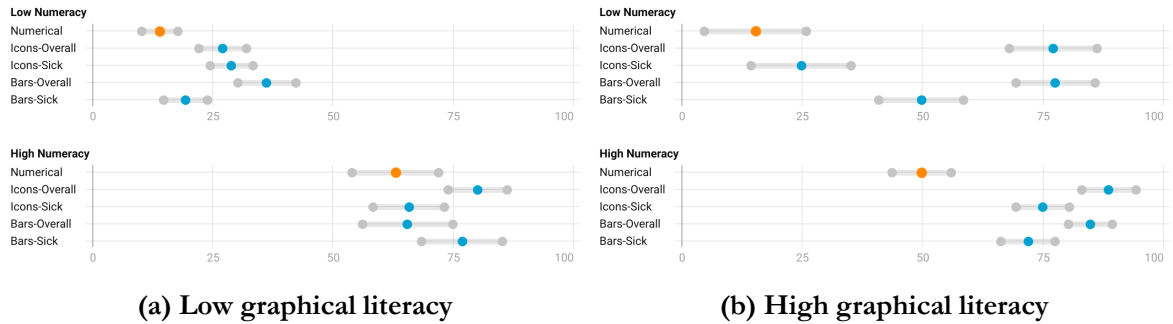


Figure 1.3: Percentage of participants with low or high graphical literacy and low or high numeracy who correctly estimated the two scenarios (by visual aids formats) in a study by Garcia-Retamero and Galesic (2010b). Orange data points indicate non-visual formats; error bars represent one standard error; adapted (see Appendix A for original plots and data extraction).

The outcomes above reveal several important considerations: it is clear that formats have diverse effects in response to different interactions between numeracy and graph literacy; icon arrays seem to contribute the most to a general positive effect on comprehension, corroborating a general belief of this research field; the overall conditions (that show the entire population at

risk and not just the affected individuals) also lead to higher, albeit general, levels of comprehension; the impact of visual aids is particularly substantial for those who have low numeracy but high graph literacy; and lastly, individuals with both low numeracy and low graph literacy display severely lower levels of comprehension than the other three subgroups, by not benefiting much from neither of the communication formats.

In essence, the proposed Meta-analysis will delve into these and the above-referred findings, hence pertaining to analyse the effects of Decision Aids of several formats on the comprehension of health-based risks in this specific subpopulation, with the purpose of quantifying these effects and hopefully pointing towards a clearer direction in the Risk Communication field.

Chapter 2

Literature Review

The search for articles and scientific content to produce this literature review started within the fields of Uncertainty Visualization/Communication, gradually shifting towards a field that, to some extent, encapsulates the former ones: Risk Communication.

Several wide searches were executed in different moments using the Web of Science database as follows:

1) *“Risk Literacy”*; 2) *“Risk Communication” AND Visual**; 3) *Risk Visual**; 4) *“Uncertainty Communication”*; 5) *“Uncertainty Visualization”*

The outputs from these searches were then filtered by Title, Keywords and Abstract. The ensuing full-text reading gave path to the analysis of citations (which references did a particular article cite, and in turn where was this particular article cited), leading to the subsequent addition of other articles and sources.

Among all consulted contents, a total of 95 sources - articles, experimental studies, reviews, meta-analyses, books, websites, programs, or programming packages - were cited in this Thesis Report. The publication year of these sources ranged from 1982 to 2021, with a median year of 2011 and a mode of 2020.

This search process was not executed following the systematic method that is to be used in the Search phase of the Meta-analysis, as its main goal was to produce an overview of several fields, filter and funnel appropriate content, and ultimately procure a relevant and feasible topic for further exploration.

2.1 Numeracy and Graph Literacy

It is now an established assumption that numeracy, or the ability to extract knowledge from numerical information and use mathematical skills, has a significant effect in the comprehension of risk information and the decision-making processes that entail it (Gardner, McMillan, Raynor, Woolf, & Knapp, 2011; Peters et al., 2006; Schwartz et al., 1997). For this reason, it is nowadays common for articles in the field to test for this condition in order to describe and classify the population sample in terms of numeracy and explore its interaction with other variables. To a lesser degree (although of growing importance), some studies measure their subjects' graph literacy, which intends to analyse how well people extract information from graphical formats.

Additionally, as seen above in Chapter 1, the interaction between graph literacy and numeracy has proved to be of great importance by revealing the particular characteristics of the subgroup of individuals under the scope of this Thesis.

The use of standard tests is of crucial importance for the sake of result comparison, and only a few stand out in regards to widespread use. Thirty-six articles were selected in the aforementioned systematic review by Garcia-Retamero and Cokely (2017), on the condition that each 1) contained empirical studies, 2) investigated the effect of numeracy and/or graph literacy and 3) explored the effect of visual aids on objective/subjective measures (standalone or compared to the effects of numbers in text). All but one tested for numeracy skills, while 33% tested for graph literacy and 31% for both. Among the studies that measured numeracy, 61% used objective tests - 86% of these used a test with questions from Lipkus et al. (2001) and Schwartz et al. (1997), while 23% used the Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) - and 39% employed a subjective test by Fagerlin et al. (2007). Every study that tested for graph literacy used the objective scale proposed by Galesic and Garcia-Retamero (2011b), while one study used this in addition to a subjective test (Garcia-Retamero, Cokely, Ghazal, & Joeris, 2016). Table 2.1 below exhibits one question from each of the mentioned studies (see also Appendix A).

Table 2.1: Example of questions in different numeracy/graph literacy tests.

Test	Skill	Example
Objective Tests		
Lipkus et al. (2001); Schwartz et al. (1997)	N	In the ACME PUBLISHING SWEEPSTAKES, the chance of winning a car is 1 in 1,000. What percent of tickets to ACME PUBLISHING SWEEPSTAKES win a car?
Cokely et al. (2012)	N	Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)?
Galesic and Garcia-Retamero (2011b)	G	Approximately what percentage of people had Adeolitis in the year 2000?*
Subjective Tests		
Fagerlin et al. (2007)	N	When reading the newspaper, how helpful do you find tables and graphs that are parts of a story? (1 = not at all, 6 = extremely)
Garcia-Retamero et al. (2016)	G	How good are you at determining the difference between 2 bars in a bar chart? (1 = not at all good; 6 = extremely good)

Note. N = Numeracy test; G = Graph Literacy test.

**See respective graph and two other questions in Appendix A.*

While the objective tests mentioned above use questions with right or wrong answers, measuring the factual abilities of an individual to deal with numerical or graphical information, subjective tests measure a subject's perception of ability and preferences. In spite of this difference, both types share the common goal of predicting objective performance. The subjective tests referred

above have been regarded as significantly correlated with their objective counterparts, with the benefit of promoting intention of participation in future research (Zikmund-Fisher et al., 2007), being far quicker to complete, and inducing smaller levels of test anxiety (Garcia-Retamero & Cokely, 2017).

The way subjects are classified into high or low numeracy/graph literacy groups depends on the executed test and the researcher's decision. For instance, the Berlin Numeracy Test (Cokely et al., 2012) divides participants into four categories in the case of its computer adaptive test (see Appendix B), as do Schwartz et al. (1997) and the short version used by Lipkus et al. (2001). In the event of severe shortage of time, the Berlin Numeracy Test proposes a single-item format, where subjects are divided by the correctness of a single answer. Regarding experimental studies there are quite a few splits put into practice: follow the groups proposed by the tests (Brick et al., 2020; Schapira et al., 2006), use splits on number of correct answers (usually following or close to the sample median) (Gaissmaier et al., 2012; Garcia-Retamero & Galesic, 2010b), or using top/bottom percentages (Galesic & Garcia-Retamero, 2010, 2013).

2.2 The impact of Numeracy

Numeracy occupies a central role in Risk Communication studies and has been found to have a deep effect on decision-making processes, affective responses, processing of information, health indicators, and other related aspects (Peters, 2008).

Less numerate individuals tend to be more persuaded by factors such as anecdotes, affective moods or trust/distrust in experts (Peters, Dieckmann, Dixon, Hibbard, & Mertz, 2007; Trevena et al., 2013), more susceptible to framing (Peters et al., 2006), and less participant in the decision-making process (Garcia-Retamero, Wicki, Cokely, & Hanson, 2014). They ask less questions in medical contexts (Paasche-Orlow & Wolf, 2007), report more negative interactions with doctors (Manganello & Clayman, 2011), and show inferior levels of comprehension and use of the information that is given to them (Peters, 2008). For these reasons, several studies report the need for different communication formats to deal with different numeracy levels (Gaissmaier et al., 2012; Peters et al., 2006; Trevena et al., 2013).

Hess, Visschers, and Siegrist (2011) and Kreuzmair, Siegrist, and Keller (2016) found that the processing of pictographs is disparate between subjects with different levels of numeracy. While highly numerate individuals tend to count each icon and thus pursue a numerical approach, the less numerate extract information by processing large areas of the icon-arrays (gist-based). The two studies found that pictographs are beneficial for both groups, which manage to attain the same risk perceptions, but Hess et al. (2011) concluded that guiding the low numerate into counting the icons might be detrimental to them. Other differences in perception, processing of information and decision-making between different levels of numeracy will be explored in Section 2.4.

As already mentioned in Section 1.1, large portions of American and German adults have surprisingly low levels of numeracy and even medical staff or highly educated individuals can exhibit the lack of these same skills (Anderson & Schulkin, 2014; Galesic & Garcia-Retamero, 2010; Lipkus et al., 2001; Peters, Meilleur, & Tompkins, 2014). Numeracy has also been reported

to be lower in older populations, which is consequential due to the larger number of health decisions that older adults face when compared to younger people (Trevena et al., 2013; van Weert, Alblas, van Dijk, & Jansen, 2021).

Numeracy can also have an impact on personal health outcomes and general quality of life: according to Garcia-Retamero, Andrade, Sharit, and Ruiz (2015), patients with less numeracy skills have a 40% higher relative risk (compared to highly numerate individuals) to develop one of the following diseases: myocardial infarction, chronic obstructive pulmonary disease, peptic ulcer disease, liver disease, diabetes, HIV/AIDS, among others. They also display a greater risk of comorbidity, lengthier delays in soliciting medical care (Petrova et al., 2017), and overestimation of both their own risks of developing various illnesses (Davids, Schapira, McAuliffe, & Nattinger, 2004) as the effectiveness of uncertain treatments (Weinfurt et al., 2003). Less numerate individuals have a greater use of emergency services (Galesic & Garcia-Retamero, 2011a), generally avoid participating in the decision-making process (Garcia-Retamero et al., 2014), and seem to interpret the risk of side effects with more difficulty (Gardner et al., 2011); and according to Berkman, Sheridan, Donahue, Halpern, and Crotty (2011), elder individuals with insufficient health literacy - which includes print literacy, numeracy and oral literacy - tend to display substandard overall health status and higher mortality.

In conclusion, the lack of numeracy has a dramatic impact across several degrees of the relationship between patients, risks, and personal health, therefore putting a large portion of society at greater risks than individuals with developed numerical skills. In broader terms, a recent analysis by Baccolini et al. (2021) found that between 33% and 50% of Europeans have insufficient levels of health literacy. While long-term solutions may benefit from investments in education (Peters et al., 2014), short-term solutions are required to address the specific needs of the less numerate.

2.3 Communication formats

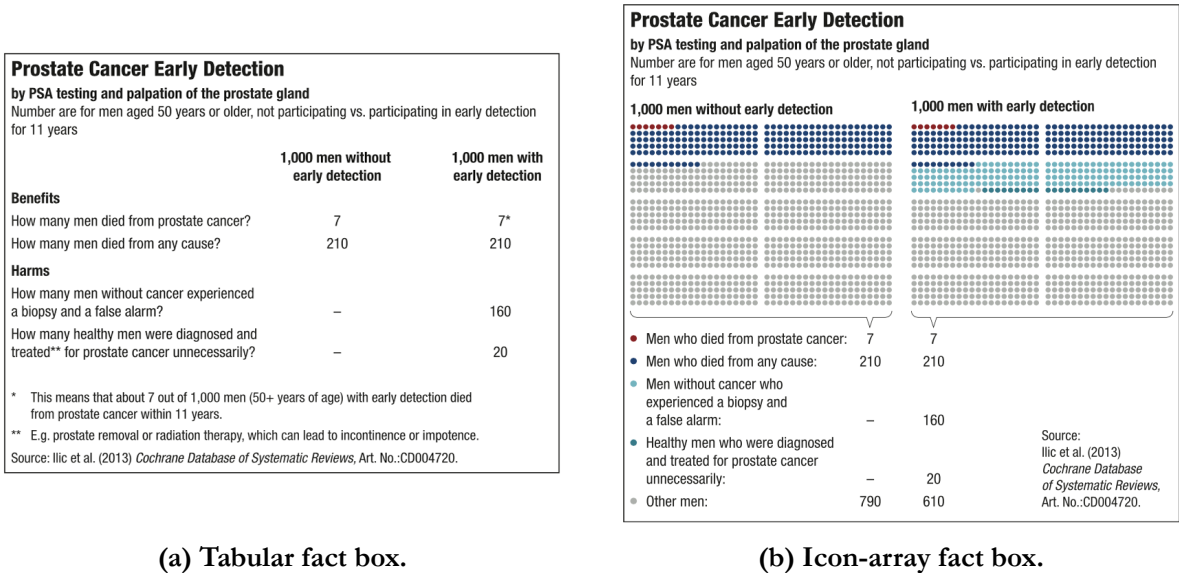
As stated in Chapter 1, numerous different formats exist when it comes to the communication of risk information in the context of health, and their impact in decision making can be so diverse that an appropriate outline should not be overlooked. Thorough analysis of each format would require a systematic search and review, which is not the scope of this Thesis report. In turn, this section will pose as an overview of different formats, along with general findings and the guidelines that exists towards their use or avoidance.

Verbal communication of numerical probabilities can be done by the use of terms such as *likely*, *certain*, *very rare*, and so on. This type of terms allows for higher fluidity in communication, to convey the extent and origin of uncertainty, and may be a better fit than other formats for the expression of personal feelings and emotions (Lipkus, 2007). Nonetheless, they are often associated with imprecision on risk estimates and high variability of interpretations (Spiegelhalter, 2017; Vromans, Pauws, Bol, van de Poll-Franse, & Krahmer, 2020) (see also Appendix C). Studies by Budescu, Por, and Broomell (2011) and Budescu, Por, Broomell, and Smithson (2014) are highly cited to illustrate this, with the former finding that while the International Panel on Climate Change attributes a probability of 90% to the term *very likely*, the mean estimate from study

subjects was 62%. Finally, while some studies report that people prefer a verbal communication of risks over numerical formats (Trevena, Davey, Barratt, Butow, & Caldwell, 2006), the opposite is also true (Gurmankin, Baron, & Armstrong, 2004; Wallsten, Budescu, Zwick, & Kemp, 1993). Finally, Freeman (2019) suggests that words should be always accompanied by numbers.

Numerical formats are in fact more precise and may be better for verbatim knowledge than graphical formats (Trevena et al., 2013), they can be computed, processed and verified, tend to convey higher trust (Visschers et al., 2009) and less risk-aversion (Zipkin et al., 2014). However, numeric formats can be especially difficult to interpret and understand by people with low numeracy levels (Schwartz et al., 1997; Woloshin, Schwartz, Moncur, Gabriel, & Tosteson, 2001). Numbers are often used as percentages or frequencies (e.g. a 2% or 2 in 100 chance an event occurring, respectively), but can also take the form of odds, classical probabilities or numbers needed to treat/harm. Among these, natural frequencies are often the preferred format as they convey both the affected individuals as the reference class (Lipkus, 2007), as it will be mentioned in Section 2.2 below. Gaissmaier et al. (2012) point that for individuals with low graph literacy, numbers may induce higher gist and verbatim knowledge than graphical formats.

A format that has recently received growing attention is the fact box (Freeman, 2019), a summary format in the shape of a table that in some cases may include icon-arrays, as can be seen in Figure 2.1 below.



(a) Tabular fact box.

(b) Icon-array fact box.

Figure 2.1: Tabular and icon-array fact boxes included in a study by McDowell et al. (2019) [corrected].

Tabular fact boxes are often linked to better comprehension and short-term recall than text formats, as they might be more engaging and easier to extract information from, even for less numerate individuals (Brick et al., 2020). McDowell et al. (2019) pointed to the same conclusions for both tabular and icon fact boxes, finding no significant difference between the two. This type of communication structure, in truth a mixture of verbal, numerical and sometimes graphical

formats, is successful at focusing on the information that is vital for a decision and allowing the user to find it with ease (Fischhoff, 2017; Schwartz & Woloshin, 2013).

Visual decision aids are generally regarded as contributing to an improvement of risk understanding, with subsequent positive impacts on comprehension, behavior, trust, risk avoidance, and recall, among others (Garcia-Retamero & Cokely, 2017). Notwithstanding, the variety of visual aids is vast - pictographs, bar graphs, line graphs, risk ladders, funnel plots, the infamous pie chart¹, and so on (examples in Figure 2.2 below) -, and the appropriate choice of style may depend on the communication goals (Ancker, Senathirajah, Kukafka, & Starren, 2006; Hawley et al., 2008).

In favour of the use of graphics is their attractiveness and ability to summarize information by putting on display otherwise hidden patterns in data; conversely a graphical format can also be regarded as distracting, misleading, unfamiliar or too complex (Lipkus, 2007). In fact, as stated above in Section 1.1 as a central topic for this Thesis, individuals with insufficient graph literacy do not benefit as much from visual aids (Garcia-Retamero & Cokely, 2017; Trevena et al., 2013). Nevertheless, it is thought that graphical formats actively contribute towards the reduction of framing effects, denominator neglect or the impact of anecdotes (Fagerlin, Wang, & Ubel, 2005; Garcia-Retamero & Galesic, 2010a; Garcia-Retamero, Galesic, & Gigerenzer, 2010). Pictographs are widely agreed upon as the best all-rounder solution when communicating individual risks, mainly due to the expression of risk as a frequency (instead of a probability), and the display of both the numerator and denominator (Fagerlin et al., 2011).

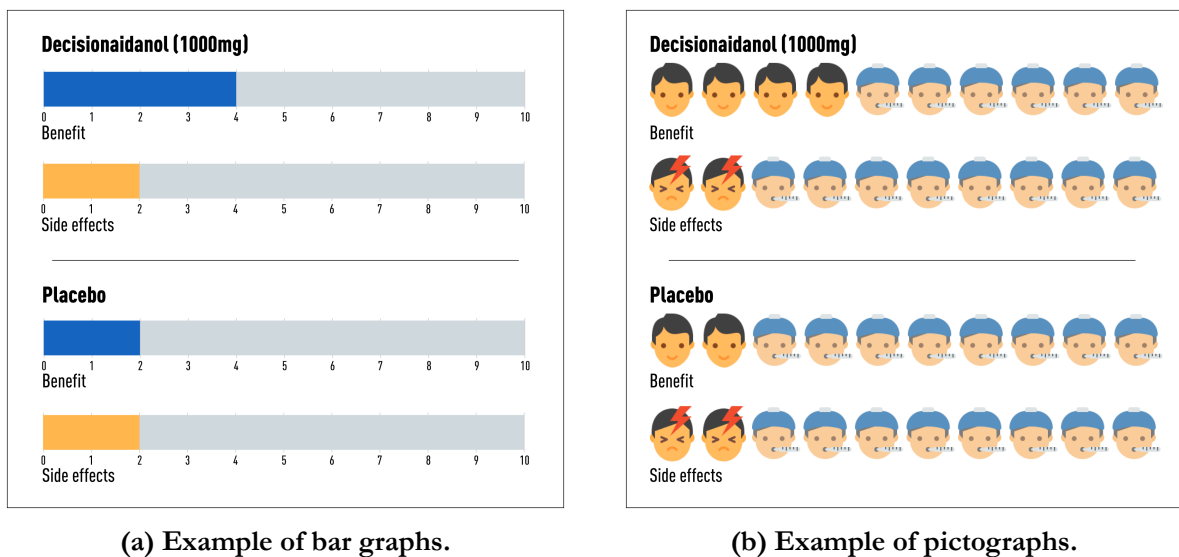


Figure 2.2: Examples of Visual Aids that can be used to inform and support a patient's decision.

Decision aids using animations have mixed results on performance (Garcia-Retamero & Cokely, 2017), with only some studies reporting improvements (with rather surprising results

¹Data Visualization pioneer Edward Tufte once said that “[g]iven their low density and failure to order numbers along a visual dimension, pie charts should never be used.” (Tufte, 1986, p. 178). It is also a format that tends to be less trusted by patients (Hawley et al., 2008).

in individuals with low graph literacy) (Okan et al., 2015). As they encourage users to engage, interactive formats are thought to improve understanding and recall, being also able to tailor information and adapt to the user (Spiegelhalter, Pearson, & Short, 2011). Even so, both animations and interactive formats may distract users from the information at stake and worsen understanding (Garcia-Retamero & Cokely, 2017).

Lastly, analogies are also a widely used format in health-related contexts and are thought to be undemanding in terms of numeracy and graph literacy (Garcia-Retamero & Cokely, 2017). Nonetheless, its effects depend on cultural context and the complexity of the information to convey: a study by Galesic and Garcia-Retamero (2013) concluded that for complex medical situations individuals with high numeracy skills benefited the most from the use of analogies, while for easy medical problems the same was true for the less numerate. One common example for an analogy is that of comparing the relationship between cancer screening and cancer (target) to a car alarm and theft (base). In fact, car alarms also experience false positives (activating the alarm when they shouldn't) and false negatives (not activating in the event of theft) (Galesic & Garcia-Retamero, 2013). Visual analogies have also been tested with positive results (Turchioe et al., 2019).

2.4 Perception of Risk

The comprehension and use of risk information does not depend exclusively on the format that is used but also on the type of information one has to convey, which of its parts are emphasized or ignored and what components are present, often in a close interaction (as with format choice) with aspects of perception, decision-making processes and modes of thinking (Garcia-Retamero & Cokely, 2017; Slovic et al., 2005). We can thus consider that in the presence of a decision, a patient's choice not only depends on what information is conveyed, but also on how this information is presented, something that is highly related with the choice architecture and nudge theory proposed by Thaler and Sunstein (2008).

According to the *dual-process theories of information processing* (Chaiken & Trope, 1999; Sloman, 1996), risk is perceived, processed and used in two distinct ways: one uses what is called the affect heuristic - when emotions and intuitions integrate the decision making process -, which tends to perceive risk as feelings; while contrary to this mode of thinking is the rational, more deliberative and logical way of processing reality and information (Epstein, 1994). Both modes are regarded as essential and enclose rationality to some extent, and despite early beliefs that the analytical was the paragon of experiential thinking, the rational is currently thought to be highly dependent on the affective (Damasio, 1994; Slovic et al., 2005). The use (or the influence) of one system over the other can vary according to individuals: while a doctor may have a larger use of the logical system to interpret risk information, a patient's mode of thinking may be more influenced by emotions (Reventlow, Hvas, & Tulinus, 2001). This was lightly addressed in Section 2.2 and will be mentioned further below.

The perception of specific features and formats of risk communication has been subject to extensive research from where several conclusions and guidelines have sprouted. For instance, the use of frequencies is usually recommended alongside percentages, unless when comparing two or

more numbers, where a percentage-only format is thought to be easier (Freeman, 2019; Trevena et al., 2013). If put into practice, frequencies should always specify a denominator (i.e. reference class), which must be kept consistent throughout the Decision Aid, as varying denominators are thought to worsen risk understanding and heighten risk perceptions (Trevena et al., 2013).

Some studies point to the hypothesis that individuals pay more attention to the numerator (Freeman, 2019), an effect commonly called denominator neglect, which is believed to be larger among the less numerate (Peters et al., 2006; Trevena et al., 2013). A study by Slovic, Monahan, and MacGregor (2000) illustrated this effect by presenting forensic psychologists and psychiatrists with the decision of approving a patient's discharge after knowing that 1) "20 out of every 100" similar patients were likely to be violent or 2) the patient had a 20% chance of being violent. The results revealed that, regardless of sharing the same probability of violent behavior, discharges were less likely in the frequency group, as participants appeared to be more affected by this framing and the vision of 20 violent individuals (see also Peters et al. (2006)). Despite these findings not all literature is consistent on whether individuals focus more on the denominator or the numerator (Lipkus, 2007).

Denominator neglect can in turn lead to ratio-bias, a situation where a lower probability ratio is perceived as higher (and vice versa) due to its absolute frequency (Lipkus, 2007). Denes-Raj and Epstein (1994) demonstrated this by offering a prize to participants in a study if they drew a red jelly bean from one of two bowls. The first bowl contained 10 beans, one of which was red (thus a 10% chance of winning); while the second had a total of 100 beans and, depending on the trial, 5 to 9 red beans (thus a 5% to 9% chance of winning); the respective probabilities were clearly displayed in each bowl. When facing the choice between 1 in 10 and 9 in 100, 61% of the participants preferred the latter option, and 21% went as far as choosing the large bowl when it had only half the chances of winning (5 in 100) than the small bowl (1 in 10). The results were analogous but less extreme when the trials meant a loss instead of a win. According to Denes-Raj and Epstein (1994, p. 819) "subjects reported that although they knew the probabilities were against them, they felt they had a better chance when there were more red beans", which is very much in line with the affect heuristic mentioned above. In a more recent study by Peters et al. (2006), numeracy was shown to have a significant effect on ratio-bias and sub-optimal choices: in a 1 in 10 vs. 9 in 100 scenario, less numerate individuals were more prone to choose the worse bowl (33%) than those with high numeracy levels (5%).

