TESIS DOCTORAL

Cómo Mejorar la Comprensión y Comunicación de Información Sobre Riesgos Médicos y de Salud

How to Improve the Comprehension and Communication of Information About Medical and Health Risks

Yasmina Okan Gil

Programa Oficial de Doctorado en Psicología
Universidad de Granada

Departamento de Psicología Experimental
CIMCYC

Editor: Editorial de la Universidad de Granada
Autor: Yasmina Okan Gil
D.L.: GR 912-2014
ISBN: 978-84-9028-948-8

Doctoral Candidate: Yasmina Okan Gil

Thesis Directors

Prof. Rocío García-Retamero Imedio
(University of Granada)

Prof. Mirta Galesic
(Max Planck Institute for Human Development Berlin)

Prof. Antonio Maldonado López
(University of Granada)

External Reviewers

Prof. Edward T. Cokely
(Michigan Technological University)

Prof. Wändi Bruine de Bruin
(University of Leeds)

Thesis Examination Committee

Prof. Encarnación Carmona Samper
(University of Almería)

Prof. Andrés Catena Martínez
(University of Granada)

Prof. Olga Kostopoulou
(King's College London)

Prof. Jose César Perales López
(University of Granada)

Prof. Daniëlle Timmermans
(VU University Amsterdam)

The studies presented in this dissertation were funded by a grant awarded by the Informed Medical Decisions Foundation (United States) to Yasmina Okan (ref. # 0230-1), by the Ministerio de Economía y Competitividad (Spain; ref. # PSI2011-22954, # PSI2008-02019, # PSI2009-12217), and were supported by the Max Planck Society.

INFORMED MEDICAL DECISIONS FOUNDATION
Partnerships for Quality Care

El doctorando …Yasmina Okan Gil… y los directores de la tesis …Rocío García-Retamero Imedio, Mirta Galesic, y Antonio Maldonado López… Garantizamos, al firmar esta tesis doctoral, que el trabajo ha sido realizado por el doctorando bajo la dirección de los directores de la tesis y hasta donde nuestro conocimiento alcanza, en la realización del trabajo, se han respetado los derechos de otros autores a ser citados, cuando se han utilizado sus resultados o publicaciones.

Granada, a 18 de Septiembre de 2013

Director/es de la Tesis

Fdo.:

Doctorando

Fdo.:

Berlin, a 16 de Septiembre de 2013

Director/es de la Tesis

Fdo.:

Acknowledgments (Agradecimientos)

En primer lugar, agradecer a mi directora Rocío García-Retamero Imedio, por haberme dado la oportunidad de llevar a cabo este trabajo. Por mostrarme lo que significa la capacidad de perseverancia y de poner el empeño suficiente en las cosas. Por enseñarme la importancia de afrontar los problemas con la actitud más positiva. Por abrirme nuevas puertas más allá de Granada. Gracias también a mi otro director en la Universidad de Granada, Antonio Maldonado López. Por todo su apoyo y sus sabios consejos. Por haber compartido conmigo su tiempo y su experiencia, y haberme dado todos los empujones necesarios para sacar este trabajo adelante.

I also want to express my deep gratitude to Mirta Galesic, my advisor at the Max Planck Institute for Human Development in Berlin, and to Edward T. Cokely, who has also offered an enormous amount of advice and support at every stage of this work. You have both been a constant source of inspiration along the way, in all possible respects. I would have been lost without your encouragement, wisdom, ideas, and enthusiasm. It has been a great honor to learn from you. Thank you for helping me to keep fighting the good fight.

Gracias también a todos los miembros del grupo de Aprendizaje, Emoción y Decisión y a todos los profesores e investigadores relacionados con el grupo por todo vuestro apoyo, sugerencias y conversaciones enriquecedoras. Mi especial agradecimiento a Andrés Catena Martínez, Antonio Cándido Ortiz, Felisa González Reyes y Jose César Perales.

A todos y cada uno de mis compañeros pre y postdoctorales, habitantes del seminario L (con su llave-reliquia), de la antigua sala de becarios, o errantes de un lado para otro. Alberto, Ana, Dafina, David, Mauro, Leandro, Steffi, Ray, Paola, Fabiano, Ana, Maryem, Eli, Francesco, Almudena, Elena, María, Manu, Sonia, Tania, Enrique, Raúl, Fran, Rocío…Gracias por hacer que el día a día en la facultad se convirtiese en algo que siempre echaré muchísimo de menos. Sin haber podido compartir con vosotros tantas penas y alegrías, cafés y medias tostadas (y algún que otro chupito de Ginja), el camino habría sido infinitamente más difícil.

I would also like to thank the other brilliant PhD students and researchers whom I met during my visits to the Max Planck Institute in Berlin and the Michigan Technological

University. Margo, Saima, Natasha, Katrina, Samantha, Patrick, Joel, Erich, Jacquelyn, Paula, Mona, Nicolai, Azzura, Henrik, Juliane, Florian, Markus, Jonathan, Nadine, Björn, Hans, Niklas, Michelle… It was a pleasure to share thoughts and experiences with you. Thank you for making me feel at home even when I was very far away… and for rescuing me from the snow when I needed it! It was also a pleasure to share ideas (and tapas) with Santiago. ¡Quién hubiese dicho que iba a acabar siguiendo tus pasos!