Framing is yet another effect that is particularly mentioned across a wide range of literature and can be easily illustrated by an apparently simple question: when presenting different treatments to a patient, should we use survival or death rates? As said by Spiegelhalter (2019, p. 32) "'5% mortality' sounds worse than '95% survival'", negative or positive framing are thought to have different effects on perception, risk understanding and decision making. While the latter is thought to induce risk-averse decisions and the former to produce risk-seeking behaviors, the effects are largely context dependent (Trevena et al., 2013). A study by McNeil, Pauker, Sox, and Tversky (1982) tested the choice between two cancer treatments among a sample of clinicians students and patients, by showing them the survival or mortality rates after 1 and 5 years. The type of framing had a significant effect in choice and influenced the treatment selection, making positively-framed treatments more attractive.

To further illustrate the effects of framing, Garcia-Retamero and Cokely (2011, 2014, 2015)

executed a set of studies on sexually active subjects with low numeracy that corroborated this effect. Participants were either shown a gain-framed message (regarding the benefits associated with a certain behavior) or a loss-framed message (drawbacks associated with not adopting the behavior), with the results showing that the gain-focused message produced a greater adherence to preventive behaviors (e.g. using a condom), while the loss-framed message was shown to generate higher rates of STD (sexually transmitted diseases) screening. These outcomes are especially relevant because not only do they attest for the distinct effect of positive vs. negative framing, as they prove that the choice between one or the other may depend on the purpose and goals of the communication: while sometimes the purpose may be full transparency, other situations may have the objective of nudging² the individual towards a certain decision or behavior. While lower numeracy is associated with higher susceptibility to framing, visual aids have been shown to reduce its effects (Trevena et al., 2013). Nonetheless, while some literature recommends the simultaneous use of both types of framing, there is currently no consistent solution (Freeman, 2019).

Given its considerable contribution to heightened and off point risk perceptions, the use of relative risks is arguably one of the aspects of Risk Communication in which consensus is larger among researchers. A paramount example is that of the UK's 1995 pill scare, when a newer (third) version of a contraceptive pill was linked to twice the risk of venous thromboembolism than its predecessor. This twofold risk was promptly communicated by the media, which persuaded thousands of women to stop taking the pill altogether. The following year 12400 additional births and 13600 additional abortions were recorded. The information was in fact true - the third version of the pill had twice the risk of this serious side effect than the previous one - but in absolute risks this meant an increase from 1 in 7000 to 2 in 7000 women, while the risk of this same event was about 4 in 7000 for pregnant women (Furedi, 1999). Thus, relative risks have been proved to magnify risk perceptions, especially when absolute risks are small, and should be avoided in almost all situations, particularly alone (Fagerlin et al., 2011; Freeman, 2019; Trevena et al., 2013). If relative risks are used, the baseline must be specified (Lipkus, 2007), and if absolute risks are very low some attention may be needed, as individuals may deem the risk as insignificant (Freeman, 2019).

2.5 Other aspects

Several other features may influence the choice architecture of a given risk communication structure or result from it, and despite covering all of them being outside the scope of this Report, some examples can be addressed. The use of informative labels is one of them, as labels are not a format *per se*, but rather an addition that can be used on top of the chosen format in order to give context to the enclosed information. With the help of labels, a patient is able to interpret quantitative information and understand the magnitude of his own personal risk (e.g. "Your 8% risk of developing side effects is lower than the average for people similar to you"), being therefore less at risk of influence from affective reasoning and non-numerical information

²Thaler and Sunstein (2008, p. 6) describe a nudge as "any aspect of the choice architecture that alters people's behavior in a predictable way without forbidding any options or significantly changing their economic incentives."

(especially for the less numerate) (Okan et al., 2015; Trevena et al., 2013). However, evaluative and comparative labels should be used with care, as they can deeply influence risk perceptions and behavior intentions (Fagerlin et al., 2011).

Reflective questions and tailoring³ are other “expansions” on risk communication with positive impacts on the use of provided information, as they may lead to higher engagement and comprehension (Brewer et al., 2007; Noar et al., 2007; Okan et al., 2015).

As mentioned above in Section 2.3, apart from objective measures such as comprehension or recall some studies also explore subjective measures as format preference or trust. Interestingly, Trevena et al. (2013) found that individual preference for a specific format is not always correlated with the best objective performance, and Lipkus (2007) concluded that accuracy of comprehension may not conduct to behavior change, casting some light onto not so obvious interactions between objective and subjective metrics.

2.6 General discussion

As can be inferred throughout the present Chapter, Risk Communication research has multiple goals, focuses and even outcomes, with several questions still open or with inconclusive results. Nonetheless, there are some main findings that share agreement from the field and provide common ground for current research.

We’ve covered the main existing tests for numeracy and graph literacy, and presented an overview of numeracy levels across European and American populations, as well as the several nefarious connections that have been found between low numeracy and negative health indicators, behaviors and objective comprehension of risk information. Lower levels of numeracy are also related with a higher propensity of being influenced by framing, ratio-bias, affective reasoning and other non-numerical information. Risk perception is concluded to be different among individuals with different numeracy or graph literacy skills, while the same can be said on the effects of risk communication formats. Visual aids are thought to induce dramatic improvements in both objective and subjective measures when compared to other formats, but their effects are rather shallow in individuals with both low numeracy and low graph literacy.

Current study limitations are several and varied. In regard to sampling, the wide majority of experimental studies uses moderately-educated samples, lacking vulnerable groups such as less educated individuals, rural populations and older people (van Weert et al., 2021). There is a need for more studies on decisions of real patients in real medical contexts (Richter, Girolodi, Jansen, & van der Weijden, 2020), and also studies outside the scope of health risks (such as environmental, public-policies or economical risks) (Spiegelhalter, 2017). Finally, only a minority of studies tests for graph literacy (Garcia-Retamero & Cokely, 2017), which is vital if researchers wish to address the subgroup of people with low numeracy and low graph literacy.

This specific subgroup - a central topic of this Thesis - is among future research directions for the Risk Communication field, as well as the exploration of risk perceptions in diverse cultural

³Tailored communication is described by Kreuter, Farrell, Olevitch, and Brennan (2000), cited by Noar, Benac, and Harris (2007, p. 674), as “any combination of strategies and information intended to reach one specific person, based on characteristics that are unique to that person, related to the outcome of interest, and derived from an individual assessment”.

contexts (Trevena et al., 2013). Other suggested future research includes: tailoring of graphical formats (Vromans et al., 2020), use of visual analogies (Turchioe et al., 2019), screening for color-blindness or the use of data science (Turchioe et al., 2020).

2.7 Conclusions

The Literature Review that builds this Chapter has allowed for an overview of the Risk Communication field, its main findings, limitations and related theories. We conclude that research is rich and evolving, with growing interest in the field and new directions sprouting in recent years. This Review also enabled the disclosure of the most adopted and promising communication formats, common biases and main challenges, an essential domain knowledge for an adequate Protocol creation and Meta-analysis execution. Lastly, it also confirmed the relevance of this Thesis main topic and question, serving as a catalyst for the work and steps to come.

Chapter 3

Meta-analysis Plan and Rationale

Knowing the sensitivity associated with the multiple steps of a Meta-analysis process, several sources were consulted in order to assemble and follow a cohesive course of action and thus mitigate any chances of bias or duplication. Among these resources, the JBI Manual for Evidence Synthesis (Aromataris & Munn, 2020) proved to be of the utmost importance by casting light onto the necessary procedures of a high-quality Meta-analysis.

Accordingly, the first step was the creation of a Protocol, which ensures that the goals and methods of the Meta-analysis are delineated *a priori*, hence contributing towards an unbiased and transparent research. This Protocol poses as a blueprint for the Meta-analysis as a whole, outlining several decisions and procedures such as the Review Questions, Inclusion Criteria, Search Strategy, Study Selection and Data Extraction, among others.

As this Section will describe the planning that preceded the Meta-analysis, going through all its necessary steps, its contents are closely related to those of the Protocol, which was made available in an OSF repository (<https://osf.io/4apx8/>) and timestamped prior to the execution of the Search strategy¹. Only in the subsequent Chapter 4, where this plan of action is put into practice, are we going to analyze and comment how it unfolded.

3.1 Meta-analysis

As mentioned in Section 1.1, the initial goal of this Thesis is to conduct a Meta-analysis on the effects of different Decision Aids in individuals that exhibit low levels of both numeracy and graph literacy. The motivation behind this particular topic comes from the lack of conclusive or well-supported solutions for this specific group as well as the nonexistence (to the best of our knowledge) of meta-analyses oriented towards this problem.

3.1.1 Steps

After certifying the feasibility and relevance of the proposed Meta-analysis, a Search strategy should be outlined, supporting a subsequent literature search through databases, references of

¹The first version of the Protocol was uploaded on March 23rd, followed by an update on April 16th.

collected articles (using both snowball and reverse-snowball searches) and, if applicable, grey literature. Studies are then filtered according to selection criteria (first by analysing the Abstract and Keywords and then executing a full-text reading), appraised and filtered. The PRISMA flow diagram (Page et al., 2021), included in Appendix E, is used to depict the information flow through these steps.

Relevant data is then extracted from the selected studies (used methods are described further below), thus building the matrix that will be inputted into subsequent analyses, executed using programming language R, mainly due to its availability and ease of result replication, using the *metafor* (Viechtbauer, 2010), *robumeta* (Fisher, Tipton, & Chipeng, 2017) and *dmetar* (Harrer, Cuijpers, Furukawa, & Ebert, 2019) packages. The software and packages will be covered in greater detail in Section 3.3. In respect to statistical models, and according to the JBI guidelines (Aromataris & Munn, 2020), we can state that a Random Effects model will be used, due to its generalization properties (Tufanaru, Munn, Stephenson, & Aromataris, 2015). Proper rationale for this choice will be provided further below. Several analyses are to be executed - common effect size, moderator analysis, sensitivity analysis, publication bias, heterogeneity between studies - and plots produced - Baugat plots and Forest plots, among others. If sufficient studies are collected, other meta-analyses on objective outcomes can be executed (effect on behavior intention, decision making, etc.). Despite having the goal of addressing the particular subpopulation of individuals with low levels of numeracy and graph literacy, data on the other three subgroups (low-high, high-low and high-high) may also be collected and analysed for richer comparisons. A significance level of $\alpha = .05$ will be considered for all analyses.

Results will then be examined and commented, hopefully reaching a conclusion on which Decision Aids have the largest positive effect when presenting risk information to individuals with low levels of both numeracy and graph literacy.

3.1.2 Rationale

Before delving further into the mechanics and intricacies of executing a meta-analysis we should first attest for its rationale and application. The goal of a meta-analysis is, similarly to a narrative review, to search for existing literature on a given topic and summarize its conclusions. As a narrative review, a meta-analysis benefits from this bird's eye view, probing a research field as a whole, being able to identify trends, patterns, consensus or even disparate findings, and with it draw diagnoses that could ultimately lead to clearer conclusions or even shape future works. It should systematize its procedures, and share as much of the decision-making processes as possible prior to its execution, thus contributing to its transparency and reliability.

Despite the similarities mentioned above, narrative reviews and meta-analyses differ in several aspects. Primarily, when executing narrative reviews, researchers are often looking at p-values, which only allow for a test of the null-hypothesis and completely neglect the effect size. This kind of approach lets us identify if an effect is found, but does not enable us to evaluate its magnitude, which is essential “[i]f a clinician or patient needs to make a decision about whether or not to employ a treatment, [as] they want to know if the treatment reduces the risk of death by 5% or 10% or 20%” (Borenstein, Hedges, Higgins, & Rothstein, 2009, p. 12). Also, using p-values to evaluate and compare the presence of statistical significance between treatments may

be misleading. Abelson (1997) gives an example of a treatment with a p -value = 0.05 in sample A and < 0.07 in sample B, and states how far-fetched it would be if we decide to address the treatment as significant in sample A but not in B and ignore the minute difference between the two². On the other hand, a meta-analysis focuses specifically on effect size, using statistical methods to synthesise available data into a common effect size and confidence interval, shared by the combination of selected studies. Moreover, in a meta-analysis we can analyse the dispersion between collected studies and execute subgroup analysis to examine the presence of any moderator variables affecting the effect size.

Additionally, narrative reviews are prone to influences from the researchers themselves, their perspectives and the studies they decide to focus on or ignore. Systematic narrative reviews try to address these issues but still have to deal with its qualitative essence and the risk of subjectiveness of its interpretations. As Borenstein et al. (2009, p. 14) mentioned - borrowing a phrase from Abelson (1997, p. 119) - “it involves doing arithmetic with words”. Contrary to this, a meta-analysis that is planned and executed transparently is shielded from these biases due to its quantitative and objective nature.

The aforementioned characteristics of a meta-analysis support our decision to conduct this type of systematic review, especially the ability to summarize collected research into a common effect size, not only detecting the direction of the effect, as its magnitude and dispersion.

3.1.3 Procedure

A random-effects model will be used due to several factors. Most importantly, a distribution of true effect sizes is expected, as collected studies will certainly exhibit variations in types of population, age, location, interventions, subgroup cut methods, to name a few. This allows us to assume the nonexistence of a common true effect size between studies, which voids the application of a fixed-effects model. Secondly, we wish to generalize and extrapolate the results and not just describe the set of collected studies. Lastly, if heterogeneity between studies is absent, the resulting outputs from a random-effects model are identical to those given by a fixed-effects model (Borenstein, Hedges, Higgins, & Rothstein, 2010). This occurs due to the nullification of the constant T^2 , which is used to calculate study weights in the random-effects model and poses as the key difference between both models. The closer this constant gets to 0, the more similar are the weights attributed to each study in both models and the closer are the final results. Weighting functions for fixed-effects (4.1) and random-effects (4.2) can be seen below:

$$W_i = \frac{1}{V_i} \quad (3.1)$$

$$W_i = \frac{1}{V_i + T^2} \quad (3.2)$$

²Cumming (2013) makes a point against this sort of black and white vision, advocating for a complete boycott of null hypothesis statistical testing.

where, for study i , W_i is the resulting weight, V_i the within-study error variance, and T^2 the global between-studies variance.

In regards to heterogeneity, a Cochran Q test will be executed in order to test for the presence of significant dispersion³. As the statistical power of Q weakens considerably with a small number of included studies, results will be interpreted with caution if less than 20 studies are included (Aromataris & Munn, 2020), acting as a reference guide - but not a guarantor - for the presence of dispersion.

The value of Q will then be inputted into the estimation of tau-squared T^2 , which represents the total between-study variance in the same scale as the effect size. In turn, this allows for the calculation of the estimated standard deviation of the effect size T , which determines the effect size distribution (or prediction interval) and respective analysis of dispersion. Additionally, and as proposed by J. P. Higgins, Thompson, Deeks, and Altman (2003), an I squared statistic I^2 - which measures the ratio of true heterogeneity to total variation in observed effects - will be calculated, with values of 25%, 50%, and 75% corresponding to low, moderate and high ratios, respectively.

The dispersion statistics mentioned above will allow for an assessment and evaluation on the consistency of results, as these are key for a justified focus on the summary effect size of the meta-analysis. In the case that effect sizes are not consistent, and depending on the magnitude of dispersion, the meta-analysis report may pivot towards variance, its causes and meanings.

If heterogeneity is found present, a Baujat plot will be produced in order to identify which studies are contributing the most to the overall heterogeneity of the meta-analysis and detect the presence of any outliers. If detected, sensitivity analyses will be conducted to report on the effects of excluding the identified outliers.

In order to explain the presence of variance, a minimum threshold of four effects should be met to allow the execution of any subgroup or meta-regression analysis (Fu et al., 2011). These analyses will be conducted to address research questions 3 to 6 - mentioned in following Section 3.1.4 - in order to detect any notable differences within the aforementioned variables on the outcomes of visual Decision Aids.

To assess study quality and publication bias a funnel plot⁴ of included studies will be produced followed by a visual analysis of symmetry and an Egger's regression test (Egger, Smith, Schneider, & Minder, 1997). If in the presence of an asymmetric funnel plot, a trim-and-fill method may be employed to produce a symmetric funnel plot and allow for the estimation of an adjusted effect.

Finally, the meta-analysis report will follow the PRISMA 2020 statement and guidelines.

3.1.4 Research Questions

As particularly highlighted in Chapter 2, there is currently no consensus on which type of Decision Aid should be provided to individuals that lack basic levels of both numeracy and graph literacy. In addition, the common effect size of different formats on comprehension or accuracy of interpretation within this subgroup is, to the best of our knowledge, yet to be studied. Ac-

³As suggested by Sutton, Abrams, Jones, Sheldon, and Song (2000), a statistical significance level of 0.10 will be used for the Cochran Q test.

⁴To test if asymmetry is due to publication bias a contour-enhanced funnel plot will be used, highlighting areas of .1, 0.05 and 0.01 statistical significance.

cordingly, executing a Meta-analysis stands out as an obvious course of action, envisioning the identification of the most effective risk communication formats and possibly its interactions with individual, study, and health behavior characteristics.

Following the conclusions presented in the narrative systematic review by Garcia-Retamero and Cokely (2017), the proposed research questions are related to the effects of visual decision aids on the subpopulation under study:

- **RQ1:** What is the direction and magnitude of the effect of Visual Decision Aids on risk comprehension in less numerate and graph-literate individuals?
- **RQ2:** Is there a general consensus between selected studies or not (low or high dispersion of reported effect sizes, respectively)?

If sufficient eligible studies are collected, a moderator analysis will be executed in order to find any relevant differences in the effect between graphical formats:

- **RQ3:** Is the effect size on comprehension of risk information by less numerate and graph-literate individuals moderated by the type of visual aid (icon array, bar graph, pie chart, etc.)?

Other analyses will be executed in order to verify for moderating variables among participant and study characteristics, as well as distinct health behaviours:

- **RQ4:** Are participant features (age, gender, education level) moderating the effect size?⁵
- **RQ5:** Do study characteristics (country of sample, numeracy/graph literacy test) act as effect size moderators?
- **RQ6:** What is the effect size variation by type of health behaviour (e.g. screening or preventive)?

3.2 Search and Screening Strategy

First and foremost, eligibility criteria were defined so as to guide the search and selection process, describing the type of articles that should be sought and ultimately included in the final analyses. We decided that eligible articles should be written in English language and feature a visual aid built for health-based contexts, comparing it with textual or numerical formats. Patients should be the target audience, not doctors or professionals. Selected studies must explore the interaction between numeracy and graph literacy and measure objective outcomes such as comprehension or decision making, among others. The type of study will pose rather as a preference than a criterion. Randomized Controlled Trials will be preferred, but if found in insufficient numbers, other designs will be considered (e.g. quasi-experimental or observational studies). Inclusion and Exclusion criteria are shown in Table below.

⁵Feasible only if selected studies display these descriptive statistics for the specific subgroup in study.

Table 3.1: Inclusion and Exclusion Criteria.

Inclusion	Exclusion
Article written in English	Written in non-English language without translation
Patients are part of target audience of the visual aid	Patients are not included in target audience
Visual Aid displays health-related data	Visual Aid does not explore health data
Explores interaction between numeracy and graph literacy	Does not explore interaction between numeracy and graph literacy
Uses textual or numerical formats as comparison	Uses other types of comparators (or none)
Explores objective outcomes (e.g. accuracy of risk understanding, comprehension, decision making and behavior intentions)	Does not explore objective outcomes
Includes effect size or data for less-numerate and graph-literate individuals	Does not include effect size for subpopulation in study

The search strategy was planned to be executed within Web of Science, Pubmed, Scopus, ACM Digital Library, IEEE Computational Index, and Google Scholar. Efforts will be made in order to find relevant studies in grey literature sources (e.g. Open Grey, Grey Literature Report).

Initial Logic Grids were constructed following the PICO mnemonic (Population, Intervention, Comparison and Outcome) and guidelines in Aromataris and Riitano (2014). After building and complexifying these Logic Grids, search strings were tweaked according to other relevant criteria and database-specific syntax. As the most used graph literacy test was published in 2011 by Galesic and Garcia-Retamero (2011b) and no other tests were found before this year, and knowing that eligible studies must explore the interaction between numeracy and graph literacy skills, the search strategy will only consider studies from 2010 onwards.

The tables that were built with initial search terms, resulting search strings and filters can be examined in Appendices G and H and within the online supplementary material (search strings: <https://osf.io/ucvz7/> and logic grids: <https://osf.io/sdtr6/>).

Flow of filtered studies per stage should be recorded and reported with the PRISMA 2020 flow diagram. The selection of articles should be executed according to the criteria presented above, initially through Title and Abstract reading, followed by full-text reading. Once this selection process is finished, further studies may be included by analysing citations of selected studies (using both backward and forward searching). Articles citing studies that proposed graphical literacy tests - such as Galesic and Garcia-Retamero (2011b) and Garcia-Retamero et al. (2016) - will also be analysed for eligibility. This will be executed by using the *Cited by* option in Google Scholar, and selected studies will be directly included in the “Identification of new studies via other methods” part of the flow diagram and identified as “Cited key study”. In the event that other key studies are analysed, these will be mentioned in the meta-analysis final document. Search and screening steps will be executed by Francisco M. Gomes (FG) and Vasco Brazão (VB) as dis-

played in Table 3.2 and further explained below.

Table 3.2: Search and Screening methods.

Step	Researcher
Search execution	FG
Title and Abstract extraction	FG
Upload to Rayyan	FG
Title and Abstract screening	FG and VB
Study extraction	FG
Full-text screening	1st FG and VB (10 articles or 25%*) 2nd FG (remaining articles)
Selected studies citation search	FG
Key studies citation search	FG
Coding	1st FG and VB (5 articles or 10%*) 2nd FG (remaining articles)

* *Whichever value is achieved first.*

FG will execute the search strategy in each of the mentioned databases, extracting Title, Abstract and other relevant information for the collected studies. Each file will be saved, uploaded first to OSF - thus being timestamped - and afterwards to Rayyan (Ouzzani, Hammady, Fedorowicz, & Elmagarmid, 2016), which will be used for Title and Abstract screening. FG and VB will execute this screening step independently, without being able to view each other's decisions until all records are screened. At this point both researcher's labels will be revealed, any conflicts recorded, discussed and resolved. Selected studies will be sought for retrieval by FG. In the first phase of the full-text screening FG and VB will go through 10 articles or 25% (whichever value is achieved first) independently, any existing conflicts will be addressed and resolved; in the second phase FG will screen the remaining articles. Selected and key studies citation searching will be performed by FG. Finally, the first phase of coding will see FG and VB code 5 articles or 10% (whichever value is achieved first), conflicts will be discussed and resolved, and FG will code the remaining articles.

3.3 Data and Software

A data extraction form, available among the supplementary material (<https://osf.io/u28ac/>), was created to synthesize the relevant data to extract from selected studies. The initial data spreadsheet will feature one article per line, including information on the study (e.g. Study ID, Authors, Title, Year of publication) and overall sample characteristics (e.g. Sample size, Average Age, % Female).

Effect sizes will be merged into a final spreadsheet with one comparison per line - one study may explore several relevant comparisons and thus feature in multiple lines -, including information on the study, tested outcome, comparison, format type, and so on. Data of interest may be

partially (or not at all) reported by one or more studies, event which will prompt the main author of this meta-analysis to seek any original, raw or missing data that could eventually enrich the research; nonetheless, we predict to use tools such as WebPlotDigitizer (Rohatgi, 2020) to extract a significant amount of data. In cases in which data is not available, extraction is not possible and attempts to contact original authors fail, the related study will be excluded from analysis. Table 3.3 below features a list of variables that were planned to be extracted.

Table 3.3: Data to extract per selected study/experiment.

Study Identification	Sample Characteristics⁶
Study ID	Sample size n
Authors	Type [student, adult, general]
Title	Average Age
Year of Publication	% Female
Doi	% with High School or Less
Source [published/unpublished]	Nationality
Numeracy Test	Numeracy Mean
Graph literacy Test	Graph literacy Mean
Method for splitting low/high groups	
Real medical decision [1,0]	
Independent / Repeated Measures	
Study Design	
Results/Effect Sizes (accordingly)	Moderators
% of correct answers (or similar)	Health Behavior ⁷
SD	Behavior Type ⁸
Subgroup sample size n	
Format ID	
Outcome ID	
Comparison ID	
Effect Size ID (if any)	
Effect Size Value (if any)	

We expect the majority of studies to report the means of correct answers (or similar) and respective standard error for each of the tested formats. If this expectation holds, and even in the event that some studies use different scales to report these values, a standardized mean difference Cohen's d will be estimated⁹ and then converted into Hedges' g (which corrects the overestimation bias that may be present in d).

⁶If available, participant characteristics will be extracted for the whole sample as well as for the particular sub-sample with low levels of numeracy and graph literacy.

⁷e.g. Diet, Mammography, Smoking cessation.

⁸e.g. Preventive, Screening, Vaccination.

⁹Cohen's d has different formulas for (1) independent groups and (2) pre-post scores or matched group samples.

However, reported results and effect sizes may vary in format between collected studies, and some conversions may be required in order to ultimately transform all effect sizes into Hedges' *g*. If existing, these conversions will be documented.

Coding of studies will be executed, collecting data for the selected variables and subsequently analysing the specific group of interest¹⁰. All the analyses will be executed in R using *robumeta* (Fisher et al., 2017), *metafor* (Viechtbauer, 2010) and *dmatar* (Harrer et al., 2019) packages. This approach was established due to the availability and accessibility of R environments and respective packages, and transparency, as results are more easily replicated.

3.4 A priori Limitations

Among possible limitations for the execution of the proposed Meta-analysis, we find that the number of studies exploring the interaction between numeracy and graph literacy as well as the variety of decision aids tested by these studies may pose as the main constraints, as they may undercut the total number of studies collected and/or the number of different decision aids analysed. The similarity between studies can also pose as a threat, mainly due to differences in sampling, numeracy/graph literacy tests and the decisions on how to divide subjects into high/low groups on both of these skills.

¹⁰As mentioned in previous Section 3.1.1, the possibility of collecting and analysing data on the other three subgroups (low-high, high-low and high-high) remained open at this stage of the research.

Chapter 4

Search, Screening and Data Extraction

As described in Chapter 3, all the decisions related with the Search, Filter and Data Extraction phases were to be disclosed prior to the execution of these steps. This was done through a Protocol, which purpose was to thoroughly address future decisions, by revealing as much of the research goals as possible, thus contributing to the transparency of the whole process. In the present Chapter we will report on the execution of these plans but also evaluate how they unfolded, giving an account of and justifying every adaptation needed. The PRISMA flow diagram, which reports the flow of studies through each of the different steps, will be presented just at the end Section 4.2 so that it may represent the whole process and its intricacies with fidelity.

4.1 Search Execution and Study Screening

As stated, the search strategy resulted in the identification of 1906 potential studies, which after download from their respective databases were uploaded to both OSF and Rayyan. Already within this screening tool, 122 studies were identified as possible duplicates, from which 120 were correctly labeled as such and thus discarded. Following this, the Title and Abstract screening was executed by Francisco Gomes (F) and Vasco Brazão (V), with both researchers going through all 1786 studies. By using the *Blind Mode* in Rayyan each researcher could only see his own decisions, thus enabling full independence between researchers.

Table 4.1 presents the decision matrix that resulted from the Title and Abstract screening, exhibiting an agreement of 98.04% and a Cohen's κ of 0.468¹.

Conflicts were recorded and individually solved with both researchers reaching an agreement. Following this, 31 studies were sought for retrieval, with all but one being successfully collected. As previously planned, both researchers executed a *blind* full-text screening of 8 articles (5% of 30), from which two studies exhibited conflicting decisions, prompting a discussion on the applicability and reasoning of some of the inclusion/exclusion criteria. Both researchers agreed that if criteria 4 and 7 - exploring the interaction of numeracy with graph literacy and including an effect size for the subgroup in study - stood as mandatory the chances of collecting a feasible number

¹There is currently no consensus on how to interpret different magnitudes of Cohen's κ . As a result, this value can denote everything from a weak (McHugh, 2012), to fair (Fleiss, 1981), or even moderate agreement (Landis & Koch, 1977).

Table 4.1: Decision matrix resulting from initial Title and Abstract screening.

F \ V	Inc	Exc
	Inc	Exc
Inc	16	15
Exc	20	1735

of studies could be cut significantly. In fact, a good portion of the collected studies collected all the necessary data: comparing visual aids with textual/numerical formats, measuring objective outcomes (mainly comprehension), and testing both numeracy and graph literacy. Nonetheless, the majority of these lacked the exploration and reporting of the interaction between numeracy and graph literacy, thus failing to analyse the impact of visual aids in the resulting subgroups. Assuming that the search strings and strategy were well built² and the screening correctly executed, it is interesting - but quite disappointing - to notice that despite Garcia-Retamero and Cokely (2017) drawing attention to the potential frailty of showing visual aids to less numerate and graph-literate individuals, recent literature has opted not to investigate this particular issue.

Facing this obstacle, we agreed to drop criteria 4 and 7 and include studies that possessed the necessary data but did not explore or report it, with the subsequent goal of contacting the authors of each study and requesting access to the raw data, thus not only enabling the execution of the meta-analysis *per se* but also potentially enabling more homogeneous decisions on group splitting. Criterion 2 was also loosened, thus allowing, for example, the inclusion of studies exploring comprehension of health decision aids by parents. Following this, the main author executed a new full-text screening, in which 10 studies were selected as eligible (two other studies tested behavior intentions and so were excluded from further analysis). A citation search³ ensued, using both snowball and reverse-snowball methods, and resulted in the selection of two more studies⁴.

At this point it was clear that this research found itself at a crossroads, with its success depending on positive answers by a fair amount of the contacted researchers, as the majority of collected studies did not report necessary data. Given this obstacle and the characteristics of some of the collected studies, we decided to explore the possibility of executing the meta-analysis using either numeracy or graph literacy scores (and not their interaction). This path would forfeit any chances of producing both an high-quality meta-analysis and a scientific article on the matter - as several steps of the search process were not executed towards this goal -, but would nonetheless enable the continuation of what could become a scoping research and meta-analysis on the theme.

²Attesting that a search strategy collected all eligible studies can be quite unreasonable, but to the avail of our own strategy, it did “catch” all the relevant studies included in the systematic review that motivated this meta-analysis.

³As stated in the Protocol, citations of two key studies by Galesic and Garcia-Retamero (2011b) and Garcia-Retamero et al. (2016) were also screened.

⁴These studies were in fact “recovered” by the citation searching, as they were initially collected by the search strategy but were excluded in the Title and Abstract screening.

In order to improve the odds of executing a meaningful meta-analysis, Garcia-Retamero and Cokely (2017) was screened once more for articles exploring numeracy or graph literacy skills independently, which resulted in the addition of four more studies. Thus, the final numbers of collected articles were as follows: 19 for numeracy, 13 for graph literacy and 12 for the interaction between both skills. Identification of the specific studies through each of these steps is available in the Meta-analysis OSF repository (<https://osf.io/62vy5/>).

4.2 Data Extraction and Handling

Following the screening phase selected articles were compiled. By using the data extraction form as a guide, each article was read through in search of the pre-established parameters of interest, whether they were reported in text, tables or plots. If the latter option was the only data source available, data was extracted through the use of the already mentioned WebPlotDigitizer tool (Rohatgi, 2020). We should denote that this tool has limited sensitivity and is thus prone to collect values with minor differences to actual data. Since Standard Errors were reported with upper and lower bounds, both values were collected and averaged, thus minimizing any potential deviation. Some data handling (e.g. transforming SE into SD) was required in order to allow the inclusion of several of the collected studies.

As stated in Section 4.1, some articles were identified as possessing the necessary data although not reporting it, and were selected as eligible with the intention of contacting respective authors to request full participant-level data. All authors were contacted by May 25th 2021.

Throughout the data extraction process, several articles failed to report essential information for the execution of the meta-analysis, such as the number of participants per subgroup⁵, standard error of outcomes, or both. All authors were contacted for the missing information or, if possible, the participant-level data.

The need for requesting data from several authors and studies prompted the postponement of the meta-analysis execution, with July 15th being agreed upon as the deadline to receive any further data. Table 4.2 contains a summary on the number of initially included studies and those to which all necessary data - through all possible means - was available by the aforementioned date.

Table 4.2: Number of studies per parameter of interest and data availability.

	Potential	Viable
Numeracy	19	10
Graph Literacy	13	3
Interaction	12	3

Due to the lack of response by the majority of the contacted authors, and as seen in the table above, only a few collected studies had enough data to be included in the meta-analysis. Among these, three studies from the numeracy branch of this analysis failed to report information on

⁵E.g. number of low-numerates that tested the visual format.

one parameter only (either n size or standard deviation), which could be estimated to render its inclusion possible. The n size of subgroups (tested format x numeracy level) was approximated for study N05, following instructions directly transmitted by the main author. Standard deviations were estimated for studies N07 and N12⁶ by using conversion factors (f) extracted from Walter and Yao (2007) but initially proposed by Tippett (1925)⁷.

Participant-level data was obtained for three studies (four experiments). While one study had its full data posted and available online, we received the remaining two upon successful requests to the respective corresponding authors. In order to extract necessary information these data sets were analysed in SPSS as follows: first a frequency analysis was executed on the total numeracy variable, which allowed us to examine its distribution and median point⁸. This variable was then recoded into a new variable named *Numeracy Split*, which assigned participants to a Low (1) or High (2) numeracy group, according to the sample median score⁹. Each file was then organized by groups based on numeracy split and format/condition¹⁰. Finally, by performing a descriptive statistics analysis of the comprehension variables we could access the n , mean and standard deviation of each numeracy group by format, which were then inputted to the data extraction matrix.

In spite of statements from Borenstein et al. (2009) and Valentine, Pigott, and Rothstein (2010) mentioning that two studies are enough to conduct a meta-analysis, this is particularly true if a fixed-effects model can be assumed and thus employed. In the case of a random-effects model the answer is not so clear and may vary according to the effect size one would deem as important to detect¹¹, the number of available studies, the number of subjects per study and the heterogeneity between studies (the four components of a power analysis).

One characteristic that makes meta-analyses so sought after is that by merging results from several studies we may attain sufficient statistical power to verify even small effect sizes, which individual studies often have difficulty in reaching. Nonetheless, several factors come at play when it comes to the power of a meta-analysis, such as number of studies, number of participants per study, expected effect size and between-study heterogeneity. The power of a test is linked with a Type II or β error (false negative), and can be defined as

$$\text{Power} = P(\text{reject } H_0 | \mu_0 \neq \mu_1) = 1 - \beta \quad (4.1)$$

A power analysis is substantially different from a significance test since it is usually performed

⁶Author shared summary data but no SD.

⁷Originally presented table can be examined in Appendix I.

⁸Prior to this step, in Study N10 four new variables were computed by summing the scores in numeracy, as well as table, bar and icon comprehension.

⁹In Study N11, which included two independent experiments, the split was executed according to the median of the combined samples. This was decided as it would seem incoherent to have the same scores (in the same numeracy test) labelled differently.

¹⁰As participants in Study N10 tested all three formats its output was organized by numeracy split only.

¹¹E.g. if testing a treatment to reduce the risk of developing a deadly disease a small effect size may be enough to be considered important, whereas in the study of a treatment to reduce back pain researchers could decided that only a medium to large effect size could be considered non-trivial.

before data is collected. Thus, and by contrast, when a power analysis is executed we are yet to possess information on values such as the effect size and its precision, having to draw some assumptions guided by available experiments and literature. Conventionally, β is set at 0.2 (Borenstein et al., 2009), which therefore means that we should plan to meet a 80% threshold. For the purposes of this meta-analysis we will execute a power analysis¹² for an expected Cohen's d around 0.2 to 0.4¹³ and 50 participants per group in each study. While Figure 4.1 plots power against effect size with individual curves for each heterogeneity level, Figure 4.2 represents power against number of studies, with curves for three different effect sizes.

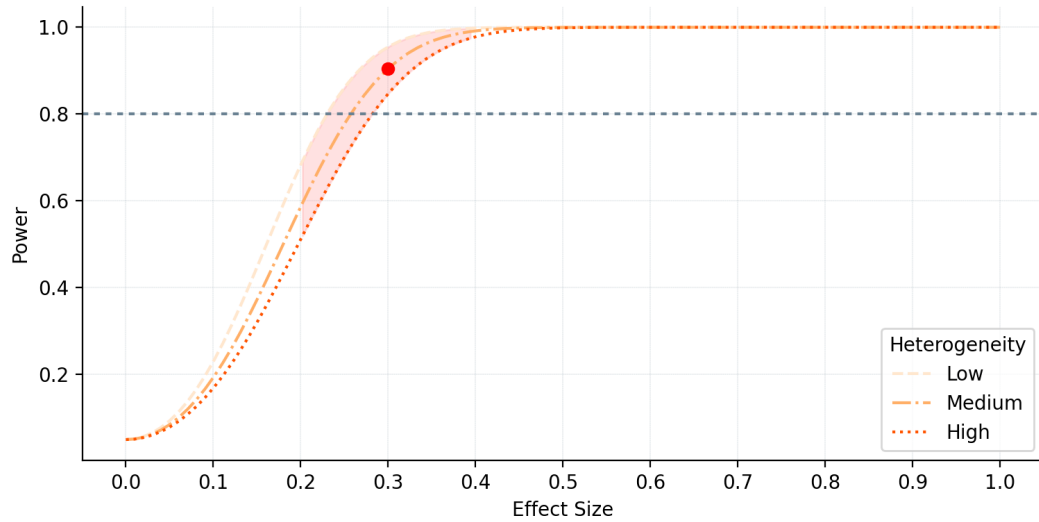


Figure 4.1: Power analysis for effect size and heterogeneity ($k = 8$, $n = 50$). Data point illustrates a power of 0.9045 for a Cohen's d of 0.3 and medium heterogeneity. Area in red highlights possible values for power for a Cohen's d between 0.2 and 0.4, with low to high heterogeneity.

Reviewing Figure 4.1 we can conclude that a meta-analysis with the presented characteristics - 8 studies and 50 participants per condition¹⁴ - should yield enough power for an effect size of 0.3, regardless of the found heterogeneity¹⁵. Nonetheless, it also points towards the possible fragility of the predicted power if the true effect size is lower than assumed.

Additionally, in Figure 4.2 we may observe the great influence that effect size has over a meta-analysis power (higher effect sizes plot much steeper curves), but also the risk of conducting a meta-analysis - with the mentioned characteristics - with fewer than 5 studies. In conclusion, these outcomes led us to decide not to pursue the meta-analysis on graphic-literacy and also its

¹²Despite not addressing here the formulas that ought to be employed, the complete Python script written for this analysis is available through the OSF repository (<https://osf.io/wt6g9/>).

¹³This reference value was chosen as it represents a small to medium effect size (Cohen, 1988), a threshold we do not expect the meta-analysis summary effect to fall above.

¹⁴Collected experiments would later have an average of approximately 167 low numerates per group.

¹⁵This could be disputed since a very low number of studies may not translate actual dispersion (Borenstein et al., 2009)

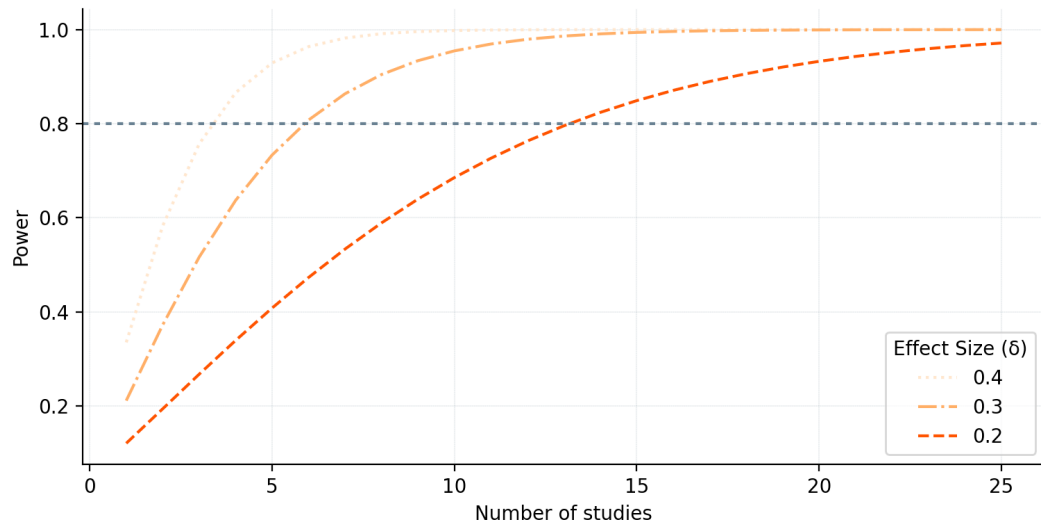


Figure 4.2: Power analysis for number of studies and effect size (heterogeneity = medium, $n = 50$).

interaction with numeracy since its results would most likely yield insufficient power due to the low number of potential and viable studies. Oppositely, with 10 collected studies (containing a total of 12 experiments), a meta-analysis on numeracy would seem to have a fair chance of attaining adequate power.

Below Figure 4.3 exhibits the PRISMA Flow Diagram of studies from the search execution to final included studies.

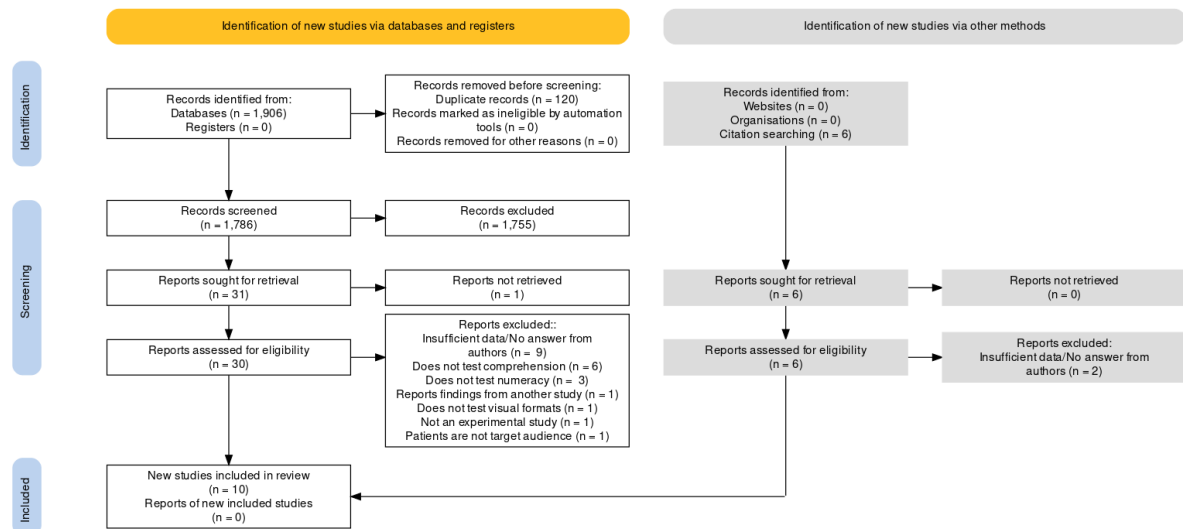


Figure 4.3: PRISMA Flow Diagram depicting the number of studies in each step.

This diagram addresses the different stages of the process with by respecting their specific goals, such that the 1906 initially collected studies were a result of a search strategy devised for a meta-analysis on the interaction between numeracy and graph literacy, while the 10 final studies are part of a numeracy-only analysis, as described above.

Table 4.3¹⁶ exhibits all included articles, listing characteristics such as the first author's name, year of publication and which Numeracy/Graph literacy tests were used. Further below, Table 4.4 displays the main sample characteristics for each experiment (note that two studies contain two independent experiments each).

Table 4.3: Simplified table with characteristics of collected studies.

Study Characteristics						
ID	First Author	Year	Source	N Test	Split	Design
N01	Danya, H.	2021	Published	Schwartz*	Median	Exp
N02	Retamero, R.	2015	Published	Schwartz & Lipkus	Median	Exp
N03	Nayak, J. G.	2016	Published	SNS	NR	Rep
N04	Galesic, M.	2009	Published	Schwartz & Lipkus	Within-group Median	Exp
N05	Tait, A. R.	2010	Published	SNS	Low = 0-35; High = 36-48	Exp
N06	Tait, A. R.	2010	Published	SNS	Low = 0-35; High = 36-48	Exp
N07	Okan, Y.	2018	Published	BNT + Lipkus	Median	Exp
N08	Durand, M.	2020	Published	SNS3	Median	Rep
N09	Petrova, D.	2015	Published	Schwartz & ABNT	≤5	Exp
N10	Turchioe, M.	2020	Published	SNS3	≤Median	Rep

Note. SNS = Short Numeracy Scale (2007); SNS3 = 3-item Subjective Numeracy Scale (2015); Schwartz = Schwartz et al. (1997); Lipkus = Lipkus et al. (2001); BNT = Berlin Numeracy Test (2012); ABNT = Adaptive Berlin Numeracy Test (2012); NR = Not Reported; Exp = Experimental; Rep = Repeated Measures

'&' implies a blend of questions from both tests, while '+' means separate execution of tests.

* In Japanese.

Despite not being able to execute a meta-analysis while exploring the initially intended subgroup - low numerates and low graph-literates -, we should note that previous literature found numeracy and graph literacy to be moderately correlated (Nayak et al., 2016). To some extent, we can thus expect that part of the low numeracy participants of the included studies also exhibit low graph literacy. In a sample of 492 Americans and 495 Germans, Galesic and Garcia-Retamero (2011b) found that while 51% of participants had low numeracy, 35% had both low numeracy as well as low graph literacy (proportions were similar in each country). We can further hypothesize

¹⁶Reduced version compared to matrix available at OSF, as Doi, Title and Real Medical Decision were omitted and information on Numeracy and Graph Literacy tests simplified.

Table 4.4: Simplified table with characteristics of collected experiments.

Sample Characteristics							
ID	n	Type	Age	Female	$\leq$ High school	Nationality	Num $\bar{x}$
N01	1062	General	47.8	49.4%	29.4%	Japanese	2.2
N02	108	Patients	52	78%	86%	Spanish	NR
N03	50	Patients with prostate cancer	71.2	0.0%	22.0%	American	4.6
N04a	59	Older adults	NR	49.0%	57.0%	German	10.3
N04b	112	University students	NR	57.0%	0.0%*	German	8.6
N05	4685	Parents	39.2	57.5%	18.9%	American	36
N06	408	Parents	35.9	74.4%	17.2%	American	39
N07	611	General	33	57.61%	8.0%	American	1.09/1.92 [†]
N08	436	Adults with low economic status	40.5	82.3%	41.5%	American	11.1
N09a	256	Adult men	36	0.0%	12.0%	American	4.48
N09b	355	Adult women	38	100%	NR	American	3.26
N10	40	Adults with heart failure condition	61.3	22%	45%	American	12.2

Note. Age = Average Age; Num $\bar{x}$ = Numeracy mean; NR = Not Reported.

* University students were interpreted as having education higher than High School.

[†] Results for BNT and Lipkus numeracy tests, respectively.

that a meta-analysis on low numerates alone, despite having the potential to cast some light on the comprehension of visual aids by individuals that are also less graph-literate, will eventually fail to grasp the granularity of the interaction between numeracy and graph literacy. As we saw above, within low numerates some people still display high graphical literacy skills, which would theoretically mean that a meta-analysis focused on low numerates should display an effect size more towards visual aids than one exploring low numerates that are also low graph literates.

4.3 Limitations

As theoretical proposals gave way to their practical execution and as expectations met reality, carrying out the steps described above induced a few changes to the initial plans, some of which already described above. Although implemented in favor of the research, these changes unveiled some of the underlying frailties of the proposed strategies.

Inclusion and exclusion criteria could have been clearer: simple tables should have been announced as valid numerical/textual formats (specifying the criteria for this validity); exploring the interaction between numeracy and graph literacy should not pose as a “hard” inclusion criteria, as articles testing both these skills - but not exploring their interaction - can be collected and later

their authors contacted; having to include an effect size for the specific subgroup of interest follows the same reasoning and should be discarded as an inclusion criteria, as some studies may have this information despite not reporting it¹⁷.

As both screening researchers have good comprehension and reading skills in Spanish and Portuguese, articles written in these languages could have been searched for and eventually included. Despite being likely insignificant when it comes to Portuguese, this limitation is more severe in the case of Spanish¹⁸, as Rocio Garcia-Retamero, one of the most prominent researchers in the field of Risk Communication/Literacy, is of Spanish nationality and lectures at the University of Granada.

Having decided to solely include experiments with visual aids designed towards patients may have undermined the total amount of collected studies, since throughout the various screening steps several articles were excluded as they explored either comprehension on medical staff (e.g. including doctors or nurses). Such studies could have been initially included and eventually contribute to richer subgroup analyses (patients/professionals/others). Ultimately, and with access to the participant-level data of all eligible studies, we could even execute the split between low and high skill groups consistently across studies, thus overcoming possible inconsistencies between median splits of samples of different nature¹⁹. Additionally, this criteria was not clear as we could also have included studies exploring health-centered data for single users and not necessarily patients.

In regard to the collected studies, not having access to the respective data sets and, in some cases, not being able to collect any data at all was one of the major limitations. In fact, this ended up shifting the focus of the entire meta-analysis and also posed as a minor limitation in this newly focused endeavour. Surpassing this hurdle meant using WebPlotDigitizer (Rohatgi, 2020) to extract data from some studies, but also imputing values like *n* size and SD of others, with both these methods adding uncertainty to the process and its results.

Expanding on this, data imputations and particularly the decisions that led to them could have been disclosed beforehand. Indeed, the Protocol lacked the forecast of partially missing data, and thus not only it did not strategize on how to deal with it, but also did not state what characteristics would allow an incomplete study to be included or not. Also, as some doubts emerged on how to classify study designs, we understand that the data extraction matrix lacked proper planning and would have benefited from preceding discussions on which characteristics to take into account.

Finally, devising and executing a search strategy oriented to collect studies that explored both graph literacy and numeracy - and having to apply it to a relatively different scope - was possibly the most severe limitation this research underwent. Without doubt, several studies were either

¹⁷In established research fields or sought-after questions, the inclusion criteria mentioned above could have worked well as a filter, but when facing a narrow research question as ours they were found to be exceedingly limiting.

¹⁸Although Spanish was not included in the search filters, at least one article written in this language was collected. Despite exhibiting some potential to be included, this article was discarded in the Title and Abstract screening in order to maintain cohesion of criteria and process integrity.

¹⁹E.g. we would expect a sample of doctors to have a higher median score of both numeracy and graph literacy when compared to a sample of adults of lower economic status.

not collected by the search strings or rejected in the Title and Abstract screening, which has left us with what is probably a subgroup of existing and eligible studies, fact that should prompt some caution in the analyses to come.