I am also indebted to Wändi Bruine de Bruin for reading my dissertation and providing thoughtful and inspiring comments that will motivate future steps to build this research programme. I am grateful as well to all the reviewers of my papers. Despite the (sometimes) bad news and challenges, I learnt a lot from all the constructive criticisms, which are always a key to keep improving.

Fuera de los mundos académicos psicológicos, la lista de personas que me han apoyado en cada uno de los altibajos de este camino es demasiado larga. En Granada, gracias a todos los que habéis conseguido que me enamore de esta ciudad y de su gente. Mi agradecimiento especial a Samu y a Isa, por su inagotable y contagioso optimismo, y a mis excompis de piso Antonio, Candi e Isa, por haberme hecho sentir que también tenía familia en Granada. En Madrid, Barcelona, Zaragoza, Sevilla y Bilbao, a todos los que habéis estado siempre ahí, a pesar de la distancia. Álvaro, Jaime, Silvia, Fer, San, Simón, Miguel, Luci (y su familia), Arantza…Mire, ¡al final parece que la cabaña en las montañas y el cultivo de tomates tendrán que esperar! Adela, ¡esas conversaciones predoctorales Zaragoza-Graná, que ahora serán postdoctorales Hannover-Leeds! Habrá muchos más cafés de reencuentro. Y a mi sunshine, que ahora mismo está al otro lado del charco. Esa “imagen en el espejo” con la que llevo años compartiendo tantas cosas, como escribió uno cuando se hacía también Señor Doctor. ¡Parece que por fin aprenderé a pelar aguacates yo sola! Aunque seguro que a ti se te seguirá dando mejor, así que espero que dentro de poco estés cerca para ayudarme. A todos y cada uno de vosotros, gracias.

Y cómo no, a mi familia. A mis tíos en Gijón, Estambul, y Alemania, y a mi primi mayor. A mis abuelos de aquí y de allí, aunque ya no estén. A mi otra familia: Viti, Yavuz y Yunus, y mis casi primos y tíos, Sara, Emma Marisa, Víctor, Ángel y Rosa. Y sobre todo, a mis padres, por vuestro apoyo incondicional. Cualquier cosa que consiga, os la debo a vosotros. No hay palabras para poder expresar lo agradecida que os estaré siempre.

CONTENTS

RESUMEN AMPLIO (IN SPANISH)

Resumen

En la actual era de la información es cada vez más común que las personas estén expuestas a información sobre riesgos médicos y de salud a través de diversas fuentes. Los pacientes cuentan con un acceso cada vez más directo a la información a través de distintos medios de comunicación. Para comprender adecuadamente esta información, las personas frecuentemente necesitan entender conceptos numéricos como porcentajes y probabilidades. Sin embargo, la investigación en toma de decisiones sobre la salud ha mostrado que los médicos y sus pacientes muestran serias limitaciones al razonar con distintos conceptos numéricos (Gardner, McMillan, Raynor, Woolf, & Knapp, 2011; Lipkus, Samsa, & Rimer, 2001; Peters et al., 2006; Peters, 2012; Schwartz, Woloshin, Black, & Welch, 1997). Estas limitaciones pueden desencadenar en una capacidad mermada en los pacientes para participar en la toma de decisiones sobre su salud y desenvolverse en el sistema sanitario (Galesic & Garcia-Retamero, 2011a; Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007).

Las representaciones gráficas de la información (p ej., gráficas de iconos o de barras) pueden mejorar sustancialmente la comunicación y la comprensión de los riesgos (Ancker, Senathirajah, Kukafka, & Starren, 2006; Fuller, Dudley, & Blacktop, 2002; Lipkus & Hollands, 1999; Lipkus, 2007). No obstante, algunas personas pueden mostrar dificultades para comprender incluso gráficos relativamente sencillos. Asimismo, los gráficos utilizados para comunicar información sobre riesgos médicos no siempre presentan los datos de un modo transparente. Ello puede sesgar los juicios sobre la información representada, alterando así las preferencias y los procesos de toma de decisiones sobre la salud.

El objetivo global de la tesis doctoral fue investigar el efecto de la manipulación sistemática de distintas características de diversos tipos de gráficos (gráficas de iconos, barras, y de líneas), con el fin de distinguir aquéllas que facilitan la comprensión de la información de aquéllas que la dificultan. Asimismo, se pretendió investigar el impacto de las *habilidades de comprensión de información gráfica* de las personas (Freedman & Shah, 2002; Galesic & Garcia-Retamero, 2011b; Shah & Freedman, 2011) en los procesos involucrados en la comprensión de los distintos gráficos. Estas cuestiones se abordaron desde un enfoque integrador, considerando diversos marcos conceptuales que permiten establecer predicciones sobre la comprensión de información presentada gráficamente. En concreto, se consideraron (1) modelos de los procesos cognitivos involucrados en la comprensión gráfica (Carpenter & Shah, 1998; Pinker, 1990), (2)

enfoques corpóreos de la cognición (*embodied cognition*) que resaltan el uso del conocimiento adquirido en interacciones con el entorno para interpretar información abstracta (Tversky, 2009; Wilson, 2002), y (3) teorías sobre la adquisición de habilidades y el conocimiento experto (Cokely & Kelly, 2009; Ericsson, Prietula, & Cokely, 2007; Haider & Frensch, 1996, 1999).