Chapter 5

Meta-analysis

The present Chapter will enclose all steps within the meta-analysis execution, from data processing to the interpretation of results and subsequent discussion. As implemented in previous chapters, all relevant matrices are available in the OSF repository.

5.1 Data Processing

In this Section we will report on the initial operations required to transform the extracted data and render it inputtable for the succeeding Meta-analysis. As previously done with in Section 4.2, matrices containing all handling and/or conversions are available at the OSF repository (<https://osf.io/y5pcx/>), granting access to all formulas and calculations.

Before drawing comparisons and therefore transforming the extracted data to the desired effect measure, some planning and data processing is required. Several of the collected studies are effectively multi-arm studies, as they explore three different formats, which prompts a decision on how to deal with this. Considering the acute complexity of carrying out a network meta-analysis and following recommendations from J. P. T. Higgins et al. (2019), multi-arm studies will be transformed into pairwise comparisons by combining data into Visual and Non-Visual categories. Formats will be combined as displayed in Table 5.1 below. Formats exploring relative risk reduction (only present in Study N05) were ignored given the inefficacy to which this format is consensually associated by available literature, as mentioned earlier in Chapter 2). Additionally, a format that combines Non-Visual and Visual formats will be labeled as Visual.

Table 5.1: Formats and the categories into which they should be combined.

Original Format	Category
Text, Numbers, Table	Non-Visual
Bar chart, Icon array, Line chart, Dot plot	Visual

Combination of groups follows two different approaches depending if the study design is within or between-subjects. In the case of independent groups, sample size, mean and SD are calculated using the following three formulas, as seen in J. P. T. Higgins et al. (2019, p. 168).

$$N_{comb} = N_1 + N_2 \quad (5.1)$$

$$\bar{X}_{comb} = \frac{N_1 \bar{X}_1 + N_2 \bar{X}_2}{N_1 + N_2} \quad (5.2)$$

$$SD_{comb} = \sqrt{\frac{(N_1 - 1)SD_1^2 + (N_2 - 1)SD_2^2 + \frac{N_1 N_2}{N_1 + N_2}(M_1^2 + M_2^2 - 2M_1 M_2)}{N_1 + N_2 - 1}} \quad (5.3)$$

When it comes to matched groups or paired samples, the formulas are

$$\bar{X}_{comb} = w\bar{X}_1 + (1 - w)\bar{X}_2 \quad (5.4)$$

$$SD_{comb} = \sqrt{w^2 SE_1^2 + (1 - w)^2 SE_2^2 + 2w(1 - w)r SE_1 SE_2} \quad (5.5)$$

where r is the correlation between the two tested formats and

$$w = \frac{SE_2^2 - r SE_1 SE_2}{SE_1^2 + SE_2^2 - 2r SE_1 SE_2} \quad (5.6)$$

Given that the mean scores for each intervention are known for every study except one, the standardized mean difference (δ) - converted into Hedges' g - is the most appropriate effect measure to use in the meta-analysis. We begin by estimating d using the same formula for both independent and pre-post scores or matched groups with

$$d = \frac{\bar{X}_1 - \bar{X}_2}{S_{within}} \quad (5.7)$$

where $\bar{X}_1$ and $\bar{X}_2$ are the mean scores for the control and intervention groups¹. S_{within} , which estimates the within-group standard deviation, is different between study designs, as we can see in below formulae 5.8 (independent) and 5.9 (pre-post or matched samples).

¹The order of the groups may be changed according to user preference, thus affecting the direction of the effect size. Nonetheless, caution is advised as the direction must be the same in every comparison.

$$S_{within} = \sqrt{\frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}} \quad (5.8)$$

$$S_{within} = \frac{S_{diff}}{\sqrt{2(1 - r)}} \quad (5.9)$$

where

$$S_{diff} = \sqrt{S_1^2 + S_2^2 - 2 \times r \times S_1 \times S_2} \quad (5.10)$$

In above formulae - extracted from Borenstein et al. (2009, p. 24-29) -, S_1 and S_2 refer to the standard deviation of the control and intervention groups, whereas n_1 and n_2 are the respective sample sizes. Finally, r is either the correlation between subjects in matched pairs or between pre and post scores. Other differences between designs occur upstream in the estimation of several parameters, and although we will not delve into every estimation or calculation, we chose to report the formulae above as they are key to understanding decisions that might sprout later in the meta-analysis process.

If studies with matched and pre-post designs do not report the value of r , only with access to the participant-level data can we calculate this value, as we are interested in the correlation between pairs of observations (Bakdash & Marusich, 2017). On these terms, Borenstein et al. (2009) point towards (1) choosing a correlation from related studies or (2) performing a sensitivity analysis with several acceptable correlations. In case such decisions are needed we expect to perform a sensitivity analysis with a range of r values between 0.25 and 0.75, subsequently reporting on the impact and degree of uncertainty related with this parameter.

5.2 Data Imputation

Apart from the eleven experiments that report mean score for each condition/numeracy level, one collected experiment chose to report odds ratio and respective confidence intervals. Odds ratio is an effect measure based in the number of events and non-events in two groups, so its resulting value already reports on a comparison between groups. Below formulas can be found in Hasselblad and Hedges (1995, p. 170) and Borenstein et al. (2009, p. 47).

$$OddsRatio = \frac{A/B}{C/D} \quad (5.11)$$

the variance of its natural logarithm

$$V_{LogOddsRatio} = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D} \quad (5.12)$$

where A and B are the number of events and non-events in the exposed group, while C and D refer to the same information in the control group. The conversion from an Odds Ratio to a standardized mean difference can be done by using

$$d = \text{LogOddsRatio} \times \frac{\sqrt{3}}{\pi} \quad (5.13)$$

and for its variance

$$V_d = V_{\text{LogOddsRatio}} \times \frac{3}{\pi^2} \quad (5.14)$$

Nonetheless, while Study N01 addresses two comparisons for each numeracy group (low/high) - Text vs. Bar graph and Text vs. Pictograph - it does not report its subgroup n nor its 2 X 2 tables, which we will need to combine both comparisons into a single Visual vs. Non Visual comparison. While Schwarzer, Carpenter, and Rücker (2015) suggest that participant numbers may be drawn at random from similar studies, another solution seemed to potentially attain much more plausible results.

Among disclosed data, Study N01 reports n per format (e.g. Text = 354), Odds Ratio per numeracy level X format comparison (e.g. Low Numeracy X Text vs. Bar graph = 0.700) and respective 95% confidence intervals. Apart from the formula used to calculate an Odds Ratio (Formula 5.11) we also know how to calculate its 95% confidence interval, as presented in Szumilas (2010)

$$95\% CI = \exp\left(\ln(OR) - 1.96 \times SE\{\ln(OR)\}\right) \text{ to } \exp\left(\ln(OR) + 1.96 \times SE\{\ln(OR)\}\right) \quad (5.15)$$

Having these formulas enables us to look at the missing 2 X 2 tables as an optimization problem with the goal of finding the proportions of events and non-events that result in the reported values. Despite having solved this problem for all four interactions between comparisons and numeracy groups at the same time, for the sake of simplicity the explanation that follows will refer to one group only. Initially, A , B , C and D are set as changing variables, since these are the values we seek to discover. The formulas for the Odds Ratio, its variance and 95% confidence interval lower limit are set in three different cells, directly influenced by the cells selected as changing variables. We set, for each one of these three cells, a constraint² to force its resulting value to match the value reported in the study (with some slack). Other constraints were added, such as integer constraints for the changing variables, or constraints that coerce the total number of participants per condition to match those divulged by the study; also, in the Text condition the number of events and non-events was forced to be the same in both comparisons. The complete model can be examined in a file uploaded to the OSF repository (<https://osf.io/m63de/>).

After executing the model with the GRG Nonlinear method a solution was found, in which all constraints were met with a maximum delta of 1% when compared to the reported OR and

²Actually one cell has to be set as the objective, but the reasoning remains.

confidence interval limits. Table 5.2 below contains the found values for A , B , C and D in all four interactions between comparisons and numeracy groups, and Table 5.3 juxtaposes reported data with data outputted by the model.

Table 5.2: Predicted events and non-events for both intervention and control groups in each interaction between numeracy and format comparisons in Study N01.

	Low Numeracy		High Numeracy	
	TB	TP	TB	TP
A	93	100	153	166
B	63	53	48	30
C	157	157	97	97
D	74	74	25	25

Note. T = Text; B = Bar graph; P = Pictograph. A and B refer to the events and non-events in the respective visual format, C and D to the corresponding values in the Text format.

Table 5.3: Descriptive statistics reported by Study N01 compared with the values predicted by the optimization model.

	Low Text/Bar		Low Text/Picto		High Text/Bar		High Text/Picto	
	Rep	Pred	Rep	Pred	Rep	Pred	Rep	Pred
OR	0.700	0.696	0.892	0.889	0.829	0.822	1.412	1.426
LCI	0.459	0.456	0.579	0.577	0.480	0.476	0.786	0.793
UCI	1.067	1.062	1.374	1.371	1.434	1.419	2.537	2.565
Max Δ	0.005		0.003		0.015		0.018	
Max % Δ	0.7%		0.5%		1%		1%	

	Text		Bar		Picto		All	
	Rep	Pred	Rep	Pred	Rep	Pred	Rep	Pred
n	354	353	358	357	350	349	1062	1059
Δ	1		1		1		3	

Note. Rep = Reported; Pred = Predicted.

Lastly, for Study N02 mean score and standard error were known for the interaction between numeracy groups and each of the two tested formats (numerical and icon-array), but information on n size was missing. Authors reported that the low numeracy group had 52 participants while the high numeracy had 56, with a between-formats study design. Following this, it was imputed that half of each numeracy group tested the numerical format while the other half tested the icon-array.

5.3 Results

All matrices containing extracted data, calculations and values inputted into the meta-analysis model are available at the OSF repository (<https://osf.io/y5pcx/>), as well as the R script built for the analyses that will follow (<https://osf.io/a64vt/>).

Our search identified 1906 potential studies, from which 6 were included in our analyses after reviewing inclusion criteria and extracting/requesting relevant data; 4 additional studies were added after citation searching procedures. Within these ten studies, 12 experiments had available data and were thus included in our meta-analysis.

Included studies were executed either in the United States ($n=7$), Europe ($n=2$), or Japan ($n=1$), all between 2009 and 2021 (median = 2015.5). Five studies employed some form of subjective numeracy tests while the remaining used objective tests. Majority of experiments ($n=11$) featured a sample of adults, only three of which worked on a sample of patients or participants with a health condition. Average n was 681, with a mean age of 39.3³, an overall share of female participants of 58.34%, and 21.1% of participants with high school or less.

Our main analysis produced the forest plot exhibited in below Figure 5.1. An overall g effect size of -0.28 was identified⁴. A shallow analysis could state that these results favour Visual aids; nonetheless, as the confidence interval overlaps the null effect and heterogeneity is found to be very high ($Q = 693.11$; $T = 0.59$; $I^2 = 95.6\%$), further inquiries are required to sustain or reject this conclusion.

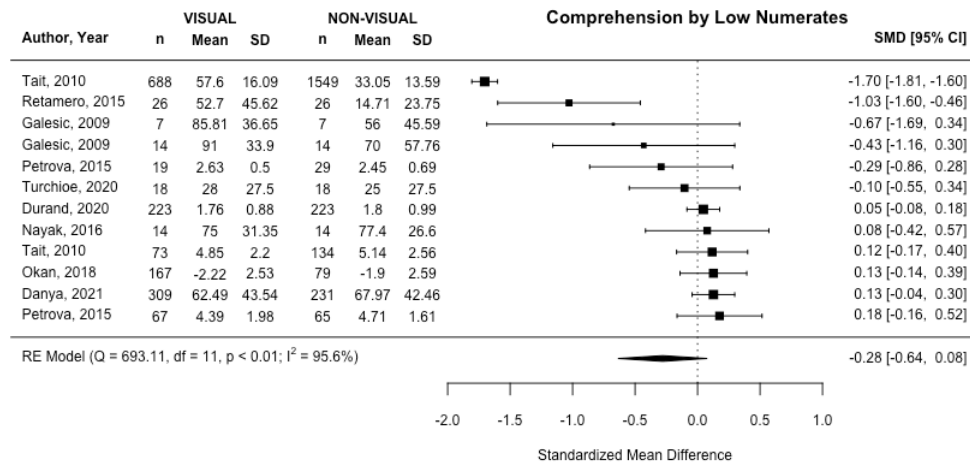


Figure 5.1: Forest plot on the comprehension of visual vs. non-visual aids by low numerates, displaying common and per study effect sizes, heterogeneity analysis and summary statistics. Square size represents individual study weight within the analysis. Ordered by effect size.

Furthermore, we can visually infer the first two experiments as potential outliers, due to the

³Incomplete data as two experiments did not report participant's age.

⁴A small sensitivity analysis was performed by varying r values between 0.25 and 0.75 in experiments with matched pair samples. No change in outcomes was found.

magnitude and confidence intervals each reports, specially the experiment included in Tait et al. (2010a). These experiments are in fact the only two that point towards a clear advantage of one format over the other, as all of the remaining experiments report confidence intervals that enclose a g of 0.

5.3.1 Heterogeneity

As mentioned previously, results point towards a significant amount of heterogeneity between collected experiments, with a T of 0.59, an I^2 of 95.6% and a Q of 693.11 ($p < .0001$).

Figure 5.2 presents a Baujat plot mapping each included experiments in accordance to its contribution to the overall result (vertical axis) and to heterogeneity (horizontal axis). From this graph we may confirm earlier inferences by denoting that experiments N05 and N02 have particular contributions to the heterogeneity of the model, while N05 is highly influential to its overall result. Thus, results from these two experiments appear once again as potential outliers. Influence diagnostics by study can be further examined in Appendix J.

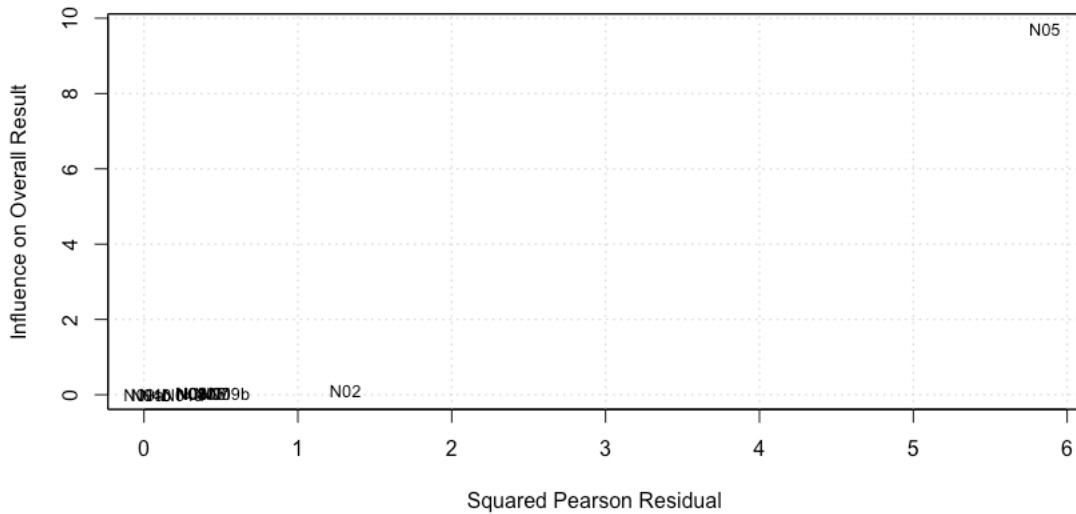


Figure 5.2: Baujat plot charting each experiment according to its influence in common effect size and heterogeneity.

The identification of experiments N02 and N05 as potential outliers prompted a more in-depth analysis on the impact each observation has on the overall results of the model. As such, a Graphic Display of Heterogeneity (GOSH) analysis, was carried out. This method, initially proposed by Olkin, Dahabreh, and Trikalinos (2012), generates all possible 2^{k-1} combinations of experiments, calculating and plotting each in accordance to its effect size and heterogeneity.

In Figure 5.3 we can observe the mapping of all generated subsets with a particular focus to those that include experiments N02 and N05 (highlighted in blue). By analysing this, we may conclude that the vast majority of subsets that feature experiment N02 exhibit an I^2 above 50% and an effect size from -1.0 to just under 0, while models that include experiment N05 are much

more condensed regarding heterogeneity, with an I^2 roughly above 80%, and have a distribution of effect sizes more towards visual aids, from -1.0 to roughly -0.2.

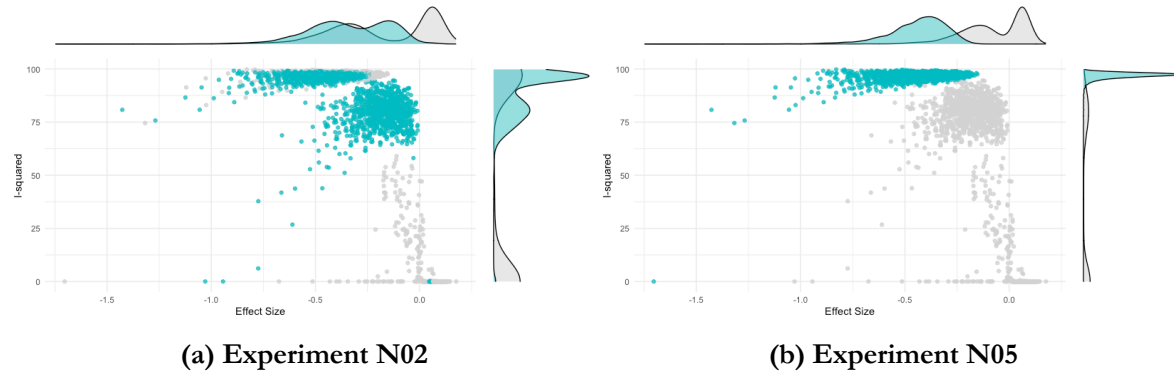


Figure 5.3: GOSH plots highlighting (in blue) models that included studies (a) N02 and (b) N05, against models that did not include these studies (in gray). Models are plotted according to their I^2 and overall effect size values.

Following this, a GOSH diagnostics was called⁵, resulting in Figure 5.4, which plots all generated models colouring each one according to its k-means cluster⁶. Additionally, we may examine the influence each study has on the imbalance of each cluster. Not surprisingly, Experiments N02 and N05 are the most prominent contributors to overall heterogeneity.

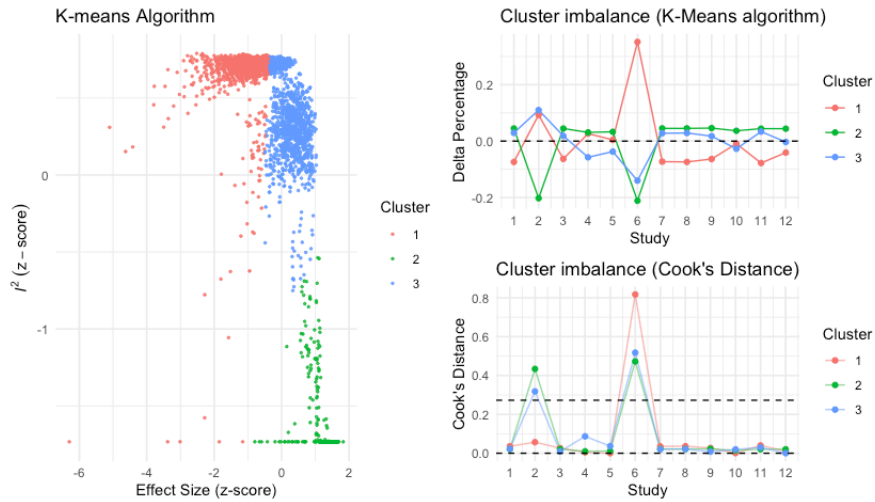


Figure 5.4: *Left:* GOSH plot mapping the 4095 models against their I^2 heterogeneity and Standardized Mean Difference z-scores (colored by K-means cluster). *Right:* Per-study analysis of influence in K-means cluster imbalance. Studies numbered 1 through 12 as ordered in Table 4.3.

⁵The function *gosh.diagnostics* in package *dmatar* runs its input through three different clustering algorithms: k-means, DBSCAN and GMM.

⁶K-means was initialized with 3 centers.

In order to cast some light onto the impact of removing these two outliers from our analysis, we executed the random effects model twice more. First we withdrew N05, by far the most evident outlier, and then we additionally removed N02, recording respective measures of heterogeneity and overall effect size on both occasions. Results are presented in Table 5.4 below.

Table 5.4: Main outcomes of main meta-analysis compared with cumulative removal of two most influential experiments.

Analysis	k	g	95% CI	p	95% PI	I^2	95%CI
Main Analysis	12	-0.28	-0.64 0.08	0.12	-1.48 0.92	95.6%	90.4 98.4
- Tait, 2010a	11	-0.03	-0.19 0.13	0.13	-0.42 0.36	58%	5.9 94.4
- Retamero, 2015	10	0.06	-0.02 0.15	0.13	-0.02 0.15	0%	0 86.7

Interestingly, removing these two outliers provokes a gravitation of the overall effect size that favours non-visual aids. First, without experiment N05, the overall effect size shifts from a g of -0.28 to -0.03 [95 % CI (-0.19 to 0.13)], while resulting I^2 drops from 95.6% to 58% (medium heterogeneity). By further removing experiment N02, g is outputted as 0.06 [95 % CI (-0.02 to 0.15)] and I^2 is non-present. Given this reduction in heterogeneity, the Prediction Intervals exhibited an expected decrease in range.

Lastly, it should be noted that researchers behind experiment N05 - in Tait et al. (2010a) - were also responsible for experiment N06 - in Tait et al. (2010b) -. While both were carried out in the United States in 2010, using fairly similar sample types (parents with at least one children), and employing the Subjective Numeracy Scale with the same cut-off between low/high skilled groups, they reported dramatically different values. This inconsistency certainly does not add confidence to the analyses and should inspire deeper scrutiny into identifying its origin.

5.3.2 Subgroup Analysis

Explaining the source of heterogeneity in a meta-analysis can sometimes be achieved by analysing available data in search for moderating variables. Therefore, a subgroup analysis was executed and respective forest plots can be examined in below Figure 5.5.

First and foremost, we should denote that the analyses in Figures 5.5a and 5.5b are limited by the size of its subgroups; not complying with the guideline of a minimum of four studies per subgroup (Fu et al., 2011) should thus prompt some caution when analysing these forest plots. The fact that two studies in our analysis contain two experiments and that some authors are in more than one study should also have an effect on the logic behind these analyses.

Figure 5.5a splits the meta-analysis according to the continent in which each study was executed. As with previous analyses, experiment N05 reports an effect size vastly different from other studies, more so when compared to those from North America. In studies from Europe heterogeneity is rather low ($Q = 1.66$; $I^2 = 6.2\%$). When it comes to both the type of health behaviour as numeracy test executed, no moderating effect is apparent. Given the already mentioned low number of total and within-subgroup studies, these analyses do not offer much more

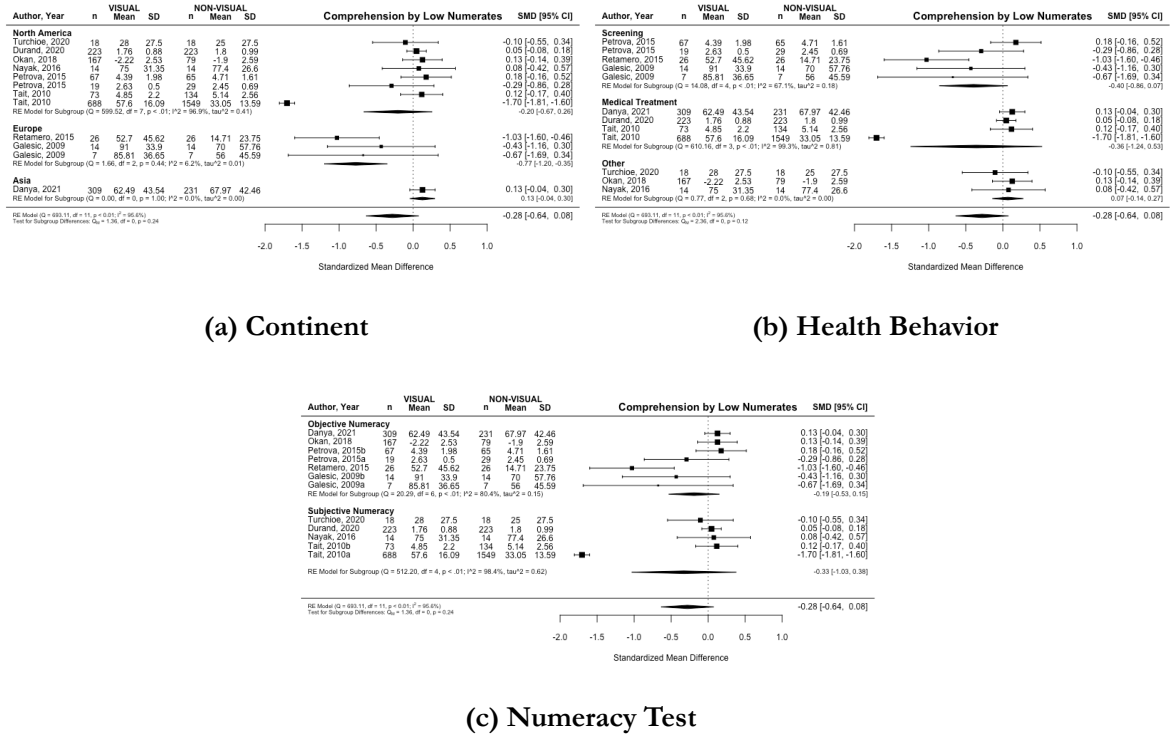


Figure 5.5: Forest Plots with subgroup analyses by (a) Continent, (b) Health Behavior and (c) type of Numeracy Test. Ordered by year, descending.

than a scoping examination of variables and would certainly be of greater value in a larger meta-analysis. Also, as experiments N02 and N05 are solely accountable for the found heterogeneity, the potential for a meaningful subgroup analysis was already limited prior to its execution.

5.3.3 Meta-regression

Meta-regression analyses were put into action regarding variables *Age*, *Year of Publication*, *% Female* and *% High School or less*. Figure 5.6 below exhibits the resulting plots.

By analysing Figure 5.6 we may infer that variables *Age* and *% Female* play no role in moderating effect size, as the Standardized Mean Difference (SMD) maintains its level - given by the regression line - regardless of the values in the aforementioned variables. On the other hand, overall effect size seems to be positively correlated with *Year of Publication*⁷ - the later the study is published the larger its effect size is - and negatively correlated with *% High School or less* - the higher its value (i.e. the lower the education levels are) the lower the effect size is. This would indicate that later studies and more educated samples slightly favour non-visual aids.

As described above in Section 5.3.1, experiments N02 and N05 were found to be outliers, as they reported materially different values from other experiments and contributed to widen overall

⁷This effect may also be denoted in the Cumulative Meta-analysis reported in Appendix K.

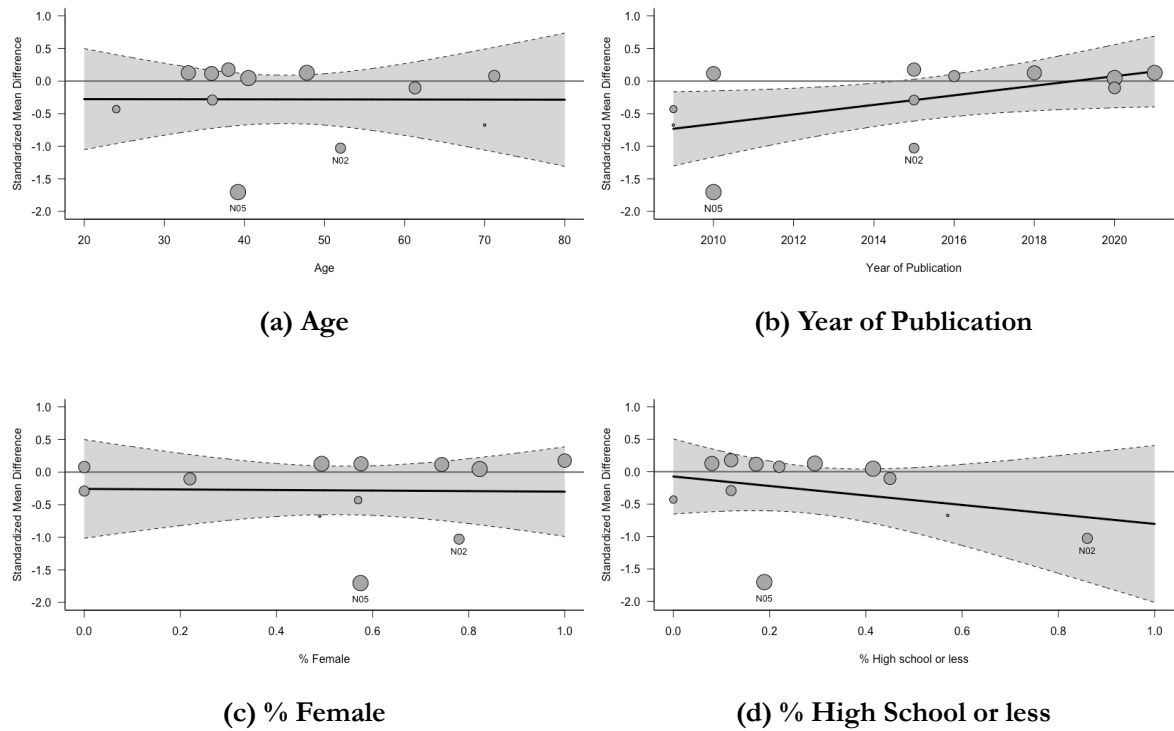


Figure 5.6: Meta-regression plotting each experiment according to its effect size and (a) participants' mean age, (b) year of publication, (c) % Female participants and (d) Education. Size of points is a function of study weight.

heterogeneity. As such, meta-regression analyses without these two experiments were also put into place, as reported by Figure 5.7.

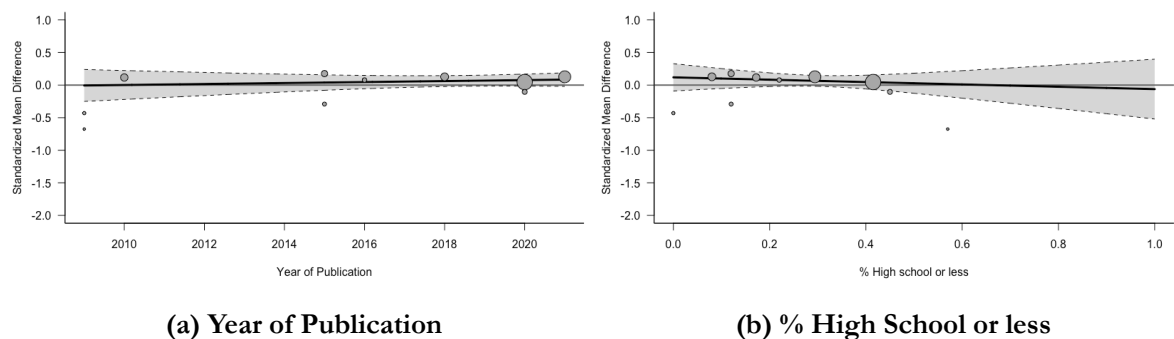


Figure 5.7: Meta-regression - excluding outliers N02 and N05 - according to effect size and (a) year of publication or (b) Education. Size of points is a function of study weight.

As expected, removing experiments N02 and N05 resulted in the erasure of the correlation between effect size and *Year of Publication* or *% High School or less*, as may be induced by the flat regression lines in both plots. Regression coefficients for all four variables, before and after the removal of outlier experiments, can be addressed in Table 5.5.

Table 5.5: Regression results for variables *Age*, *Year of Publication*, *% Female* and *% High School or less*, indicating regression coefficients and respective p-values.

Variable	Main Analysis		Without Outliers	
	Coef.	<i>p</i>	Coef.	<i>p</i>
Age	-0.0001	0.9918	-0.0014	0.7894
Year	0.0733	0.0583	0.0075	0.5471
Female	-0.0042 [†]	0.9468	0.0134 [†]	0.4635
School	-0.0732 [†]	0.3729	-0.0181 [†]	0.5821

Note. *Female* = % Female; *School* = % High School or less; [†] = Refers to change of 10% in respective variable.

From the results in above Figures 5.6 and 5.7 as well as Table 5.5 we may conclude that no variable played a consequential role in moderating overall effect size. Only *Year* exhibits some impact, albeit borderline statistically non-significant, which in turn is completely mitigated by the removal of experiments N02 and N05. The withdrawal of these experiments did not incur in major changes in the other variables under scrutiny. Notwithstanding, these plots and table should be examined with some caution, as significant within-study differences between low and high numerates in *Age*, *% Female* or *% High School or less* are potentially being ignored in these analyses, as included data refers to the experiment sample and not the low numeracy subgroup.

5.3.4 Publication Bias

Despite our efforts in the Search phase, we could not access any unpublished studies on the topic under examination, which therefore can raise the risk of Publication Bias. To assess this we began by generating a funnel plot for our main analysis. This tool charts included experiments according to their respective effect sizes and standard errors⁸. We may then perform a visual analysis on the resulting plot, specially looking for any asymmetry near its bottom part, which could therefore mean that small non-significant studies may be missing.

Figure 5.8 exhibits the plot generated by our analysis⁹, alongside limit estimates for effect size¹⁰. By examining this funnel plot we may denote that, as expected, experiments N02 and N05, which were deemed as outliers in previous analyses, are plotted with considerable separation to other studies, being in fact the only two experiments in the $p \leq 0.01$ area. A visual analysis of the plot could suggest some asymmetry, as the bottom right part of the funnel is completely void of points. Since the overall effect size leans towards the left, the asymmetry, if present, was to be expected exactly in this section of the graph.

⁸Other precision measures may be used, such as sampling variance or inverse standard error.

⁹Appendix M contains a variation of this funnel plot centered in the found effect size.

¹⁰By using a precision measure as a predictor we may estimate effect size as error approaches 0 (infinite precision).

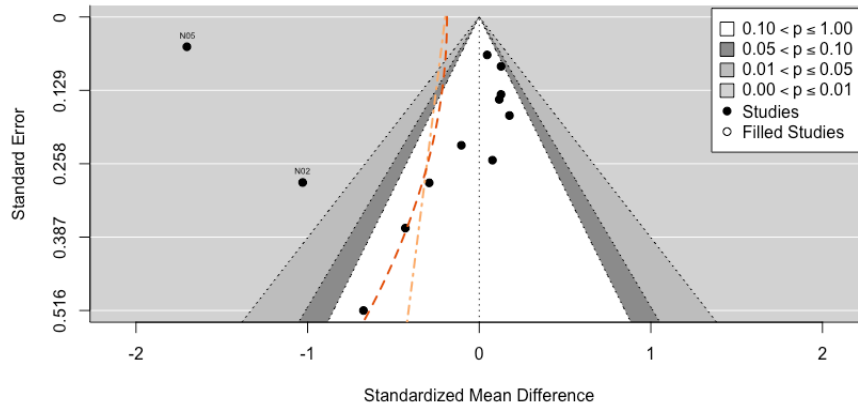


Figure 5.8: Funnel plot for main analysis. Limit estimates given by lines using standard error (orange) or sampling variance (red) as predictors. Contour lines for several statistical significance levels. No experiment was added or removed by trim-and-fill method.

Considering the results from the regression test for funnel plot asymmetry [SE as predictor ($p = 0.7893$), sampling error as predictor ($p = 0.5614$)] no studies were trimmed nor filled, albeit the rank test - which analyses the correlation between observed effect sizes and their variances - reflecting significant asymmetry ($p = 0.0138$). Nonetheless, given the subjectiveness of the visual analyses and the rather low total number of included experiments, any conclusion or interpretation on funnel plot asymmetries should be taken with caution, if at all.

Following the previous course of action, we decided to explore the impact of experiments N02 and N05 by removing them from the analyses. Figure 5.9 is the result of such exploration, presenting funnel plots without the aforementioned experiments (a) before and (b) after the use of the trim-and-fill method, which imputes studies to bring symmetry to the funnel plot.

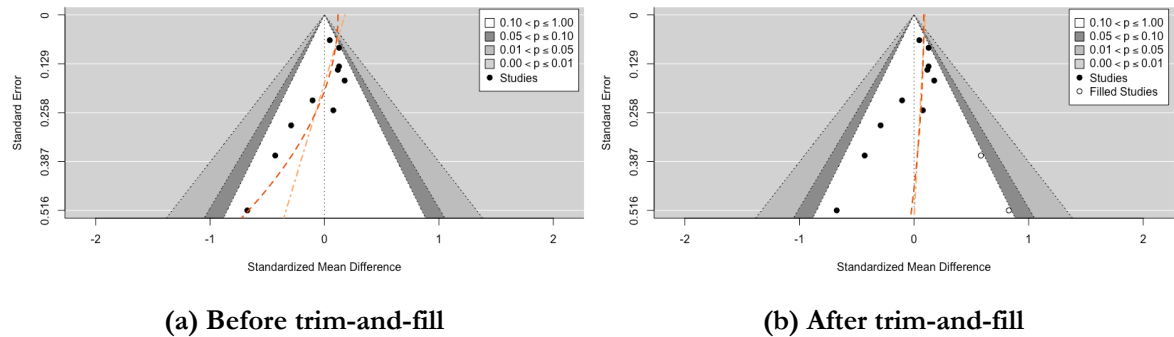


Figure 5.9: Funnel plots for main analysis after removal of outlier experiments N02 and N05, (a) before and (b) after trim-and-fill. Limit estimates given by lines using standard error (orange) or sampling variance (red) as predictors. Contour lines for several statistical significance levels.

Naturally, removing experiments N02 and N05 skews the overall effect size towards the right and into a range of values that favours non-visual aids, similarly to what was presented in above Table 5.4. However, the removal of these experiments further emphasized the presence of asymmetry, now statistically significant in the regression test [sampling variance as predictor ($p = 0.0375$)]. As such, the trim-and-fill method resulted in the addition of two new studies in bottom right corner of the funnel (hollow circles).

Table 5.6 below presents the limit estimates for overall effect size in the above funnel plots.

Table 5.6: Regression limit estimates as precision approaches infinity in presented funnel plots, using standard error or sampling variance as predictors.

Predictor	Overall effect size limit		
	Fig. 5.8	Fig. 5.9a	Fig 5.9b
Std. Error	-0.1978	0.1788	0.0985
Variance	-0.1879	0.1179	0.0850

Given the low number of studies, particularly low number of statistically significant experiments and the presence of heterogeneity (in the first plot), the relevance of executing a publication bias analysis is severely undercut. Nonetheless, this tool allows for a visual analysis of studies given their effect size and measure of precision, as well as a calculation of the effect size limits and the shifts these undergo upon changes to the number of studies (outlier removal and trim-and-fill method), adding one more layer to the analyses executed previously in Section 5.3.1.

5.3.5 High Numeracy

Since all included experiments explored comprehension in low and high numerates we decided to collect data for both, envisioning minor analyses beyond the subgroup in focus. Accordingly, we executed a random-effects model on high numeracy data, which resulted in the plot in Figure 5.10.

The forest plot in Figure 5.10 indicates a slightly better performance in visual aids ($g = -0.12$), although confidence interval overlaps null effect [95 % CI (-0.38 to 0.13)]. Heterogeneity is lower than what was found in the low numeracy group ($Q = 399.36$; $T = 0.4$), with 91.8% of it being assigned to the inconsistency between individual study effects (I^2).

As with the analysis presented in above Section 5.3 we may visually infer that experiment N05 behaves as an outlier, once again skewing overall effect size towards the visual condition. The GOSH analysis - available in Appendix L - confirmed this inference, with a K-means algorithm singling out the mentioned experiment as an outlier. Removing N05 from the analysis results in a shift towards non-visual aids of both the overall effect ($g = 0.02$), as its confidence interval [95 % CI (-0.12 to 0.16)].

From the results presented above we may conclude that, as with individuals with low numeracy skills, those who are high numerates do not seem to have a clear and unequivocal benefit from receiving risk information in either visual or non-visual aids. Unsurprisingly, and as will follow in Section 5.3.6, high numerates do have far better comprehension values in both these formats when compared to low numerates.

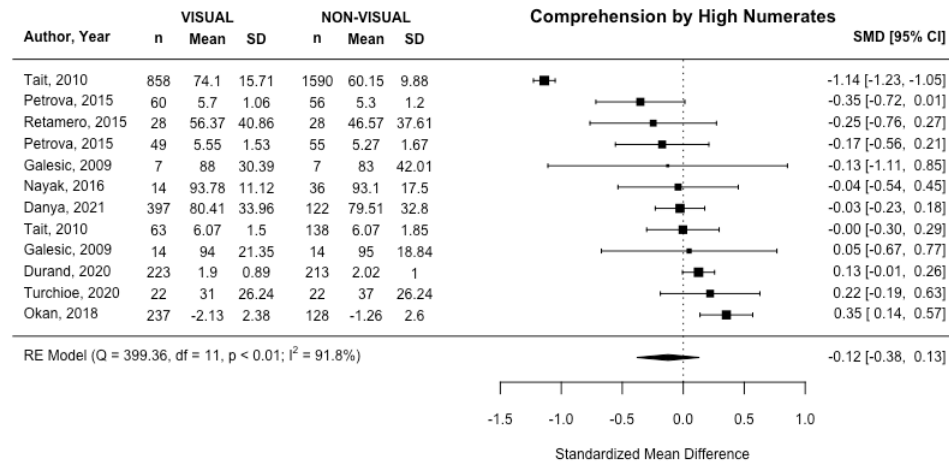


Figure 5.10: Forest plot on the comprehension of visual vs. non-visual aids by high numerates, displaying common and per study effect sizes, heterogeneity analysis and summary statistics. Square size represents individual study weight within the analysis. Ordered by effect size.

5.3.6 High vs. Low Numeracy

As a sort of confirmation of the differences between high and low numerates we decided to pursue one last analysis, opposing these two subgroups regarding the comprehension of either visual or non-visual aids. Resulting forest plots may be examined in below Figure 5.11.

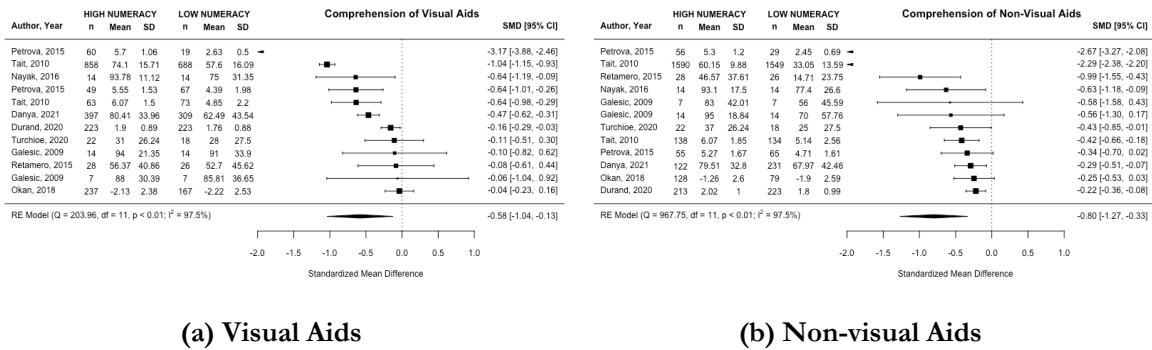


Figure 5.11: Forest plots on the comprehension of (a) visual and (b) non-visual aids by high vs. low numerates, displaying effect sizes, heterogeneity analysis and summary statistics. Square size represents individual study weight within the analysis. Ordered by effect size.

Despite large confidence intervals in some cases, all included experiments report an estimate in favour of high numerates, both in the visual aids as in the non-visual aids condition. As reported, the overall effect size is -0.58 for visual aids and -0.8 for non-visual aids, with confidence

intervals well under the null effect size threshold. Very high heterogeneity was found as both analyses reported an I^2 of 97.5%, despite the visual aids condition reporting a much lower Q , while T remained at 0.77 and 0.79 for visual and non-visual conditions, respectively.

These analyses should attest for the medium to large difference between subgroups when it comes to the comprehension of decision aids and, to some extent, the ability of numeracy tests to predict these discrepancies.

5.4 Discussion

This meta-analysis pursued a synthesis of research on risk communication focused in the comprehension of health decision aids by individuals with low numeracy skills. The 10 included studies were published from 2009 onwards and explored textual, numerical, bar charts, icon arrays and other communication formats. Developing on a question raised by Garcia-Retamero and Cokely (2017), we aggregated these formats into visual and non-visual aids, aspiring to probe how participants of lower abilities performed in each category.

Overall we did not find a significant difference in performance favouring one type of risk communication format. These findings were obtained for both low as high numerates and show how numeracy may not be enough to predict potential disparities in comprehension across formats. Interestingly, in both conditions ten experiments had confidence intervals enclosing null-effect and only two showed significant differences; both towards visual aids in the low numeracy group and one to each category in high numerates. We reckon that, had we the chance to analyse our initially sought interaction, a subgroup of participants with low numeracy as well as low graph literacy would likely exhibit a comprehension score more skewed towards non-visual aids than the results presented above. However, the degree and precision of that overall effect will remain unknown until a systematic review and meta-analysis is executed on the matter.

In our main analysis, virtually all heterogeneity found present was due to the only two significant studies, which, as consequence, were deemed as outliers, while the majority of experiments exhibited quite large confidence intervals. This meant that any chance of executing meaningful meta-regression or subgroup analyses were cut short, and no variable was found moderating overall effect size. Nonetheless, both *Year of Publication* and *% High School or Less* showed interesting results before the removal of outliers and should gather further scrutiny in subsequent studies.

Overall effect size veered from -0.28 to 0.06 through different analyses, while effect size limits went from -0.20 to 0.18. In all situations, confidence intervals overlapped null-effect, further confirming the lack of a clear advantage for any format category. Still, given that different formats were aggregated into categories, polarities may have been mitigated and additional analyses would be of great benefit to understand the potential of specific risk communication formats.

The results generated by the present meta-analysis reinforce the need for more studies on the comprehension of risk communication by more vulnerable individuals and certainly bring up a reasonable doubt over visual aids, their clarity and applicability. Despite visual information being widely preferred over non-visual formats, our research concludes that visual aids may not be better understood and more data is needed in order to find which format fits better to whom.

Unfortunately, we were not able to explore the subgroup with lower skills in both numeracy and graph literacy and we suspect that if future research is executed on this interaction, effect sizes will lean towards non-visual aids. This could be an important step towards better grasping how to communicate clearly and transparently to this particular subgroup of individuals, with the same need as everyone else to undertake so important decisions but with far less tools to do it properly.

Chapter 6

Limitations and Future Work

While both Chapters 3 and 4 featured a section where some of the *a priori* and search strategy/execution weaknesses were addressed, the present Section will provide an outline of the most prominent drawbacks of our analysis as a whole, afterwards concluding with what we think are some of the key aspects future researchers should aim for in studies to come.

6.1 Limitations

More than nine months in since the beginning of this investigation in mid-December 2020, the resulting conclusions are undeniably a consequence of early expectations, faced obstacles and necessary adaptations, and, as such, are not immune to errors, flaws or shortfalls.

Among such pitfalls is, naturally, the way this investigation was initially built. With hindsight, we may now agree on how ambitious were our early goals of analysing the interaction between numeracy and graph literacy. This resulted in the planning and execution of a search strategy that was not oriented towards our final goals, which in turn has most certainly undercut the number of collected articles that could have met our later adapted objective of executing this meta-analysis on numeracy alone. The possibility of having missed some eligible and potentially key articles poses as a major limitation, not only to the solidity of the results but also to the quality of the whole process. Regarding this, we may only suspect that domain knowledge and research expertise would have been paramount for better informed and guided decisions.

Not having access to the participant-level data for the majority of included studies (only 4 experiments made their data available), meant that high and low numeracy groups were defined within each experiment and in accordance to the respective researchers' decisions. Most studies opted for a median split on numeracy scores, which depends directly on the numeracy levels of the given sample. Ultimately this within-study discretization may mean that, overall, participants with similar numeracy skills may have been labeled differently depending on how their score relates to the median score of the sample they are part of. As such, one person may be labeled low numerate in one study and high numerate in another. This adds uncertainty and noise to the data and subsequent results, not allowing for an analysis on standardized low and high numeracy groups.

6.2 Future Work

Having the aforementioned limitations in consideration, it is clear that future research could benefit greatly with some changes and improvements to the way data is reported and shared. If complete data were available for all initially eligible studies, a meta-analysis on the interaction between numeracy and graph literacy would have been possible. Nonetheless, the lack of participant-level data or even necessary summary statistics for the subgroup in question severely limited our use of available research and consequently the quality of our own.

Making data available is not only vital when it comes to transparency but is also of paramount importance to the impact it might have in future research. While in some cases summary statistics might be enough, in others the specific data that was reported might be of no use; e.g. some studies might have collected all the necessary data to be included in a particular meta-analysis but chose not to report these as they were outside of its main scope. This was the case of several collected articles, as these studies tested both numeracy and graph literacy but did not report on their interaction, analysing comprehension against one of these measures or both separately.

The plurality of numeracy and graph literacy tests adds some complexity to the data analysis as there is no clear guidance on how to convert results among tests. More research on the correlation between different tests would therefore help bridge this gap and enable richer comparisons. Also, given the lack of available participant-level data, and the chance that little to no changes will be witnessed on this matter, some standardization would be beneficial for future research. As much as the rationale behind median splits is comprehensible, it has the risk of provoking heterogeneity between results as the thresholds that define low/high numeracy groups are dependant on scores of each sample and thus highly volatile. Having a standard score defining the cut between these two levels (or even a scale of scores) would make results more comprehensible and comparable.

From our perspective, risk communication research appears to be highly concentrated in Health, overlooking other potential areas as Climate, Finance or Public Policies. As expected, research is also focused on developed countries, and even among these they are executed in a specific subset (in our meta-analysis: United States, Japan, Germany and Spain). A higher plurality of nationalities would therefore enable the exploration of potential differences between countries leading to a higher and richer granularity of data.

Additionally, an intensive network meta-analysis could be interesting in order to pinpoint which specific formats should be employed to different numeracy/graph literacy levels of skill. Other interesting analyses might explore format preference or comparisons between gist and verbatim comprehension.

Finally, we understand that more investigations should be focused in individuals with low numeracy and graph literacy levels. As pointed by Garcia-Retamero and Cokely (2017) - which acted as a starting point for our research -, this particular subgroup appears to be the most at risk and in higher need for solutions, as they face unequal medical and health challenges and exhibit much lower comprehension of decision aids when compared to other individuals. Thus, we believe future research should be focused in tackling these disparities by assessing which techniques, formats or elements should be employed to improve decision-making in this vulnerable population.