El proyecto de la tesis

En primer lugar se llevó a cabo una revisión exhaustiva de la literatura (**Capítulo I**), y posteriormente se llevaron a cabo seis experimentos diseñados con el fin de investigar las cuestiones mencionadas anteriormente. En todos los experimentos se evaluaron las habilidades gráficas de los participantes mediante la escala desarrollada por Galesic y Garcia-Retamero (2011b). Los dos primeros experimentos investigaron el efecto de la representación de la información mediante gráficas de iconos en la comprensión de la reducción del riesgo de morir asociada a la ingesta de distintos fármacos (**Capítulos II y III**). Ambos experimentos se llevaron a cabo en condiciones controladas de laboratorio, y recogieron medidas conductuales. Los resultados del primer experimento (**Capítulo II**) mostraron que las gráficas de iconos facilitan la comprensión de la información, reduciendo sesgos en la comprensión de información numérica. Sin embargo, la mejora en la comprensión fue significativamente mayor en participantes con altas habilidades gráficas, resaltando el papel moderador de las habilidades gráficas en la efectividad de este tipo de apoyo visual. Los participantes con altas habilidades gráficas mostraron también un mayor grado de confianza en sus estimaciones sobre la información de riesgo cuando recibieron gráficas de iconos. En cambio, la confianza de los participantes con bajas habilidades gráficas no se vio afectada por la presencia de gráficas.

En el **Capítulo III** se investigó el efecto de la manipulación de distintos aspectos dinámicos de las gráficas de iconos, con el fin de determinar cómo potenciar el efecto de las mismas en personas con bajas habilidades gráficas. Las distintas manipulaciones de las gráficas se diseñaron con el objetivo de fomentar distintos procesos involucrados en la comprensión gráfica (atención y codificación de los patrones visuales vs. identificación de los referentes de los distintos elementos), según el marco conceptual de Carpenter y Shah (1998). Se observó una mejora significativa en la comprensión de la información a través una manipulación diseñada para fomentar el procesamiento activo de la información representada en los iconos, incluso en participantes con bajas habilidades gráficas. Sin embargo, los otros tipos de iconos dinámicos no llevaron consigo mejoras en la comprensión. Los resultados pusieron también de manifiesto la ausencia de una relación directa entre el grado en que las distintas gráficas mejoran la comprensión y el grado en que se perciben como útiles.

Los siguientes cuatro experimentos investigaron la comprensión de gráficos de barras y de líneas. El primero (**Capítulo IV**) se llevó a cabo en una plataforma online, mientras que los dos siguientes (**Capítulo V**) utilizaron registro de movimientos oculares, con el fin de analizar diferencias en los procesos de comprensión gráfica en participantes con altas y bajas habilidades gráficas. La mayor parte de los gráficos investigados en los experimentos de esta serie se diseñaron de modo que el mensaje transmitido por elementos visuoespaciales (p ej., la altura de las barras) entraba en conflicto con el mensaje transmitido por elementos ligados a convenciones arbitrarias (e.j., los números de las escalas y las etiquetas de los ejes). De este modo, se pretendió determinar el efecto de las habilidades gráficas en la tendencia de las personas a basar sus interpretaciones en un tipo de información frente a otra. Se investigaron gráficos con dos tipos de conflictos: (1) aquéllos que involucran las escalas numéricas, y (2) aquéllos que involucran elementos textuales (títulos y etiquetas de los ejes). En el estudio incluído en el **Capítulo IV** se manipuló también la orientación de los gráficos (i.e., horizontal vs. vertical).

Los resultados mostraron que los participantes con bajas habilidades gráficas basaron sus interpretaciones con mayor frecuencia en elementos visuoespaciales, mostrando por tanto interpretaciones erróneas de la información. Estos resultados apoyan la hipótesis de que las personas con bajas habilidades gráficas se guían en mayor medida por el conocimiento adquirido en interacciones con el entorno para interpretar información en gráficos (p ej., asociaciones entre altura y cantidad; Tversky, 2009; Wilson, 2002). Se observó también que los participantes con altas habilidades gráficas mostraron un menor número de errores de interpretación en la condición de orientación horizontal que en la vertical, para aquellos gráficos que contenían información esencial en las escalas. No obstante, la metodología empleada no permitió determinar hasta qué punto los resultados observados se podían explicar en base a diferencias atencionales a los distintos elementos de los gráficos, o a diferencias en el conocimiento conceptual necesario para interpretar la información y llevar a cabo las inferencias apropiadas.

Esta cuestión se abordó en los experimentos incluídos en el **Capítulo V**, los cuales mostraron que el tiempo de fijación visual en las escalas numéricas tuvo un papel mediador entre las habilidades gráficas y la ejecución. Los participantes con bajas habilidades gráficas dedicaron un menor tiempo a inspeccionar las escalas, indicando que las diferencias en la comprensión entre personas con altas y bajas habilidades gráficas se deben, al menos parcialmente, a diferencias atencionales. Sin embargo, el tiempo de fijación en elementos textuales no mostró un papel mediador. Ello indica que las personas con bajas habilidades gráficas se podrían beneficiar de intervenciones orientadas a dirigir la atención hacia las escalas, mientras que la mejora de la comprensión de los gráficos con información esencial en elementos textuales podría requerir de un entrenamiento a nivel más conceptual.

Por último, el **Capítulo VI** incluye un experimento en el que se investigó un sesgo derivado de características específicas del procesamiento perceptual de las gráficas de barras. Este experimento se planteó en base a un estudio reciente que demostró la existencia de una tendencia a asumir que los datos localizados gráficamente dentro de las barras tienen una mayor probabilidad de formar parte de la distribución que aquéllos que se encuentran fuera de las barras (*within-the-bar bias*; Newman & Scholl, 2012). Los resultados mostraron que este sesgo puede alterar las preferencias de las personas por distintos tratamientos médicos, llevando a los participantes a querer modificar sus niveles de glucosa en sangre aún sin existir razones para ello. Los participantes con altas habilidades gráficas mostraron este tipo de sesgo con mayor frecuencia que aquéllos con bajas habilidades gráficas, indicando posibles diferencias en la atención dirigida a la información textual frente a la gráfica. No obstante, los participantes con altas habilidades gráficas mostraron una reducción en el sesgo cuando los gráficos contenían barras de error bidireccionales. Estos resultados amplían el conocimiento de los factores que pueden dificultar la comprensión de información médica presentada gráficamente, mostrando el impacto de los sesgos ocasionados por principios básicos de procesamiento perceptual.

La discusión general de la tesis (**Capítulo VII**) gira en torno a las implicaciones teóricas y prácticas de los resultados. Entre las aportaciones teóricas destaca el avance en la comprensión de los procesos involucrados en la comprensión de distintos tipos de gráficos, así como del papel de las habilidades gráficas. Se discuten también los resultados en relación a teorías predominantes en la literatura de juicios probabilísticos y toma de decisiones (teoría de la representación borrosa; Brainerd & Reyna, 1990; Reyna & Brainerd, 1995; Reyna, Nelson, Han, & Dieckmann, 2009). Entre las implicaciones prácticas destaca la necesidad de tener en cuenta el nivel de habilidades gráficas de las personas a la hora de evaluar la eficacia de distintos tipos de apoyos visuales. De cara al diseño de gráficas de iconos, se resalta la importancia de fomentar un procesamiento de tipo activo y elaborativo, que permita generalizar la información proporcionada sobre los riesgos. En cuanto al diseño de gráficos de barras y de líneas, se plantea como principio fundamental la necesidad de preservar la compatibilidad entre los elementos visuoespaciales, los elementos ligados a convenciones arbitrarias, y las preguntas que las personas deberán responder en base a los gráficos. Por último, se exponen limitaciones del presente trabajo y se plantean cuestiones a abordar en investigaciones futuras.

CHAPTER I.

INTRODUCTION AND AIMS

Introduction

In today's information age people need to understand health-relevant information in day-to-day medical results (growing charts, cholesterol level, etc.), commercial advertisements, and in the news (Glazer, 2011). People increasingly look online for health information, including search for weight loss diets, vaccination, and immunization (Lewandowsky, Ecker, Seifert, Schwarz, & Cook, 2012). Doctors' offices are full of brochures with information about ways to control risks, and many widespread campaigns have been launched by the authorities to inform the public about risks relating to drugs or diseases such as AIDS (Morgan, Fischhoff, Bostrom, & Atman, 2002). Effective risk communication should involve sharing of information that improves risk understanding and allows shared decision making (Ahmed, Naik, Willoughby, & Edwards, 2012; Edwards, Elwyn, & Mulley, 2002). However, research on health literacy and medical decision making has shown that doctors and patients have severe problems grasping a host of numerical concepts that are prerequisites for understanding health-relevant statistical information (Garcia-Retamero & Galesic, 2009, 2013; Gardner, McMillan, Raynor, Woolf, & Knapp, 2011; Lipkus, Samsa, & Rimer, 2001; Peters, 2012; Peters et al., 2006; Schwartz, Woloshin, Black, & Welch, 1997). Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, and Woloshin (2007) provided evidence that "collective statistical illiteracy" is widespread and can have serious consequences for health. Such consequences include an increase in patients' susceptibility to emotional manipulation and the inability to effectively participate in decision making (Galesic & Garcia-Retamero, 2011a; Gigerenzer et al., 2007).

Fortunately, pictures and graphical displays can facilitate the communication and comprehension of information concerning risks, benefits, and harms of different treatments. Tools such as visual displays—including line plots, bar charts or icon arrays—can represent information in accessible ways, helping to reduce the errors made by professionals and the public alike (Ancker, Senathirajah, Kukafka, & Starren, 2006; Fuller, Dudley, & Blacktop, 2002; Lipkus & Hollands, 1999; Lipkus, 2007). Accordingly, graphical displays are increasingly being used and recommended for the communication of medical risks to the public (Ancker et al., 2006; Fuller et al., 2002; Lipkus, 2007), on the basis of the assumption that they are transparent to the viewer (Glazer, 2011). However, graph comprehension is not always intuitive, and some individuals can have problems understanding even simple graphs. To illustrate, one third of the population in Germany and in the United States have both *low numeracy* (i.e., the ability to use basic probability and numerical concepts; Lipkus & Peters, 2009; Peters, Hibbard, Slovic, & Dieckmann, 2007; Reyna, Nelson, Han, & Dieckmann, 2009) and *low graph literacy* (i.e., the ability to understand graphically presented information; Galesic & Garcia-Retamero, 2011b; Garcia-Retamero & Galesic, 2010b). These troubling findings indicate that inadequately designed graphical displays may not only be unhelpful for many patients, but that such formats could even deceive them.