Bibliography

The articles marked with an asterisk indicate studies included in the meta-analysis.

- Abelson, R. (1997). A retrospective on the significance test ban of 1999 (if there were no significance tests, they would be invented). In L. Harlow, S. Mulaik, & J. Steiger (Eds.), *What if there were no significance tests?* Mahwah, NJ: Lawrence Erlbaum Associates.
- Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: a systematic review [Journal Article]. *J Am Med Inform Assoc*, 13(6), 608-18. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/16929039> doi: 10.1197/jamia.M2115
- Anderson, B. L., & Schulkin, J. (2014). Physicians' understanding and use of numeric information [Book Section]. In J. Anderson Britta L.Schulkin (Ed.), *Numerical reasoning in judgments and decision making about health* (p. 59-79). Cambridge: Cambridge University Press. Retrieved from <https://www.cambridge.org/core/books/numerical-reasoning-in-judgments-and-decision-making-about-health/physicians-understanding-and-use-of-numeric-information/F43F17FF804A76302D108047C66F9CE8> doi: 10.1017/cbo9781139644358.004
- Aromataris, E., & Munn, Z. (2020). *Jbi manual for evidence synthesis* [Edited Book]. JBI. Retrieved from Availablefrom<https://synthesismanual.jbi.global> doi: <https://doi.org/10.46658/JBIMES-20-01>
- Aromataris, E., & Riitano, D. (2014). Constructing a search strategy and searching for evidence. a guide to the literature search for a systematic review [Journal Article]. *Am J Nurs*, 114(5), 49-56. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/24759479> doi: 10.1097/01.NAJ.0000446779.99522.f6
- Baccolini, V., Rosso, A., Di Paolo, C., Isonne, C., Salerno, C., Migliara, G., ... Romano, F. (2021). What is the prevalence of low health literacy in european union member states? a systematic review and meta-analysis [Journal Article]. *Journal of General Internal Medicine*, 9. Retrieved from <GotoISI>://WOS:000605107800003 doi: 10.1007/s11606-020-06407-8
- Bakdash, J. Z., & Marusich, L. R. (2017). Repeated measures correlation [Journal Article]. *Frontiers in psychology*, 8, 456-456. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/28439244https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5383908/> doi: 10.3389/fpsyg.2017.00456
- Berkman, N. D., Sheridan, S. L., Donahue, K. E., Halpern, D. J., & Crotty, K. (2011). Low health literacy and health outcomes: An updated systematic review [Journal Article]. *Annals*

- of *Internal Medicine*, 155(2), 97-+. Retrieved from <GotoISI>://WOS:000292822000016
doi: 10.7326/0003-4819-155-2-201107190-00005
- Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2009). *Introduction to meta-analysis* [Book]. Chichester: Wiley.
- Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis [Journal Article]. *Res Synth Methods*, 1(2), 97-111. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/26061376>
doi: 10.1002/jrsm.12
- Brewer, N. T., Chapman, G. B., Gibbons, F. X., Gerrard, M., McCaul, K. D., & Weinstein, N. D. (2007). Meta-analysis of the relationship between risk perception and health behavior: the example of vaccination [Journal Article]. *Health Psychol*, 26(2), 136-45. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/17385964> doi: 10.1037/0278-6133.26.2.136
- Brick, C., McDowell, M., & Freeman, A. L. J. (2020). Risk communication in tables versus text: a registered report randomized trial on 'fact boxes' [Journal Article]. *R Soc Open Sci*, 7(3), 190876. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/32269779> doi: 10.1098/rsos.190876
- Budescu, D. V., Por, H.-H., & Broomell, S. B. (2011). Effective communication of uncertainty in the ipcc reports [Journal Article]. *Climatic Change*, 113(2), 181-200. Retrieved from <https://doi.org/10.1007/s10584-011-0330-3><https://link.springer.com/article/10.1007/s10584-011-0330-3> doi: 10.1007/s10584-011-0330-3
- Budescu, D. V., Por, H.-H., Broomell, S. B., & Smithson, M. (2014). The interpretation of ipcc probabilistic statements around the world [Journal Article]. *Nature Climate Change*, 4(6), 508-512. Retrieved from <https://doi.org/10.1038/nclimate2194><https://www.nature.com/articles/nclimate2194> doi: 10.1038/nclimate2194
- Chaiken, S., & Trope, Y. (1999). *Dual-process theories in social psychology* [Book]. New York, NY, US: The Guilford Press.
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences (2nd ed.)* [book]. Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.
- Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The berlin numeracy test [Journal Article]. *Judgment and Decision Making*, 7(1), 25-47.
- Cumming, G. (2013). The new statistics: Why and how [Journal Article]. *Psychological Science*, 25(1), 7-29. Retrieved from <https://doi.org/10.1177/0956797613504966> doi: 10.1177/0956797613504966
- Damasio, A. R. (1994). *Descartes' error: Emotion, reason, and the human brain* [Book]. Putnam.
- *Danya, H., Yonekura, Y., & Nakayama, K. (2021). Effects of graphic presentation on understanding medical risks and benefits among japanese adults [Journal Article]. *Cogent Medicine*, 8, 1907894. doi: 10.1080/2331205X.2021.1907894
- Davids, S. L., Schapira, M. M., McAuliffe, T. L., & Nattinger, A. B. (2004). Predictors of pessimistic breast cancer risk perceptions in a primary care population [Journal Article]. *Journal of General Internal Medicine*, 19(4), 310-315. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1525-1497.2004.20801.x> doi: <https://doi.org/10.1111/j.1525-1497.2004.20801.x>

- Denes-Raj, V., & Epstein, S. (1994). Conflict between intuitive and rational processing: When people behave against their better judgment [Journal Article]. *Journal of Personality and Social Psychology*, 66(5), 819-829. doi: 10.1037/0022-3514.66.5.819
- *Durand, M.-A., Yen, R. W., O'Malley, J., Elwyn, G., & Mancini, J. (2020). Graph literacy matters: Examining the association between graph literacy, health literacy, and numeracy in a medicaid eligible population [Journal Article]. *PloS one*, 15(11), e0241844-e0241844. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/33175891><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7657552/> doi: 10.1371/journal.pone.0241844
- Egger, M., Smith, G. D., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test [Journal Article]. *BMJ*, 315(7109), 629-634. doi: 10.1136/bmj.315.7109.629
- Epstein, S. (1994). Integration of the cognitive and the psychodynamic unconscious [Journal Article]. *American Psychologist*, 49(8), 709-724. doi: 10.1037/0003-066X.49.8.709
- Fagerlin, A., Wang, C., & Ubel, P. A. (2005). Reducing the influence of anecdotal reasoning on people's health care decisions: is a picture worth a thousand statistics? [Journal Article]. *Med Decis Making*, 25(4), 398-405. doi: 10.1177/0272989X05278931
- Fagerlin, A., Zikmund-Fisher, B. J., & Ubel, P. A. (2011). Helping patients decide: ten steps to better risk communication [Journal Article]. *J Natl Cancer Inst*, 103(19), 1436-43. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/21931068> doi: 10.1093/jnci/djr318
- Fagerlin, A., Zikmund-Fisher, B. J., Ubel, P. A., Jankovic, A., Derry, H. A., & Smith, D. M. (2007). Measuring numeracy without a math test: development of the subjective numeracy scale [Journal Article]. *Med Decis Making*, 27(5), 672-80. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/17641137><https://journals.sagepub.com/doi/pdf/10.1177/0272989X07304449> doi: 10.1177/0272989X07304449
- Fischhoff, B. (2017). Breaking ground for psychological science: The u.s. food and drug administration [Journal Article]. *Am Psychol*, 72(2), 118-125. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/28221064> doi: 10.1037/a0040438
- Fisher, Z., Tipton, E., & Chipeng, H. (2017). *robumeta: Robust variance meta-regression* [Journal Article]. Retrieved from <https://CRAN.R-project.org/package=robumeta>
- Fleiss, J. L. (1981). *Statistical methods for rates and proportions. second edition* [Book]. Wiley, John and Sons, Inc., New York, N.Y.
- Freeman, A. L. J. (2019). How to communicate evidence to patients [Journal Article]. *Drug Ther Bull*, 57(8), 119-124. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/31345957> doi: 10.1136/dtb.2019.000008
- Fu, R., Gartlehner, G., Grant, M., Shamliyan, T., Sedrakyan, A., Wilt, T. J., ... Trikalinos, T. A. (2011). Conducting quantitative synthesis when comparing medical interventions: Ahrq and the effective health care program [Journal Article]. *J Clin Epidemiol*, 64(11), 1187-97. doi: 10.1016/j.jclinepi.2010.08.010
- Furedi, A. (1999). The public health implications of the 1995 'pill scare' [Journal Article]. *Hum Reprod Update*, 5(6), 621-6. doi: 10.1093/humupd/5.6.621
- Gaissmaier, W., Wegwarth, O., Skopec, D., Muller, A. S., Broschinski, S., & Politi, M. C. (2012). Numbers can be worth a thousand pictures: individual differences in understanding graphical and numerical representations of health-related information [Journal Article]. *Health*

- Psychol*, 31(3), 286-96. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/21842998> doi: 10.1037/a0024850
- Galesic, M., & Garcia-Retamero, R. (2010). Statistical numeracy for health: a cross-cultural comparison with probabilistic national samples [Journal Article]. *Arch Intern Med*, 170(5), 462-8. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20212183> doi: 10.1001/archinternmed.2009.481
- Galesic, M., & Garcia-Retamero, R. (2011a). Do low-numeracy people avoid shared decision making? [Journal Article]. *Health Psychol*, 30(3), 336-41. doi: 10.1037/a0022723
- Galesic, M., & Garcia-Retamero, R. (2011b). Graph literacy: a cross-cultural comparison [Journal Article]. *Med Decis Making*, 31(3), 444-57. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20671213> doi: 10.1177/0272989X10373805
- Galesic, M., & Garcia-Retamero, R. (2013). Using analogies to communicate information about health risks [Journal Article]. *Applied Cognitive Psychology*, 27(1), 33-42. Retrieved from <https://onlinelibrary.wiley.com/doi/abs/10.1002/acp.2866> doi: 10.1002/acp.2866
- *Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: overcoming low numeracy [Journal Article]. *Health Psychol*, 28(2), 210-6. doi: 10.1037/a0014474
- Garcia-Retamero, R., Andrade, A., Sharit, J., & Ruiz, J. G. (2015). Is patients' numeracy related to physical and mental health? [Journal Article]. *Medical Decision Making*, 35(4), 501-511. Retrieved from <https://journals.sagepub.com/doi/abs/10.1177/0272989X15578126> doi: 10.1177/0272989X15578126
- Garcia-Retamero, R., & Cokely, E. T. (2013). Communicating health risks with visual aids [Journal Article]. *Current Directions in Psychological Science*, 22(5), 392-399. Retrieved from <https://journals.sagepub.com/doi/abs/10.1177/0963721413491570> doi: 10.1177/0963721413491570
- Garcia-Retamero, R., & Cokely, E. T. (2017). Designing visual aids that promote risk literacy: A systematic review of health research and evidence-based design heuristics [Journal Article]. *Hum Factors*, 59(4), 582-627. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/28192674> doi: 10.1177/0018720817690634
- Garcia-Retamero, R., Cokely, E. T., Ghazal, S., & Joeris, A. (2016). Measuring graph literacy without a test: A brief subjective assessment [Journal Article]. *Med Decis Making*, 36(7), 854-67. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/27353824><https://journals.sagepub.com/doi/pdf/10.1177/0272989X16655334> doi: 10.1177/0272989X16655334
- Garcia-Retamero, R., & Galesic, M. (2010a). How to reduce the effect of framing on messages about health [Journal Article]. *Journal of General Internal Medicine*, 25(12), 1323-1329. Retrieved from <https://doi.org/10.1007/s11606-010-1484-9> doi: 10.1007/s11606-010-1484-9
- Garcia-Retamero, R., & Galesic, M. (2010b). Who profits from visual aids: overcoming challenges in people's understanding of risks [corrected] [Journal Article]. *Soc Sci Med*, 70(7), 1019-25. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20116159> doi: 10.1016/j.socscimed.2009.11.031
- Garcia-Retamero, R., Galesic, M., & Gigerenzer, G. (2010). Do icon arrays help reduce denominator neglect? [Journal Article]. *Medical Decision Making*, 30(6), 672-684. Re-

- trieved from <https://journals.sagepub.com/doi/abs/10.1177/0272989X10369000> doi: 10.1177/0272989X10369000
- Garcia-Retamero, R., Okan, Y., & Cokely, E. T. (2012). Using visual aids to improve communication of risks about health: a review [Journal Article]. *ScientificWorldJournal*, 2012, 562637. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/22629146> doi: 10.1100/2012/562637
- Garcia-Retamero, R., Wicki, B., Cokely, E. T., & Hanson, B. (2014). Factors predicting surgeons' preferred and actual roles in interactions with their patients [Journal Article]. *Health Psychology*, 33(8), 920-928. doi: 10.1037/hea0000061
- Garcia-Retamero, R., & Cokely, E. T. (2011). Effective communication of risks to young adults: using message framing and visual aids to increase condom use and std screening [Journal Article]. *J Exp Psychol Appl*, 17(3), 270-87. doi: 10.1037/a0023677
- Garcia-Retamero, R., & Cokely, E. (2014). The influence of skills, message frame, and visual aids on prevention of sexually transmitted diseases [Journal Article]. *Journal of Behavioral Decision Making*, 27. doi: 10.1002/bdm.1797
- Garcia-Retamero, R., & Cokely, E. T. (2015). Simple but powerful health messages for increasing condom use in young adults [Journal Article]. *J Sex Res*, 52(1), 30-42. doi: 10.1080/00224499.2013.806647
- *Garcia-Retamero, R., Cokely, E. T., & Hoffrage, U. (2015). Visual aids improve diagnostic inferences and metacognitive judgment calibration [Journal Article]. *Frontiers in Psychology*, 6(932). Retrieved from <https://www.frontiersin.org/article/10.3389/fpsyg.2015.00932> doi: 10.3389/fpsyg.2015.00932
- Gardner, P. H., McMillan, B., Raynor, D. K., Woolf, E., & Knapp, P. (2011). The effect of numeracy on the comprehension of information about medicines in users of a patient information website [Journal Article]. *Patient Educ Couns*, 83(3), 398-403. doi: 10.1016/j.pec.2011.05.006
- Gurmankin, A. D., Baron, J., & Armstrong, K. (2004). The effect of numerical statements of risk on trust and comfort with hypothetical physician risk communication [Journal Article]. *Medical Decision Making*, 24(3), 265-271. doi: 10.1177/0272989X04265482
- Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2019). dmetar: Companion r package for the guide 'doing meta-analysis in r' [Computer software manual]. Retrieved from <http://dmetar.protectlab.org/> (R package version 0.0.9000)
- Hasselblad, V., & Hedges, L. V. (1995). Meta-analysis of screening and diagnostic tests [Journal Article]. *Psychol Bull*, 117(1), 167-78. doi: 10.1037/0033-2909.117.1.167
- Hawley, S. T., Zikmund-Fisher, B., Ubel, P., Jancovic, A., Lucas, T., & Fagerlin, A. (2008). The impact of the format of graphical presentation on health-related knowledge and treatment choices [Journal Article]. *Patient Educ Couns*, 73(3), 448-55. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/18755566> doi: 10.1016/j.pec.2008.07.023
- Heckler, A., Mikula, B., & Rosenblatt, R. (2013). *Student accuracy in reading logarithmic plots: The problem and how to fix it* [Book]. doi: 10.1109/FIE.2013.6684990
- Hess, R., Visschers, V. H., & Siegrist, M. (2011). Risk communication with pictographs: The role of numeracy and graph processing [Journal Article]. *Judgment and Decision Making*, 6(3), 263-274.

- Higgins, J. P., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analyses [Journal Article]. *Bmj*, 327(7414), 557-60. doi: 10.1136/bmj.327.7414.557
- Higgins, J. P. T., Thomas, J., Chandler, J., Cumpston, M., Li, T., Page, M. J., & Welch, V. A. (2019). *Cochrane handbook for systematic reviews of interventions* (2nd Edition ed.) [Book]. Chichester (UK): John Wiley & Sons. Retrieved from www.training.cochrane.org/handbook
- Jahn, N. (2020). europepmc: R interface to the europe pubmed central restful web service [Computer software manual]. Retrieved from <https://CRAN.R-project.org/package=europepmc> (R package version 0.4)
- Kreuzmair, C., Siegrist, M., & Keller, C. (2016). High numerates count icons and low numerates process large areas in pictographs: Results of an eye-tracking study [Journal Article]. *Risk Anal*, 36(8), 1599-614. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/26800006> doi: 10.1111/risa.12531
- Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data [Journal Article]. *Biometrics*, 33(1), 159-174. Retrieved from <http://www.jstor.org/stable/2529310> doi: 10.2307/2529310
- Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: suggested best practices and future recommendations [Journal Article]. *Med Decis Making*, 27(5), 696-713. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/17873259> doi: 10.1177/0272989X07307271
- Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples [Journal Article]. *Med Decis Making*, 21(1), 37-44. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/11206945> doi: 10.1177/0272989X0102100105
- Manganello, J. A., & Clayman, M. L. (2011). The association of understanding of medical statistics with health information seeking and health provider interaction in a national sample of young adults [Journal Article]. *J Health Commun*, 16 Suppl 3, 163-76. doi: 10.1080/10810730.2011.604704
- McDowell, M., Gigerenzer, G., Wegwarth, O., & Rebitschek, F. G. (2019). Effect of tabular and icon fact box formats on comprehension of benefits and harms of prostate cancer screening: A randomized trial [Journal Article]. *Med Decis Making*, 39(1), 41-56. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/30799691> doi: 10.1177/0272989X18818166
- McHugh, M. L. (2012). Interrater reliability: the kappa statistic [Journal Article]. *Biochemia medica*, 22(3), 276-282. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/23092060https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3900052/>
- McNeil, B. J., Pauker, S. G., Sox, J., H. C., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies [Journal Article]. *N Engl J Med*, 306(21), 1259-62. doi: 10.1056/nejm198205273062103
- Menge, D. N. L., MacPherson, A. C., Bytnerowicz, T. A., Quebbeman, A. W., Schwartz, N. B., Taylor, B. N., & Wolf, A. A. (2018). Logarithmic scales in ecological data presentation may cause misinterpretation [Journal Article]. *Nature Ecology & Evolution*, 2(9), 1393-1402. Retrieved from <https://doi.org/10.1038/s41559-018-0610-7> doi: 10.1038/s41559-018-0610-7
- Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G. (2009). Preferred reporting items for

- systematic reviews and meta-analyses: the prisma statement [Journal Article]. *BMJ*, 339, b2535. Retrieved from <https://www.bmj.com/content/bmj/339/bmj.b2535.full.pdf> doi: 10.1136/bmj.b2535
- *Nayak, J. G., Hartzler, A. L., Macleod, L. C., Izard, J. P., Dalkin, B. M., & Gore, J. L. (2016). Relevance of graph literacy in the development of patient-centered communication tools [Journal Article]. *Patient Education and Counseling*, 99(3), 448-454. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0738399115300756> doi: <https://doi.org/10.1016/j.pec.2015.09.009>
- Noar, S. M., Benac, C. N., & Harris, M. S. (2007). Does tailoring matter? meta-analytic review of tailored print health behavior change interventions [Journal Article]. *Psychol Bull*, 133(4), 673-93. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/17592961> doi: 10.1037/0033-2909.133.4.673
- Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2015). Improving risk understanding across ability levels: Encouraging active processing with dynamic icon arrays [Journal Article]. *Journal of Experimental Psychology: Applied*, 21(2), 178-194. doi: 10.1037/xap0000045
- *Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2018). Biasing and debiasing health decisions with bar graphs: Costs and benefits of graph literacy [Journal Article]. *Quarterly Journal of Experimental Psychology*, 71(12), 2506-2519. Retrieved from <https://doi.org/10.1177/1747021817744546> doi: 10.1177/1747021817744546
- Olkin, I., Dahabreh, I. J., & Trikalinos, T. A. (2012). Gosh - a graphical display of study heterogeneity [Journal Article]. *Res Synth Methods*, 3(3), 214-23. doi: 10.1002/jrsm.1053
- Ouzzani, M., Hammady, H., Fedorowicz, Z., & Elmagarmid, A. (2016). Rayyan—a web and mobile app for systematic reviews. *Systematic Reviews*, 5(1), 210. Retrieved from <http://dx.doi.org/10.1186/s13643-016-0384-4> doi: 10.1186/s13643-016-0384-4
- Paasche-Orlow, M. K., & Wolf, M. S. (2007). The causal pathways linking health literacy to health outcomes [Journal Article]. *Am J Health Behav*, 31 Suppl 1, S19-26. doi: 10.5555/ajhb.2007.31.supp.S19
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... Moher, D. (2021). The prisma 2020 statement: an updated guideline for reporting systematic reviews [Journal Article]. *BMJ*, 372, n71. Retrieved from <https://www.bmj.com/content/bmj/372/bmj.n71.full.pdf> doi: 10.1136/bmj.n71
- Peters, E. (2008). Numeracy and the perception and communication of risk [Journal Article]. *Ann N Y Acad Sci*, 1128, 1-7. doi: 10.1196/annals.1399.001
- Peters, E., Dieckmann, N., Dixon, A., Hibbard, J. H., & Mertz, C. K. (2007). Less is more in presenting quality information to consumers [Journal Article]. *Med Care Res Rev*, 64(2), 169-90. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/17406019> doi: 10.1177/10775587070640020301
- Peters, E., Meilleur, L. R., & Tompkins, M. K. (2014). Numeracy and the affordable care act: Opportunities and challenges [Conference Proceedings].
- Peters, E., Vastfjall, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making [Journal Article]. *Psychol Sci*, 17(5), 407-13. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/16683928> doi: 10.1111/j.1467-9280.2006.01720.x

- Petrova, D., Garcia-Retamero, R., Catena, A., Cokely, E., Heredia Carrasco, A., Arrebola Moreno, A., & Ramírez Hernández, J. A. (2017). Numeracy predicts risk of pre-hospital decision delay: a retrospective study of acute coronary syndrome survival [Journal Article]. *Ann Behav Med*, 51(2), 292-306. doi: 10.1007/s12160-016-9853-1
- *Petrova, D., Garcia-Retamero, R., & Cokely, E. T. (2015). Understanding the harms and benefits of cancer screening: A model of factors that shape informed decision making [Journal Article]. *Med Decis Making*, 35(7), 847-58. doi: 10.1177/0272989x15587676
- Reventlow, S., Hvas, A. C., & Tulinius, C. (2001). “in really great danger...” the concept of risk in general practice [Journal Article]. *Scand J Prim Health Care*, 19(2), 71-5. doi: 10.1080/028134301750235268
- Richter, R., Girolodi, E., Jansen, J., & van der Weijden, T. (2020). A qualitative exploration of clinicians’ strategies to communicate risks to patients in the complex reality of clinical practice [Journal Article]. *PLoS One*, 15(8), e0236751. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/32790675> doi: 10.1371/journal.pone.0236751
- Rohatgi, A. (2020, July, 2020). *Webplotdigitizer* [Web Page]. Retrieved from <https://apps.automeris.io/wpd/>
- Romano, A., Sotis, C., Dominioni, G., & Guidi, S. (2020). The scale of covid-19 graphs affects understanding, attitudes, and policy preferences [Journal Article]. *Health Econ*, 29(11), 1482-1494. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/32844495> doi: 10.1002/hecc.4143
- Rost, L. C. (2020). (Vol. 2020) (Web Page No. 24/01/2020). Retrieved from <https://blog.datawrapper.de/weeklychart-charttypes-2020/>
- Ruiz, J. G., Andrade, A. D., Garcia-Retamero, R., Anam, R., Rodriguez, R., & Sharit, J. (2013). Communicating global cardiovascular risk: are icon arrays better than numerical estimates in improving understanding, recall and perception of risk? [Journal Article]. *Patient Educ Couns*, 93(3), 394-402. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/23916416> doi: 10.1016/j.pec.2013.06.026
- Schapira, M. M., Nattinger, A. B., & McAuliffe, T. L. (2006). The influence of graphic format on breast cancer risk communication [Journal Article]. *J Health Commun*, 11(6), 569-82. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/16950729> doi: 10.1080/10810730600829916
- Schutt, J. (2020). (Web Page No. 24/01/2020). Retrieved from <https://engineering.tableau.com/handling-the-covid-19-traffic-spike-in-tableau-public-54303a59ea47>
- Schwartz, L. M., & Woloshin, S. (2013). The drug facts box: Improving the communication of prescription drug information [Journal Article]. *Proceedings of the National Academy of Sciences*, 110(Supplement 3), 14069-14074. Retrieved from https://www.pnas.org/content/pnas/110/Supplement_3/14069.full.pdf doi: 10.1073/pnas.1214646110
- Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography [Journal Article]. *Ann Intern Med*, 127(11), 966-72. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/9412301> doi: 10.7326/0003-4819-127-11-199712010-00003
- Schwarzer, G., Carpenter, J., & Rücker, G. (2015). *Meta-analysis with r* [Book]. doi: 10.1007/978-3-319-21416-0