At the same time, data is not always presented clearly in graphs used to communicate medical information. Graphs containing misleading features such as improperly scaled or split axes can alter peoples' preferences and decisions in substantial ways, leading to biases in judgment and decision making (Arunachalam, Pei, & Steinbart, 2002; Cooper, Schriger, Wallace, Mikulich, & Wilkes, 2003; Kosslyn, 2006; Tufte, 2001). This entails that patients may often need to make potentially life changing decisions using information depicted in graphical displays that can lead to important errors in the comprehension.

The overarching goal of the present dissertation was to achieve a better understanding of the design features in graphs that can either enhance or hinder the comprehension of health-related statistics, and of the different factors that moderate the effect of such features. Special emphasis was placed on the impact of individual differences in graph literacy—an often neglected skill that can influence people's susceptibility to misinterpreting graphically presented data. An essential step towards accomplishing this global aim is to achieve a precise understanding of the difficulties and errors that are more prominent among less graph literate individuals, and of the cognitive processes underlying such difficulties. Six experiments are reported which sought to address these questions using different methodologies (e.g., behavioral data collected through surveys, eye-tracking) and different types of graphs.

Different theoretical frameworks are relevant to anticipate if and how people understand quantitative information presented in graphs, including (1) models of the cognitive processes involved in graph comprehension and usability of graph designs (e.g., Carpenter & Shah, 1998; Kosslyn, 2006; Pinker, 1990; Tufte, 2001), (2) perspectives on embodied cognition and the use of knowledge acquired in the environment to interpret abstract information (Tversky, 2009; Wilson, 2002) and (3) theories on skill acquisition and expert performance (Cokely & Kelley, 2009; Ericsson, Prietula, & Cokely, 2007; Haider & Frensch, 1996, 1999). Here, I sought to bridge relevant notions borrowed from these frameworks, aiming to achieve a novel and comprehensive understanding of how various kinds of displays are processed by different viewers.

In what follows, I will review key milestones in the development and use of different types of graphs to communicate statistical information. Next, I will review recent research investigating the efficacy of graphs to communicate medical risks, as well as work examining different design features that can mislead viewers. I will then discuss prominent models of the processes involved in graph comprehension, and how eye-tracking methodologies have been used to test predictions derived from such models. Lastly, I will discuss the top-down influence of different types of prior knowledge in graph comprehension, focusing on the impact of graph literacy. I will conclude providing an overview of the detailed goals of the experiments reported in this dissertation, as well as highlighting how they contribute to expand previous research.

A brief history of statistical graphs

One of the earliest graphical representations of quantitative information is a graph depicting the changing positions of the seven most prominent planets over space and time, which dates back to the 10th century (Friendly, 2008; Tufte, 2001). However, it was not until the 17th century that the first representation of statistical data was produced, which showed different estimates of the difference in longitude between two cities (Friendly, 2008; Tufte, 2001). In the 18th century numerous forms of data representation were invented and abstract graphs became more widespread, coupled with an increasing interest in the systematic collection of demographic and economic data (Friendly, 2008). The Scottish economist William Playfair played a crucial role in the development of the mostly widely used graphs at present, creating the line graph and the bar chart towards the end of the 18th century. He then developed the pie chart at the beginning of the 19th century, which first appeared in *The Statistical Breviary*, along with other charts depicting statistical data for European countries (e.g., areas of the countries, populations and revenues; Friendly, 2008; Spence, 2005).

An exponential growth in the development of different kinds of statistical graphs and their use followed during the first half of the 19th century, including histograms, time-series plots, and scatterplots (Friendly, 2008; Wainer & Velleman, 2001). Interestingly, English statisticians did not become interested in the use of graphs until late 19th century and early 20th century, and it was Karl Pearson who was responsible for making graphs respectable among statisticians (Spence, 2005). Furthermore, the term *graph* only started to be used in its modern sense about 1910–1920 (Friendly & Denis, 2005). Later, John W. Tukey (1977) proposed various innovations for exploratory data analysis, including stem-leaf plots and box-and-whisker plots.

Another kind of graphical display that gained momentum in the 20th century is the pictograph. Although iconic representations have been used since ancient times, it was not until the early 20th century that the philosopher and economist Otto Neurath considered employing pictographs ("isotypes") as a means to describe statistics such as the employment of women (Kurz-Milcke, Gigerenzer, & Martignon, 2008; Spiegelhalter, Pearson, & Short, 2011). Although the pictographs developed by Neurath have not yet been adapted to represent health statistics (Gigerenzer et al., 2007), they served as an inspiration for many different forms of icon arrays (i.e., matrices of stick figures, faces, circles, squares), which have been used to communicate different kinds of statistical information, with a predominant role in the communication of medical risks. Icon arrays enable to symbolize an at-risk population, and simultaneously display the individuals who are affected and not affected by the risk (Bodemer & Gaissmaier, 2012; Galesic, Garcia-Retamero, & Gigerenzer, 2009; Zikmund-Fisher et al., 2008). As noted by Hess, Visschers, and Siegrist (2011), icon arrays are a special type of graph in the sense that they show frequency information and at the same time convey numbers in a graphical way, enabling a one-to-one match between the individuals represented and the icons. Kurz-Milcke et al. (2008) contend that the fact that icon arrays represent frequency in an analog fashion prompts an identification of the viewer with the individuals represented.

In sum, the history of statistical graphs highlights that such displays were not invented and used commonly until relatively recently. It follows that there is no clear reason for which understanding of such graphs should be intuitive or ingrained in people's minds (Galesic & Garcia-Retamero, 2011b). This emphasizes the need to achieve a better understanding of the comparative efficacy of different kinds of graphs, as well as of the factors affecting such efficacy. In the next section I will discuss relevant findings from recent studies that have examined this issue in the context of the communication of health-related statistics.

Statistical graphs in medicine: Evaluating the effectiveness of graphs to communicate medical risks

Graphical displays that have been used to communicate health-related statistics include pie charts, bar graphs, line graphs (e.g., survival and mortality curves), tree diagrams, risk ladders and scales, and icon arrays (Ancker et al., 2006; Fortin, Hirota, Bond, O'Connor, & Col, 2001; Kurz-Milcke et al., 2008; Lipkus & Hollands, 1999; for a review, see Mt-Isa et al., 2013). A number of such graphs have been used both in paper-based formats and in websites to educate patients, communicate risks, and for decision support (e.g., online decision aids developed by health authorities; Ahmed et al., 2012; Ancker et al., 2006; Edwards et al., 2002; Edwards, Elwyn, & Gwyn, 1999; Lipkus & Hollands, 1999; Lipkus, 2007). However, research on the effectiveness of different graphs to convey benefits and risks is relatively new in the medical literature (Mt-Isa et al., 2013).

Graphs have been proposed to be well-suited to communicate health-related statistics due to a number of reasons, including their capacity to reveal data patterns that may otherwise go undetected, to evoke specific mathematical operations, and to attract and hold people's attention to a larger extent than textual formats (Lipkus & Hollands, 1999; Tversky, Morrison, & Betrancourt, 2002; Tversky, 2001). By using space to organize information, memory and inferences can be facilitated, as the burden on viewers' memory can be reduced (Tversky et al., 2002; Tversky, 2009). Additionally, graphs allow saving time in general practice consultations and can avoid the need to explain the range of different potential options available (Edwards et al., 1999).

Two types of graphs that have received special attention in empirical studies on the communication of medical risks are icon arrays and bar graphs (McCaffery et al., 2012; Timmermans, Molewijk, Stiggelbout, & Kievit, 2004). Icon arrays have been found to improve risk understanding by eliminating errors induced by anecdotal narratives (Fagerlin, Wang, & Ubel, 2005), reducing aversion to side effects which can lead to refuse beneficial therapies (Waters, Weinstein, Colditz, & Emmons, 2007), reducing common judgment biases (e.g., denominator neglect; Garcia-Retamero, Galesic, & Gigerenzer, 2010; Garcia-Retamero & Galesic, 2009), and facilitating the comprehension of complex concepts such as incremental risk of side effects (Zikmund-Fisher et al., 2008). Icon arrays can also enhance gist and verbatim understanding of risk information among parents of children scheduled to undergo elective surgical procedure (Tait, Voepel-Lewis, Zikmund-Fisher, & Fagerlin, 2010).

Bar graphs have also been found to improve the communication of medical risks, for instance, by enhancing the accuracy of evaluations of risk trade offs (Waters, Weinstein, Colditz, & Emmons, 2006). Bar graphs are well-suited for depicting proportions (Lipkus & Hollands, 1999) and for making comparisons between risks by using height to depict quantities (Lipkus, 2007). Moreover, bar graphs in some occasions are preferred over other types of graphs such as line graphs, survival curves, or icon arrays (Fortin et al., 2001).

However, results of studies comparing the efficacy of icon arrays and bar graphs, as well as other kinds of graphs, have often been mixed. For instance, Feldman-Stewart, Kocovski, McConnell, Brundage, and Mackillop (2000) compared accuracy for vertical bar graphs, horizontal bar graphs, icon arrays including either systematically or randomly arranged ovals, and pie charts. They found that vertical bar graphs were the best graphical format for making choices (e.g., selecting the treatment with larger chances of survival or smaller chances of side effects), followed by icon arrays with systematic ovals (see also Feldman-Stewart, Brundage, & Zotov, 2007). The latter format led to the best performance for more specific tasks such as estimating the magnitude of a difference. In contrast, other studies found bar graphs to be most effective for conveying verbatim knowledge (i.e., specific numerical estimations), while pie charts were associated with better gist knowledge (i.e., general impression; Hawley et al., 2008). Other studies have found a superiority of icon arrays over stacked bar graphs to improve people's ability to evaluate risk tradeoffs (Waters et al., 2007) and to improve decision making concerning clinical trials among physicians (Elting, Martin, Cantor, & Rubenstein, 1999).

Finally, other studies have found that risk comprehension does not vary significantly as a function of the specific kind of graph, or even that graphs are of little help to improve accuracy. For instance, Garcia-Retamero and Galesic (2010b) found similar increases in accuracy of treatment risk reduction when either icon arrays or bar graphs were added to numerical information. Tait, Voepel-Lewis, Brennan-Martinez, McGonegal and Levine (2012) reported that people's understanding and perceptions of risks and benefits of treatments was similar for text-based formats and for different types of animated graphs, including pie charts, bar graphs, and icon arrays with human figures. Other authors have also found that different types of graphs including Euler circles, icon arrays, and hybrid graphs combining both formats did not improve accuracy of Bayesian reasoning, as compared to numerical information alone (Micallef, Dragicevic, & Fekete, 2012; Ottley, Metevier, Han, & Chang, 2012).