- Slooman, S. A. (1996). The empirical case for two systems of reasoning [Journal Article]. *Psychological Bulletin*, 119(1), 3-22. doi: 10.1037/0033-2909.119.1.3
- Slovic, P., Monahan, J., & MacGregor, D. G. (2000). Violence risk assessment and risk communication: The effects of using actual cases, providing instruction, and employing probability versus frequency formats [Journal Article]. *Law and Human Behavior*, 24(3), 271-296. doi: 10.1023/A:1005595519944
- Slovic, P., Peters, E., Finucane, M. L., & Macgregor, D. G. (2005). Affect, risk, and decision making [Journal Article]. *Health Psychol*, 24(4S), S35-40. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/16045417> doi: 10.1037/0278-6133.24.4.S35
- Spiegelhalter, D. (2017). Risk and uncertainty communication [Journal Article]. *Annual Review of Statistics and Its Application*, 4(1), 31-60. Retrieved from <https://www.annualreviews.org/doi/abs/10.1146/annurev-statistics-010814-020148> doi: 10.1146/annurev-statistics-010814-020148
- Spiegelhalter, D. (2019). *The art of statistics : learning from data* [Book].
- Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future [Journal Article]. *Science*, 333(6048), 1393-400. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/21903802> doi: 10.1126/science.1191181
- Sutton, A., Abrams, K. R., Jones, D. R., Sheldon, T., & Song, F. (2000). *Methods for meta-analysis in medical research* [Book].
- Szumilas, M. (2010). Explaining odds ratios [Journal Article]. *Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent*, 19(3), 227-229. Retrieved from <https://pubmed.ncbi.nlm.nih.gov/20842279https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2938757/>
- *Tait, A. R., Voepel-Lewis, T., Zikmund-Fisher, B. J., & Fagerlin, A. (2010a). The effect of format on parents' understanding of the risks and benefits of clinical research: a comparison between text, tables, and graphics [Journal Article]. *J Health Commun*, 15(5), 487-501. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20677054> doi: 10.1080/10810730.2010.492560
- *Tait, A. R., Voepel-Lewis, T., Zikmund-Fisher, B. J., & Fagerlin, A. (2010b). Presenting research risks and benefits to parents: does format matter? [Journal Article]. *Anesth Analg*, 111(3), 718-23. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/20686011> doi: 10.1213/ANE.0b013e3181e8570a
- Thaler, R. H., & Sunstein, C. R. (2008). *Nudge: Improving decisions about health, wealth, and happiness* [Book]. New Haven, CT, US: Yale University Press.
- Tippett, L. H. C. (1925). On the extreme individuals and the range of samples taken from a normal population [Journal Article]. *Biometrika*, 17(3/4), 364-387. Retrieved from <http://www.jstor.org/stable/2332087> doi: 10.2307/2332087
- Trevena, L. J., Davey, H. M., Barratt, A., Butow, P., & Caldwell, P. (2006). A systematic review on communicating with patients about evidence [Journal Article]. *J Eval Clin Pract*, 12(1), 13-23. doi: 10.1111/j.1365-2753.2005.00596.x
- Trevena, L. J., Zikmund-Fisher, B. J., Edwards, A., Gaissmaier, W., Galesic, M., Han, P. K., ... Woloshin, S. (2013). Presenting quantitative information about decision outcomes: a risk communication primer for patient decision aid developers [Journal Article]. *BMC Med*

- Inform Decis Mak*, 13 Suppl 2(Suppl 2), S7. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/24625237> doi: 10.1186/1472-6947-13-S2-S7
- Tufanaru, C., Munn, Z., Stephenson, M., & Aromataris, E. (2015). Fixed or random effects meta-analysis? common methodological issues in systematic reviews of effectiveness [Journal Article]. *Int J Evid Based Healthc*, 13(3), 196-207. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/26355603> doi: 10.1097/XEB.0000000000000065
- Tufte, E. R. (1986). *The visual display of quantitative information* [Book].
- Turchioe, M. R., Myers, A., Isaac, S., Baik, D., Grossman, L. V., Ancker, J. S., & Creber, R. M. (2019). A systematic review of patient-facing visualizations of personal health data [Journal Article]. *Appl Clin Inform*, 10(4), 751-770. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/31597182> doi: 10.1055/s-0039-1697592
- *Turchioe, M. R., Grossman, L. V., Myers, A. C., Baik, D., Goyal, P., & Masterson Creber, R. M. (2020). Visual analogies, not graphs, increase patients' comprehension of changes in their health status [Journal Article]. *J Am Med Inform Assoc*, 27(5), 677-689. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/31999316> doi: 10.1093/jamia/ocz217
- Valentine, J. C., Pigott, T. D., & Rothstein, H. R. (2010). How many studies do you need?: A primer on statistical power for meta-analysis [Journal Article]. *Journal of Educational and Behavioral Statistics*, 35(2), 215-247. Retrieved from <https://doi.org/10.3102/1076998609346961> doi: 10.3102/1076998609346961
- van Weert, J. C. M., Alblas, M. C., van Dijk, L., & Jansen, J. (2021). Preference for and understanding of graphs presenting health risk information. the role of age, health literacy, numeracy and graph literacy [Journal Article]. *Patient Educ Couns*, 104(1), 109-117. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/32727670> doi: 10.1016/j.pec.2020.06.031
- Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package [Journal Article]. *Journal of Statistical Software*, 36(3), 1—48. Retrieved from <https://www.jstatsoft.org/v36/i03/>
- Vischers, V. H., Meertens, R. M., Passchier, W. W., & de Vries, N. N. (2009). Probability information in risk communication: a review of the research literature [Journal Article]. *Risk Anal*, 29(2), 267-87. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/19000070> doi: 10.1111/j.1539-6924.2008.01137.x
- Vromans, R. D., Pauws, S. C., Bol, N., van de Poll-Franse, L. V., & Krahmer, E. J. (2020). Communicating tailored risk information of cancer treatment side effects: Only words or also numbers? [Journal Article]. *BMC Med Inform Decis Mak*, 20(1), 277. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/33109175> doi: 10.1186/s12911-020-01296-7
- Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms [Journal Article]. *Bulletin of the Psychonomic Society*, 31(2), 135-138. Retrieved from <https://doi.org/10.3758/BF03334162> doi: 10.3758/BF03334162
- Walter, S. D., & Yao, X. (2007). Effect sizes can be calculated for studies reporting ranges for outcome variables in systematic reviews [Journal Article]. *J Clin Epidemiol*, 60(8), 849-52. doi: 10.1016/j.jclinepi.2006.11.003
- Weinfurt, K. P., Castel, L. D., Li, Y., Sulmasy, D. P., Balshem, A. M., Benson, r., A. B., ... Meropol, N. J. (2003). The correlation between patient characteristics and expectations of benefit

- from phase i clinical trials [Journal Article]. *Cancer*, 98(1), 166-75. doi: 10.1002/cncr.11483
- Willems, S., Albers, C., & Smeets, I. (2020). Variability in the interpretation of probability phrases used in dutch news articles — a risk for miscommunication [Journal Article]. *Journal of Science Communication*. Retrieved from https://jcom.sissa.it/archive/19/02/JCOM_1902_2020_A03.pdf doi: <https://doi.org/10.22323/2.19020203>
- Woloshin, S., Schwartz, L. M., Moncur, M., Gabriel, S., & Tosteson, A. N. A. (2001). Assessing values for health: Numeracy matters [Journal Article]. *Medical Decision Making*, 21(5), 382-390. Retrieved from <https://journals.sagepub.com/doi/abs/10.1177/0272989X0102100505> doi: 10.1177/0272989x0102100505
- Zikmund-Fisher, B. J., Smith, D. M., Ubel, P. A., & Fagerlin, A. (2007). Validation of the subjective numeracy scale: effects of low numeracy on comprehension of risk communications and utility elicitations [Journal Article]. *Med Decis Making*, 27(5), 663-71. Retrieved from <https://www.ncbi.nlm.nih.gov/pubmed/17652180> doi: 10.1177/0272989X07303824
- Zipkin, D. A., Umscheid, C. A., Keating, N. L., Allen, E., Aung, K., Beyth, R., ... Feldstein, D. A. (2014). Evidence-based risk communication: a systematic review [Journal Article]. *Ann Intern Med*, 161(4), 270-80. doi: 10.7326/m14-0295

Appendix

A Data extraction with plot digitizer

The original plots in Figure A1 were print-screened and inputted as images into WebPlot-Digitizer (Rohatgi, 2020), where data was extracted by identifying and calibrating each vertical axis and selecting all data points. This online tool then calculates the coordinates of each point along the identified axis and outputs the respective values. A matrix was built and uploaded into Datawrapper, where resulting plots were designed and stored¹² (data matrix is available for download).

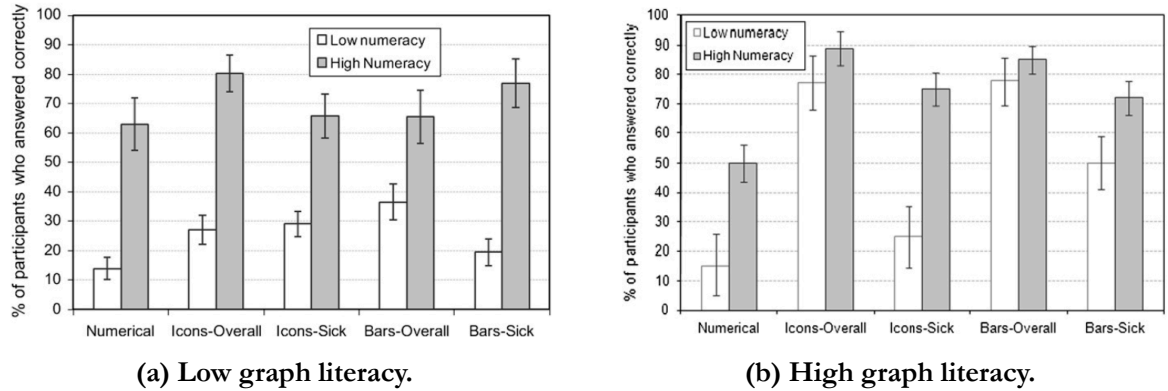


Figure A1: Percentage of participants with low or high graphical literacy and low or high numeracy who correctly estimated the two scenarios (by visual aids formats) in a study by Garcia-Retamero and Galesic (2010b). Error bars represent one standard error. Original plots.

¹www.datawrapper.de/_/3DGSP/

²www.datawrapper.de/_/KFIOg/

B Objective Graph Literacy Test

Figure B1 exhibits part of the graph literacy test developed by Galesic and Garcia-Retamero (2011b), featuring questions for three levels of comprehension: Level 1 - ability to read the data; Level 2 - ability to read between the data; Level 3 - ability to read beyond the data.

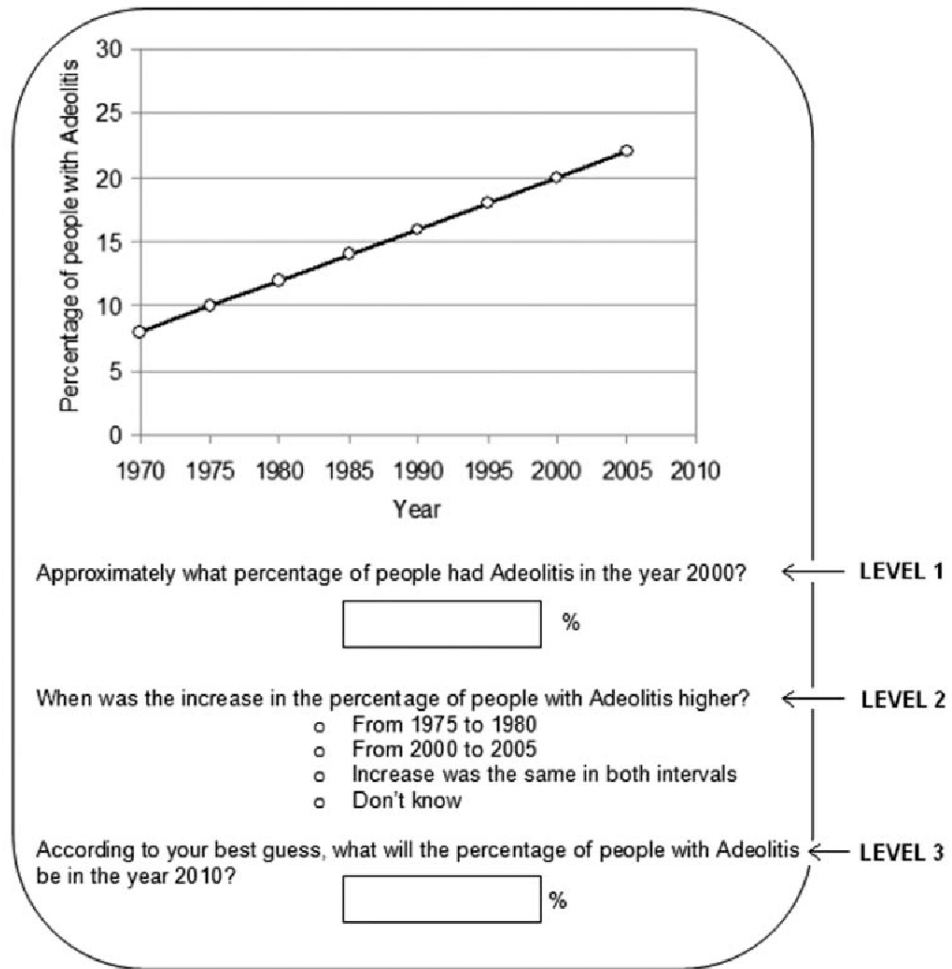


Figure B1: Section of objective graph literacy test developed by Galesic and Garcia-Retamero (2011b) with three levels of graphical comprehension.

C Computer Adaptive Berlin Numeracy Test

Figure C1 shows the structure of the Computer Adaptive Berlin Numeracy Test developed by Cokely et al. (2012).

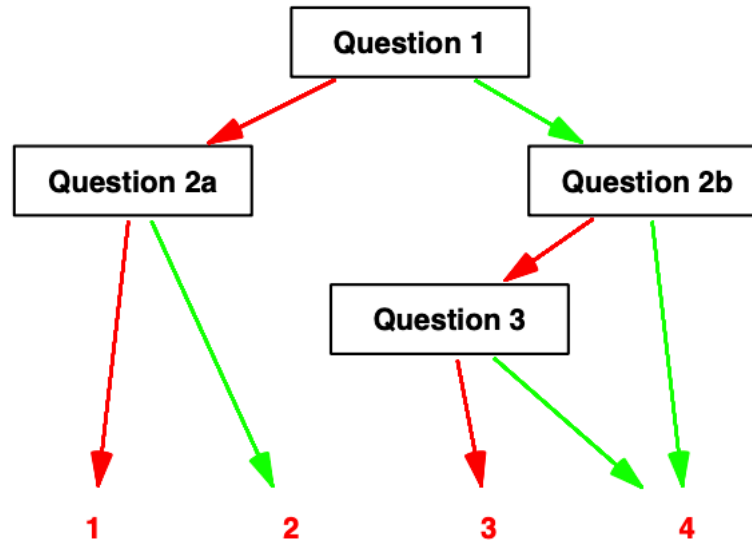


Figure C1: Structure of the Computer Adaptive Berlin Numeracy Test proposed by Cokely et al. (2012). Flow of test and classification of participants is given by green (correct answer) and red (incorrect answer) arrows.

The test questions³, scoring and answers are as follows:

1. Out of 1,000 people in a small town 500 are members of a choir. Out of these 500 members in the choir 100 are men. Out of the 500 inhabitants that are not in the choir 300 are men. What is the probability that a randomly drawn man is a member of the choir? Please indicate the probability in percent. ____%

2a. Imagine we are throwing a five-sided die 50 times. On average, out of these 50 throws how many times would this five-sided die show an odd number (1, 3 or 5)? ____ out of 50 throws.

2b. Imagine we are throwing a loaded die (6 sides). The probability that the die shows a 6 is twice as high as the probability of each of the other numbers. On average, out of these 70 throws how many times would the die show the number 6? ____ out of 70 throws.

3. In a forest 20% of mushrooms are red, 50% brown and 30% white. A red mushroom is poisonous with a probability of 20%. A mushroom that is not red is poisonous with a probability of 5%. What is the probability that a poisonous mushroom in the forest is red? ____

Scoring = Based on answers to 2-3 questions following the adaptive structure.

Correct answers are as follows: 1 = 25; 2a = 30; 2b = 20; 4 = 50.

³Other test versions can be found at www.riskliteracy.org.

D Interpretation of Probability Phrases

Figure D1 displays the distributions of interpreted percentages for different probability phrases in a study by Willems, Albers, and Smeets (2020).

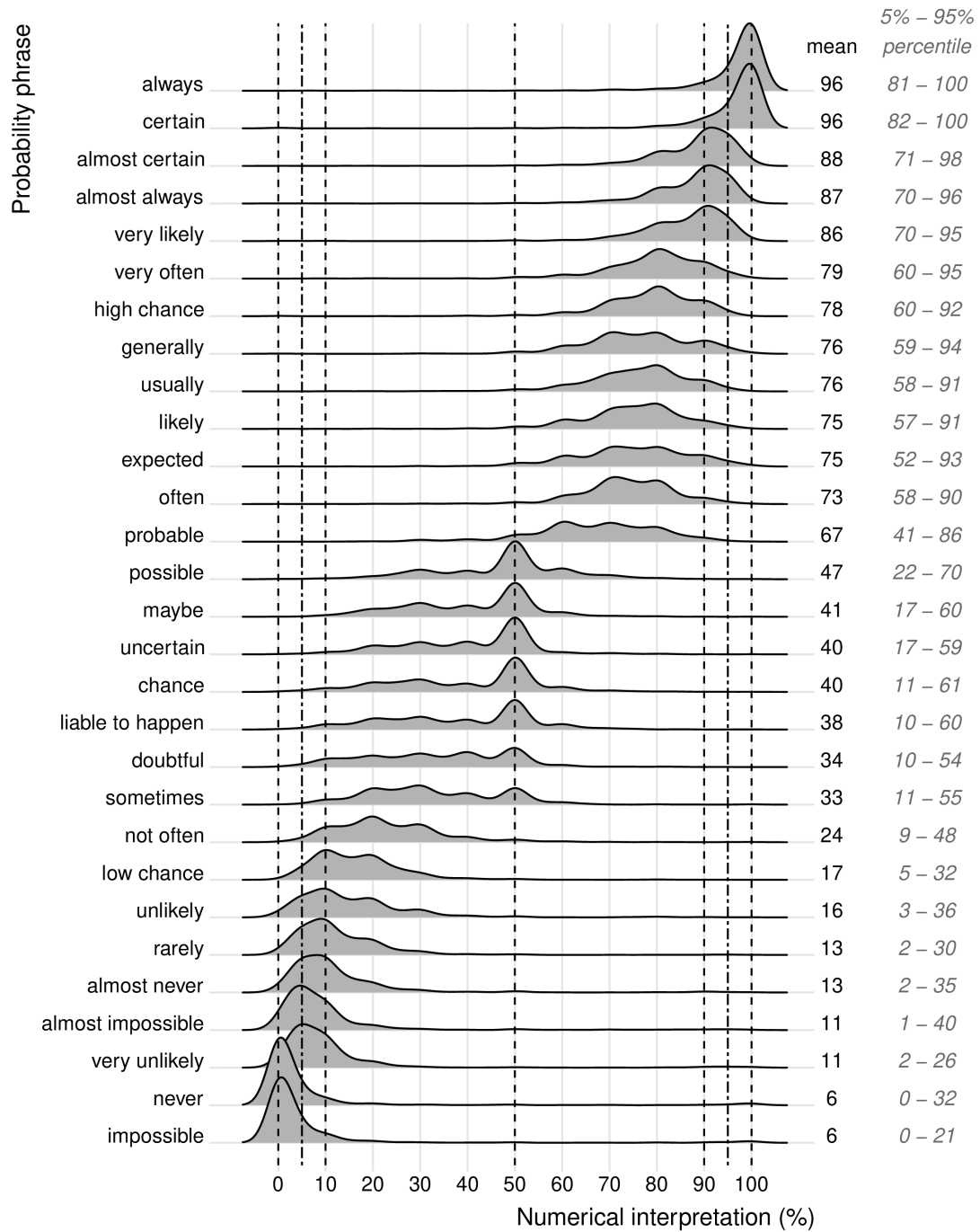


Figure D1: Distributions of interpreted percentages for probability terms tested by Willems et al. (2020).

E PRISMA Flow Diagram

Figure E1 exhibits the flow diagram to address and report the Meta-analysis flow of information.

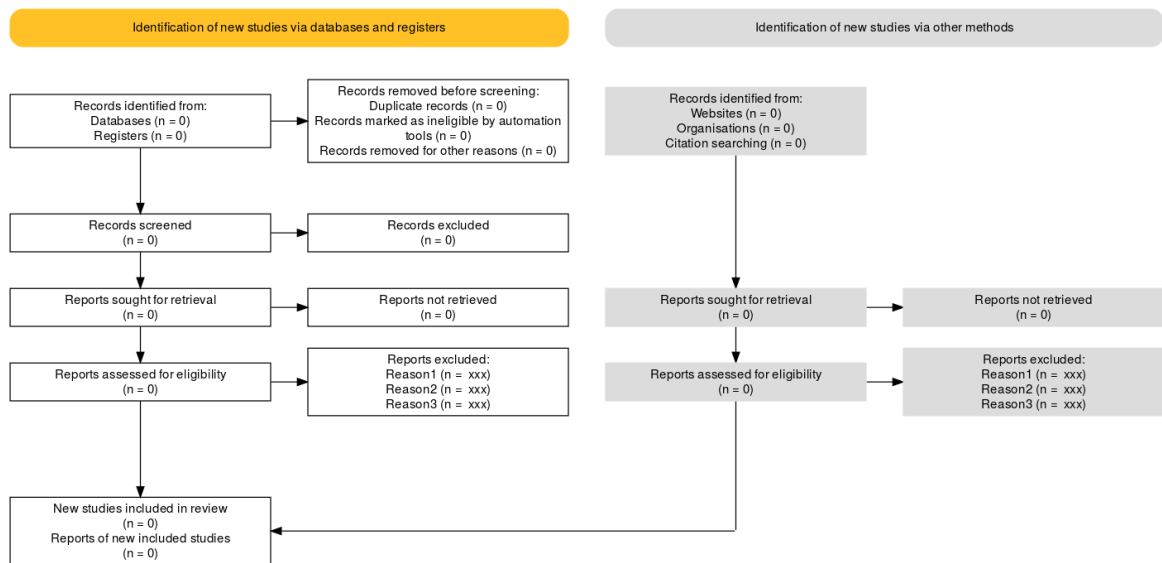


Figure E1: PRISMA 2020 Flow Diagram for depicting flow of information through different steps of the Meta-analysis (Moher et al., 2009).

F PRISMA-P Checklist

Table F1 lists the necessary features of a Protocol according to PRISMA guidelines.

Table F1: PRISMA-P Checklist (Moher et al., 2009).

Section and topic	Item No	Checklist item
ADMINISTRATIVE INFORMATION		
Title:		
Identification	1a	Identify the report as a protocol of a systematic review
Update	1b	If the protocol is for an update of a previous systematic review, identify as such
Registration	2	If registered, provide the name of the registry (such as PROSPERO) and registration number
Authors:		
Contact	3a	Provide name, institutional affiliation, e-mail address of all protocol authors; provide physical mailing address of corresponding author
Contributions	3b	Describe contributions of protocol authors and identify the guarantor of the review
Amendments	4	If the protocol represents an amendment of a previously completed or published protocol, identify as such and list changes; otherwise, state plan for documenting important protocol amendments
Support:		
Sources	5a	Indicate sources of financial or other support for the review
Sponsor	5b	Provide name for the review funder and/or sponsor
Role of sponsor or funder	5c	Describe roles of funder(s), sponsor(s), and/or institution(s), if any, in developing the protocol
INTRODUCTION		
Rationale	6	Describe the rationale for the review in the context of what is already known
Objectives	7	Provide an explicit statement of the question(s) the review will address with reference to participants, interventions, comparators, and outcomes (PICO)
METHODS		
Eligibility criteria	8	Specify the study characteristics (such as PICO, study design, setting, time frame) and report characteristics (such as years considered, language, publication status) to be used as criteria for eligibility for the review
Information sources	9	Describe all intended information sources (such as electronic databases, contact with study authors, trial registers or other grey literature sources) with planned dates of coverage

Continuation of Table F1		
Search strategy	10	Present draft of search strategy to be used for at least one electronic database, including planned limits, such that it could be repeated
Study records:		
Data management	11a	Describe the mechanism(s) that will be used to manage records and data throughout the review
Selection process	11b	State the process that will be used for selecting studies (such as two independent reviewers) through each phase of the review (that is, screening, eligibility and inclusion in meta-analysis)
Data collection process	11c	Describe planned method of extracting data from reports (such as piloting forms, done independently, in duplicate), any processes for obtaining and confirming data from investigators
Data items	12	List and define all variables for which data will be sought (such as PICO items, funding sources), any pre-planned data assumptions and simplifications
Outcomes and prioritization	13	List and define all outcomes for which data will be sought, including prioritization of main and additional outcomes, with rationale
Risk of bias in individual studies	14	Describe anticipated methods for assessing risk of bias of individual studies, including whether this will be done at the outcome or study level, or both; state how this information will be used in data synthesis
Data synthesis	15a	Describe criteria under which study data will be quantitatively synthesised
	15b	If data are appropriate for quantitative synthesis, describe planned summary measures, methods of handling data and methods of combining data from studies, including any planned exploration of consistency (such as I ² , Kendall's τ)
	15c	Describe any proposed additional analyses (such as sensitivity or subgroup analyses, meta-regression)
	15d	If quantitative synthesis is not appropriate, describe the type of summary planned
Meta-bias(es)	16	Specify any planned assessment of meta-bias(es) (such as publication bias across studies, selective reporting within studies)
Confidence in cumulative evidence	17	Describe how the strength of the body of evidence will be assessed (such as GRADE)

G Logic Grids

Logic Grids

Table G1: Logic Grid with initial elements of Review Question.