In sum, previous research has not identified one single type of graph that is consistently superior to communicate medical risks. This is partly due to the variety existing in previous research in the kind of information depicted, display characteristics, tasks types and complexity (Feldman-Stewart et al., 2000; Lipkus, 2007; Mt-Isa et al., 2013; Schonlau & Peters, 2008, 2012). For instance, icon arrays can be more effective than bar graphs when numerators are small (McCaffery et al., 2012) and when outcome probabilities range between 1% and 50% (Dolan, Qian, & Veazie, 2012). Additionally, as noted by Lipkus (2007), research on the graphical communication of medical risks has often been atheoretical. This hinders efforts to explain and integrate mixed findings.

Moreover, people’s preferences for different types of graphs or the extent to which they are perceived as helpful can also vary contingent on display and task characteristics. To illustrate, Timmermans et al. (2004) found that icon arrays were perceived to be more helpful than bar graphs or numbers for decisions concerning a surgical treatment, while bar graphs were evaluated as the most complex format to comprehend. Similar results were obtained by Schapira, Nattinger, and McHorney (2001) in a qualitative study using focus groups. However, Timmermans, Ockhuysen-Vermey, and Henneman (2008) found that icon arrays were not evaluated as easier to understand than numerical formats, and Schapira, Nattinger and McAuliffe (2006) showed that icon arrays were preferred to bar graphs only when presenting single risks. Moreover, physicians have been shown to prefer tables, bar graphs, or pie charts over icon arrays, even though icon arrays are often preferred by nurses and students (Elting et al., 1999). Taken together, these findings highlight the importance of examining how preferences of different populations relate to their risk understanding with different graphical displays, as well as the effect of specific task characteristics on both outcomes.

Finally, it should be noted that improving accuracy of risk understanding constitutes only one communication goal among others such as affecting risk perceptions and risk aversion, or increasing acceptance of interventions. As emphasized by Lipkus (2007), graphical displays that promote accuracy will not necessarily be those that affect risk perceptions or lead to changes in behavior. For instance, when risks are presented in graphs (bar graphs or icons arrays) which do not display the number of people at risk (i.e., background information), people are more risk averse than when risks are presented in numerical formats (Stone et al., 2003; Stone, Yates, & Parker, 1997). While displaying only the number of people harmed graphically (i.e., foreground information) might be most effective to induce people to take actions that protect them from harm and to reduce people’s tendency to reduce low-probability events to zero, this can be at cost of a less complete understanding of risk information (Stone et al., 2003). Thus, important questions concerning the legitimacy and ethical desirability of employing such design features arise (Kurz-Milcke et al., 2008). Some of these issues will be discussed in the next section.

Misleading features in graphs. Towards the development of a set of unified standards of effective design

Graphs can be a double edged sword. While the reviewed evidence suggests that well designed graphs can often support informed decision making in medicine, there is increasing evidence indicating that graphs can also mislead viewers and lead to important judgment errors. Tufte (1997) referred to the graph used by engineers involved in the Space Shuttle Challenger disaster as an unfortunate example of a poorly designed display that led to dramatic consequences. Although poorly designed graphs fortunately do not always lead to such striking consequences, they can alter preferences and decisions in important, and often undesired, ways.

Comprehensive lists of design features in graphs that can distort the data depicted or mislead viewers have been documented by several authors (Arunachalam et al., 2002; Penrose, 2008). A paradigmatic example is Wainer's (1984) sarcastically dubbed "the dirty dozen", which includes 12 techniques to display data badly such as minimizing the data density, graphing data out of context (i.e., choosing to depict specific time intervals), and changing scales in mid-axis (e.g., to make large differences look small). As Penrose (2008) notes, Tufte's (2001) term *lie factor* includes the connotation that the graph designer had a deliberate intention to mislead the viewer. Such intentions seem to be particularly prevalent in financial contexts, as graphs in accounting may often be used with manipulative purposes, in order to create favorable impressions (Beattie & Jones, 1992). To illustrate, Johnson, Rice and Roemmich (1980) noted that, from a sample of Fortune 500 companies, almost half of the annual reports contained at least one misleading graph (e.g., distortions of recent trends), while Beattie and Jones (1992) found that 30 percent of 240 annual reports of UK companies contained distorted graphs.

It could be argued that in medical contexts such distortions may be less prevalent, as motivations to create favorable impressions may not be as widespread as in the financial domain. However, graphs in pharmaceutical advertisements in medical journals do not always reach satisfactory levels of quality, and often include improperly scaled or improperly split axes and improper baselines (Cooper et al., 2003). Misleading graphs can also be found in pharmaceutical marketing campaigns accessible via the Internet (Woller-Carter, Okan, Cokely, & Garcia-Retamero, 2012). In some cases, graphs may not include clearly distinguishable misleading features, but may still alter viewers' preferences and decisions. For instance, if a graph representing incidences of a disease fails to show the number of people at risk (i.e., the reference class), this can give the impression that differences in incidences between groups with and without a treatment are larger than they really are (Kurz-Milcke et al., 2008). Calling attention to certain elements away from others can impact judgments and decisions and result in lower accuracy of risk understanding (Lipkus, 2007).