Population	Intervention
Patients	Visual Decision Aids
Comparison	Outcome
Numerical and verbal formats	Accuracy of risk understanding

Table G2: Logic Grid with inclusion of similar identified keywords.

Population	Intervention
Patients	Visual Decision Aids
Consumer	Visual Aids
User	Decision Aids
Numeracy	Risk Communication
Graph Literacy	Graph
	Visualization
	Health Record
	Test Results
	Icon arrays
	Patient-centered communication
Comparison	Outcome
Numerical and verbal formats	Accuracy of risk understanding
Numerical	Risk Literacy
Verbal	Health Literacy
Numbers	Risk Comprehension
	Decision Making
	Behavior
	Literacy

Table G3: Logic Grid with added Index MeSH terms.

Population	Intervention
Patients	Visual Decision Aids
Consumer	Visual Aids
User	Decision Aids
Numeracy	Risk Communication
Graph Literacy	Graph
	Visualization
	Health Record
	Test Results
	Icon arrays
	Patient-centered communication
	Health Communication
	Health Communication [mh]
	Audiovisual Aids [mh]
	Decision Support Techniques [mh]
	Medical Records [mh]
	Electronic Health Records [mh]
Comparison	Outcome
Numerical and verbal formats	Accuracy of risk understanding
Numerical	Risk Literacy
Verbal	Health Literacy
Numbers	Risk Comprehension
	Decision Making
	Behavior
	Literacy
	Risk understanding
	Health Literacy [mh]
	Patient Medication Knowledge [mh]
	Decision Making [mh]
	Clinical Decision Making [mh]
	Decision Making, Shared [mh]
	Risk Assessment [mh]

Table G4: Logic Grid including field codes and wildcard characters.

Population	Intervention
Patient* [tiab]	Visual Decision Aids [tiab]
Consumer [tiab]	Visual Aids [tiab]
User [tiab]	Decision Aids [tiab]
Numeracy [tiab]	Risk Communication [tiab]
Graph Literacy [tiab]	Graph* [tiab]
	Visual* [tiab]
	Health Record [tiab]
	Test Results [tiab]
	Icon arrays [tiab]
	Patient-centered communication [tiab]
	Health Communication [mh]
	Health Communication [tw]
	Audiovisual Aids [mh]
	Decision Support Techniques [mh]
	Medical Records [mh]
	Eletronic Health Records [mh]
Comparison	Outcome
Numerical and verbal format* [tiab]	Accuracy of risk understanding [tw]
Numerical [tw]	Risk Literacy [tw]
Verbal [tw]	Health Literacy [tw]
Numbers [tiab]	Risk Comprehension [tw]
Numeric* [tiab]	Decision Making [tw]
Verbal Communicat* [tw]	Behavior* [tiab]
	Behaviour* [tiab]
	Literacy [tiab]
	Risk understanding [tw]
	Health Literacy [mh]
	Patient Medication Knowledge [mh]
	Decision Making [mh]
	Clinical Decision Making [mh]
	Decision Making, Shared [mh]
	Risk Assessment [mh]

H Search Strings

PubMed: (((“patient*”[Title/Abstract] OR “Consumer”[Title/Abstract] OR “User”[Title/Abstract] OR “Numeracy”[Title/Abstract] OR “Graph-literacy”[Title/Abstract]) AND (“visual*”[Title/Abstract]) AND ((“visual decision aids”[Title/Abstract] OR “visual aids”[Title/Abstract] OR “decision aids”[Title/Abstract] OR “graph*”[Title/Abstract] OR “visual*”[Title/Abstract] OR “icon array*”[Title/Abstract] OR “patient-centered communication”[Title/Abstract] OR “health communication”[MeSH Terms] OR “health communication”[Title/Abstract] OR “audiovisual aids”[MeSH Terms] OR ((“Health”[MeSH Terms] OR “Health Risks”[Title/Abstract]) AND “records”[MeSH Terms])) AND (“numerical format*”[Text Word] OR “Numerical”[Text Word] OR “Verbal”[Title/Abstract] OR “Numbers”[Title/Abstract] OR “Numeric*”[Title/Abstract] OR “verbal communicat*”[Title/Abstract]) AND (“risk literacy”[Title/Abstract] OR “health literacy”[Title/Abstract] OR “risk comprehension”[Title/Abstract] OR (“Decision”[Title/Abstract] OR “decision s”[Title/Abstract] OR “decisions”[Title/Abstract]) OR “behavior*”[Title/Abstract] OR “Literacy”[Title/Abstract] OR “health literacy”[MeSH Terms] OR “patient medication knowledge”[MeSH Terms] OR “decision making”[MeSH Terms] OR “clinical decision making”[MeSH Terms] OR “decision making, shared”[MeSH Terms] OR “risk assessment”[MeSH Terms] OR “behaviour*”[Title/Abstract] OR “risk understanding”[Title/Abstract]) AND 2010/01/01:2021/12/31[Date - Publication])) AND (“randomized controlled trial”[Publication Type] OR “research support, american recovery and reinvestment act”[Publication Type] OR “research support, n i h , extramural”[Publication Type] OR “research support, n i h , intramural”[Publication Type] OR “research support non u s gov t”[All Fields] OR “research support u s gov t non p h s”[All Fields] OR “research support u s gov t p h s”[All Fields] OR “research support, u s government”[Publication Type]))

Web of Science: ((TS=(Patient* OR Numeracy OR Graph-literacy OR Graph* Literacy) AND TS=(Visual Decision Aids OR Visual Aids OR Decision Aids OR Graph* OR Visual* OR Patient-centered communication OR Health Communication OR Audiovisual Aids) AND TS=(Numerical format* OR Numerical OR Numeric* OR Verbal Communicat*) AND TS=(Risk Literacy OR Health Literacy OR Risk Comprehension OR Health Literacy OR Patient Medication Knowledge OR Decision Making OR Clinical Decision Making) AND MH=(Health Literacy OR Decision Making OR Patient Medication Knowledge OR Clinical Decision Making OR Audiovisual Aids OR Decision Support Techniques))) AND LANGUAGE: (English) AND DOCUMENT TYPES: (Adaptive Clinical Trial OR Clinical Trial OR Comparative Study OR Controlled Clinical Trial OR Equivalence Trial OR Evaluation Studies OR Evaluation Study OR Observational Study OR Randomized Controlled Trial OR Research Support American Recovery And Reinvestment Act OR Research Support N I H Extramural OR Research Support N I H Intramural OR Research Support Non U S Gov T OR Research Support U S Gov T Non P H S OR Research Support U S Gov T P H S OR Validation Studies OR Validation Study) AND SPECIES: (Humans) Indexes=MEDLINE Timespan=2010-2021

Scopus: TITLE-ABS("patient*" OR "consumer*" OR "user*" OR "numeracy" OR "graph-literacy" OR "graph* literacy") AND TITLE-ABS(visual*) AND TITLE-ABS("visual decision aids" OR "visual aids" OR "decision aids" OR "graph*" OR "icon array*" OR "graphical format*" OR "patient-centered communication" OR "health communication" OR "audiovisual aids" OR "health risks") AND TITLE-ABS("numerical format*" OR "Numerical" OR "Numbers" OR "Numeric*" OR "verbal communicat*" OR "Risk communication" OR "health literacy") AND TITLE-ABS("risk literacy" OR "Health Literacy" OR "Risk comprehension" OR "decision" OR "decision making" OR "behaviour*" OR "behavior*" OR "risk understanding") AND KEY("Health*" OR "Data Visualization" OR "Graph literacy" OR "Numeracy" OR "medical decision making" OR "clinical decision making" OR "Risk literacy" OR "Visual Aids" OR "Shared decision making" OR "patient engagement" OR "Graph literacy in health" OR "consumer health information") AND LANGUAGE("English") AND PUBYEAR > 2009

ACM Digital Library: [[Abstract: "patient*"] OR [Abstract: "consumer*"] OR [Abstract: "user*"] OR [Abstract: "numeracy"] OR [Abstract: "graph-literacy"] OR [Abstract: "graph* literacy"]]] AND [Abstract: visual*] AND [[Full Text: "visual decision aids"] OR [Full Text: "visual aids"] OR [Full Text: "decision aids"] OR [Full Text: "graph*"] OR [Full Text: "icon array*"] OR [Full Text: "graphical format*"] OR [Full Text: "patient-centered communication"] OR [Full Text: "health communication"] OR [Full Text: "audiovisual aids"] OR [Full Text: "health risks"]]] AND [[Full Text: "numerical format*"] OR [Full Text: "numerical"] OR [Full Text: "numbers"] OR [Full Text: "numeric*"] OR [Full Text: "verbal communicat*"] OR [Full Text: "risk communication"] OR [Full Text: "health literacy"]]] AND [[Full Text: "risk literacy"] OR [Full Text: "health literacy"] OR [Full Text: "risk comprehension"] OR [Full Text: "decision"] OR [Full Text: "decision making"] OR [Full Text: "behaviour*"] OR [Full Text: "behavior*"] OR [Full Text: "risk understanding"]]] AND [[Full Text: "health*"] OR [Full Text: "data visualization"] OR [Full Text: "graph literacy"] OR [Full Text: "numeracy"] OR [Full Text: "medical decision making"] OR [Full Text: "clinical decision making"] OR [Full Text: "risk literacy"] OR [Full Text: "visual aids"] OR [Full Text: "shared decision making"] OR [Full Text: "patient engagement"] OR [Full Text: "graph literacy in health"] OR [Full Text: "consumer health information"]]] AND [Publication Date: (01/01/2010 TO *)]

IEEE Computational Index: ("Abstract": "patient*" OR "Abstract": "consumer*" OR "Abstract": "user*" OR "Abstract": "numeracy" OR "Abstract": "graph-literacy" OR "Abstract": "graph* literacy") AND ("Abstract": visual*) AND ("Full Text Only": "visual decision aids" OR "Full Text Only": "visual aids" OR "Full Text Only": "decision aids" OR "Full Text Only": "graph*" OR "Full Text Only": "icon array*" OR "Full Text Only": "graphical format*" OR "Full Text Only": "patient-centered communication" OR "Full Text Only": "health communication" OR "Full Text Only": "audiovisual aids" OR "Full Text Only": "health risks") AND ("Full Text Only": "numerical format*" OR "Full Text Only": "Numerical" OR "Full Text Only": "Numbers" OR "Full Text Only": "Numeric*" OR "Full Text Only": "verbal communicat*" OR "Full Text Only": "Risk communication" OR "Full

Text Only”：“health literacy”) AND (“Full Text Only”：“risk literacy” OR “Full Text Only”：“Health Literacy” OR “Full Text Only”：“Risk comprehension” OR “Full Text Only”：“decision” OR “Full Text Only”：“decision making” OR “Full Text Only”：“behaviour*” OR “Full Text Only”：“behavior*” OR “Full Text Only”：“risk understanding”) AND (“Abstract”：“medical” OR “Abstract”：“health”)
Filter: Publication date: 2010-2021

Google Scholar (using Publish or Perish): visual AND numeracy AND health AND patient AND risk AND “graph literacy”

I SD Estimation Conversion Factors

Table I1 contains conversion factors initially proposed by Tippett (1925) in order to estimate a Standard Deviation for a given sample size.

Table I1: Conversion factors for estimating SD for a sample size N . Extracted from Walter and Yao (2007); proposed by Tippett (1925).

Sample size (N)	Conversion factor (f)	Sample size (N)	Conversion factor (f)	Sample size (N)	Conversion factor (f)
2	0.886	13	0.300	50	0.222
3	0.591	14	0.294	60	0.216
4	0.486	15	0.288	70	0.210
5	0.430	16	0.283	80	0.206
6	0.395	17	0.279	90	0.202
7	0.370	18	0.275	100	0.199
8	0.351	19	0.271	150	0.189
9	0.337	20	0.268	200	0.182
10	0.325	25	0.254	300	0.174
11	0.315	30	0.245	500	0.165
12	0.307	40	0.231	1000	0.154

Note. Linear interpolation can be used to calculate SD for any N value between those presented in table.

J Influence Diagnostics

Below Figure J1 exhibits the plot outputted by *metafor* function *influence*, which calculates several metrics following a leave-one-out method. Experiment N05, here labeled as 6, displays deviant values for the following metrics:

- *dfits* - change in standard deviations upon the removal of the *i*th experiment;
- *cook.d* - Mahalanobis distance between predicted values with and without the *i*th experiment;
- *cov.r* - determinant of the covariance matrix without the *i*th experiment divided by the determinant of the covariance matrix with all experiments. A value under 1 signals that removing the *i*th experiment produces more precise estimates;
- *tau2.del* - residual τ^2 heterogeneity left upon the removal of the *i*th experiment;
- *QE.del* - value for the test statistic for residual heterogeneity when the *i*th experiment is removed.

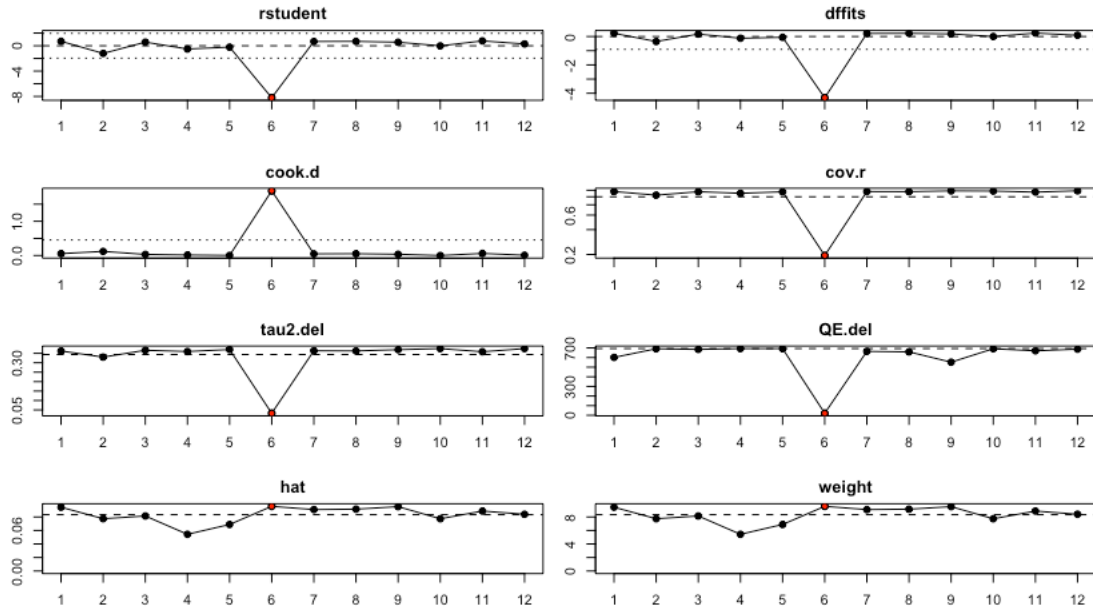


Figure J1: Per study influence in several diagnostic metrics.

K Cumulative Meta-analysis

Figure K1 exhibits a cumulative Forest Plot ordered by Year of Publication. Through its analysis we may denote how the overall effect size tends to gradually shift towards the right as new experiments are added; thus meaning that later research tends to point to non-visual aids as better tools for the comprehension of risk information by low numerates.

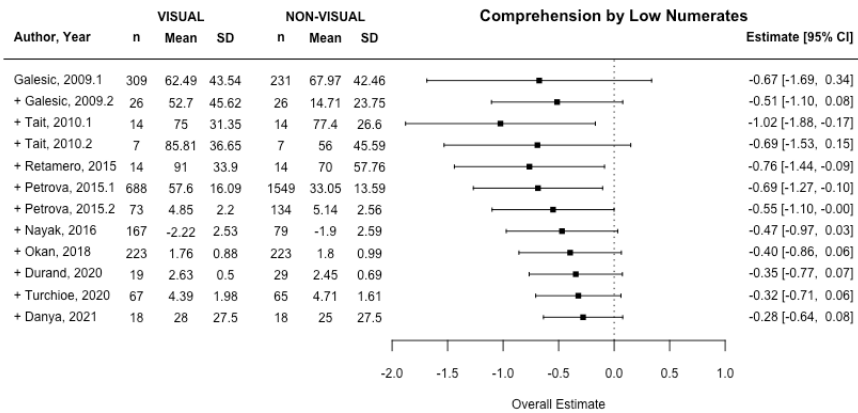


Figure K1: Cumulative forest plot on the comprehension of visual vs. non-visual aids by low numerates by year of publication, displaying common and per study effect sizes, heterogeneity analysis and summary statistics.

Figure K2 displays the evolution of overall effect size and I^2 point estimates throughout the cumulative meta-analysis presented in above Figure K1.

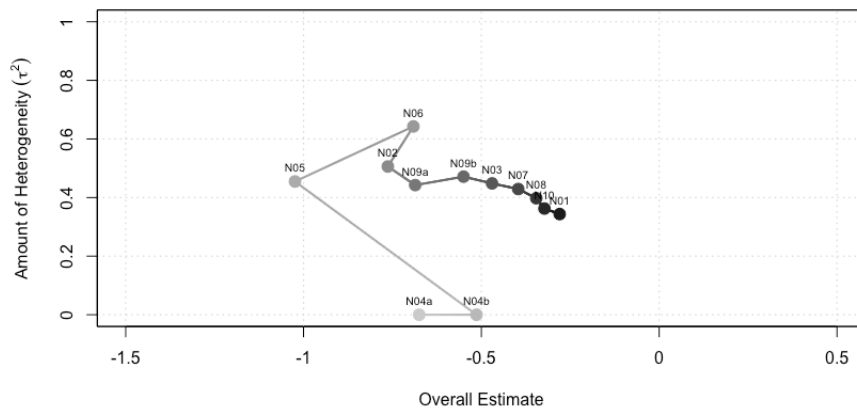


Figure K2: Plot of cumulative results of Figure K1, exhibiting change in overall estimate of effect size and heterogeneity with the sequential introduction of new experiments in the analysis.

L High Numeracy GOSH Analysis

Figures L1 and L2 exhibit GOSH analysis for study N05 and cluster analysis with per study influence in imbalance, respectively.

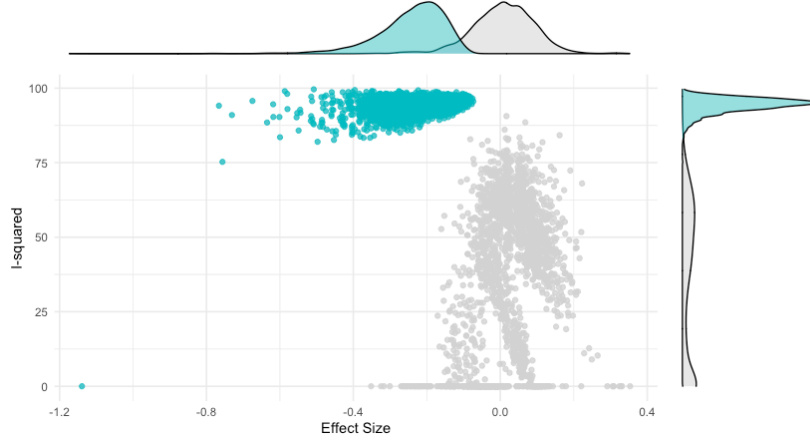


Figure L1: GOSH plot highlighting (in blue) models that included study N05, against models that did not include this study (in gray). Models are plotted according to their I^2 and overall effect size values.

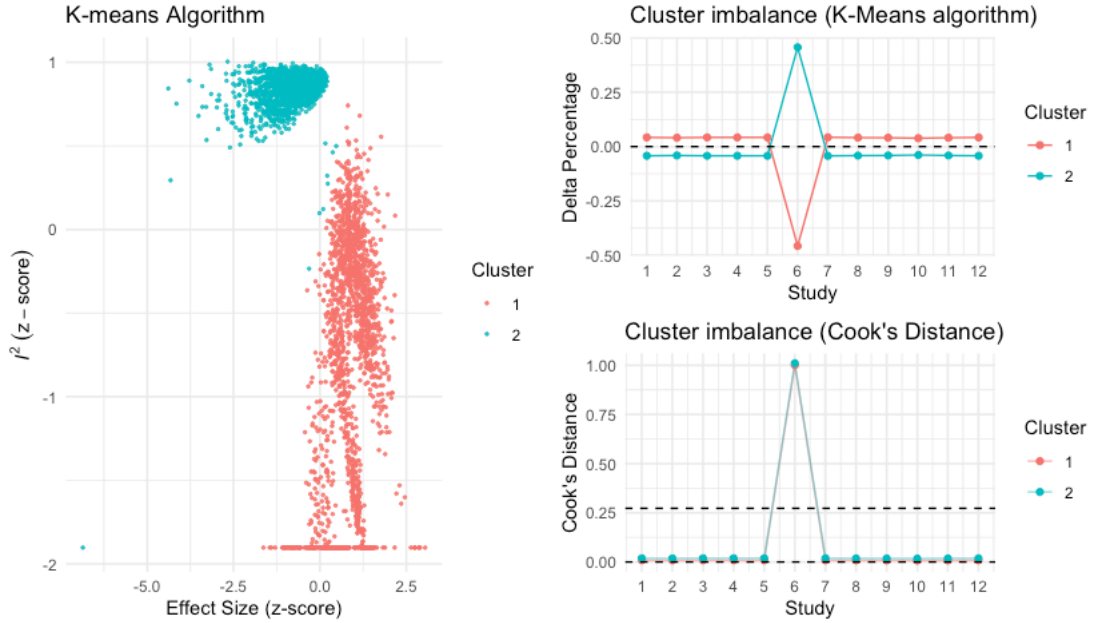


Figure L2: *Left:* GOSH plot mapping the 4095 models against their I^2 heterogeneity and Standardized Mean Difference z-scores (colored by K-means cluster). *Right:* Per-study analysis of influence in K-means cluster imbalance. Experiments numbered 1 through 12 as ordered in Table 4.3.

M Funnel Plot Variations

Figures M1 and M2 exhibit variations on the funnel plot presented in Section 5.3.4, by using standard error and sampling variance as precision measures and centering the funnel around the common found effect size.

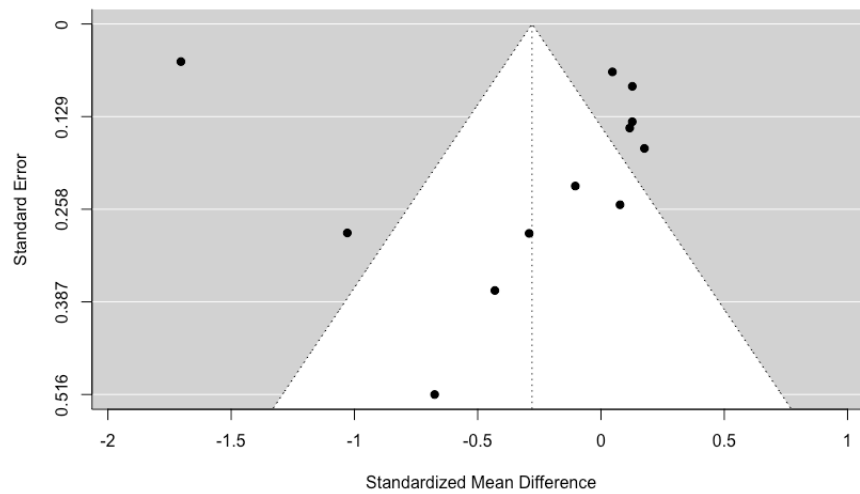


Figure M1: Funnel plot for main analysis using standard error as precision measure.

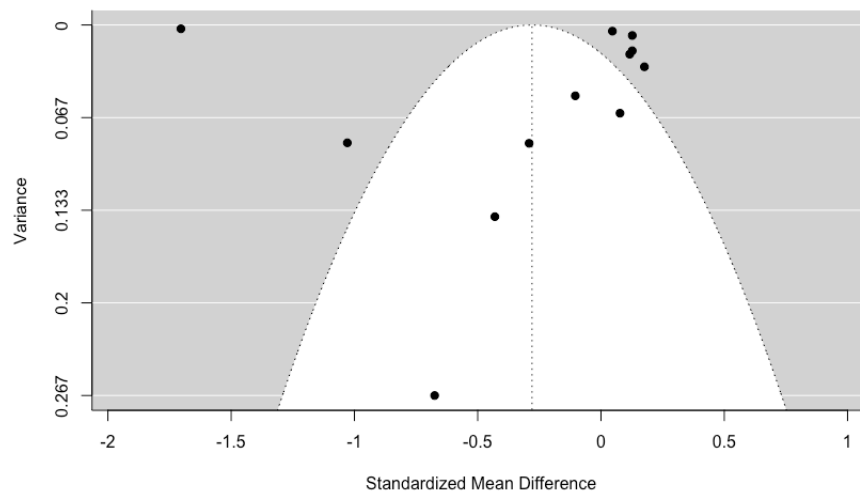


Figure M2: Funnel plot for main analysis using sampling variance as precision measure.