Distinguishing the instances in which misleading graphs reflect conflicts of interest (Gigerenzer et al., 2007) from the instances in which they merely reflect a lack of sufficient knowledge of effective graph design or of procedures to adjust default outputs provided by statistical software, is not a straightforward endeavor. This task can become particularly challenging if one takes into account that a consensus regarding the adequacy or effective implementation of published graph design guidelines does not always exist (e.g., Gillan, Wickens, Hollands, & Carswell, 1998; Kosslyn, 2006; Shah & Hoeffner, 2002; Tufte, 2001). For instance, Cleveland (1994) argues that the scale in the y-axis of bar charts should start at zero, in order for the proportion between the lengths of the bars to reflect the proportion between the quantitative data. In contrast, Kosslyn (2006) contends that it is adequate not to do so, provided there are marks to indicate discontinuities. Concerning the amount of information to be included in graphs, Tufte (2001) strongly recommends that all elements that fail to communicate anything new to the viewer should be removed, in order to maximize the data:ink ratio (i.e., the proportion of a graphic's ink devoted to the non-redundant display of data information). In contrast, Kosslyn (2006) argues that designers should not always insist on bars that minimize ink, as research indicates that the data:ink ratio has different effects for different displays and tasks.

In sum, the reviewed evidence emphasizes that the assumption that graphs will generally facilitate risk understanding is a naïve one, and that it is not always clear which design features will enhance or hinder understanding. Thus, it is imperative to achieve a precise understanding of the impact of different graph design features on medical judgments and decisions in viewers with varying skill levels. Efforts to develop a systematic and theoretically-grounded conceptualization of different sources of graphic-related biases would significantly contribute to achieve such understanding. Such efforts can be informed by theories and models of graph comprehension developed within the frameworks of cognitive psychology and human–computer interaction. Such models serve as a benchmark for most of the studies reported in the present dissertation, and are discussed next.

Cognitive processes involved in graph comprehension. Testing and refining graph comprehension models using eye-tracking data

Shah, Freedman, and Vekiri (2005) distinguished between two broad classes of models of graph comprehension. The first group of models has focused on providing descriptions of simple graph interpretation tasks or subtasks. These models have emphasized perceptual processes, and are oriented to yielding precise predictions concerning the time required to retrieve specific facts from different types of graphs. In their seminal work, Cleveland and McGill (1984, 1986) developed an empirically-based hierarchy of different perceptual judgments, ranging from the ones that people performed more accurately (i.e., judging position along a common scale), to those that people performed less accurately (i.e., area judgments). Shortly after, Simkin and Hastie (1987) put forward an information-processing analysis of graph perception based on this taxonomy, emphasizing that the accuracy of different perceptual judgments can be linked to the task to be conducted (e.g., comparison judgments vs. estimates of the proportion of the whole). They proposed a series of elementary processes that are thought to operate on mental representations of graphs, enabling people to perform perceptual inferences. The proposed processes received empirical support in some subsequent studies (e.g., Spence & Lewandowsky, 1990; but see also Carswell, 1992). Also focusing on simple graph interpretation tasks, Lohse (1993) developed a model to simulate graphical perception, which predicted response time for fact-retrieval questions about line graphs. However, the model included only a limited range of tasks and types of graphs.

A second class of models has addressed general processes involved in graph comprehension more broadly, beyond the specific perceptual processes or sub-processes involved in decoding visuospatial information. These models have placed more emphasis on the interpretations that people form of the data depicted, and the interactions between features of the graphs (e.g., their content), the viewers' prior knowledge, and different task requirements. A key source of inspiration for such models was Bertin's (1983) work, in which he established a distinction between three major component processes involved in graph comprehension, namely (1) encoding the visual pattern and identifying the principal features (e.g., lines with different slopes), (2) translating the identified features into conceptual relations, including quantitative relations between variables, and (3) determining the referents of the concepts identified by associating them with the specific variables depicted and their numerical values. This characterization of graph comprehension processes was borrowed by Pinker (1990) and by Carpenter and Shah (1998), who articulated further the different aspects of each of the three types of component processes. Importantly, Carpenter and Shah (1998) demonstrated that such processes are performed in a cyclic manner. That is, the processes occur serially and incrementally, although some visual features can also be encoded in parallel.

Carpenter and Shah's (1998) model of graph comprehension has been highly influential until present, and will serve as a framework for most of the experiments reported in this dissertation. The model will be discussed in detail in the coming chapters, in connection with accounts of spatial cognition based on embodied cognition perspectives (Tversky et al., 2002; Tversky, 2009). Carpenter and Shah's model has received empirical support from studies using eye-tracking methodologies, and related graph comprehension processes have also been examined via the analyses of eye tracking metrics, for different kinds of displays (Burns, Elzer, & Carberry, 2008; Carpenter & Shah, 1998; Elzer, Green, Carberry, & Hoffman, 2006; Huestegge & Philipp, 2011; Peebles & Cheng, 2001, 2003; Trafton, Marshall, Mintz, & Trickett, 2002). Relatedly, studies on human-computer interaction have also employed eye-tracking to examine how graph usability is affected by compliance with existing principles of effective design (e.g., Renshaw, Finlay, Tyfa, & Ward, 2003, 2004).

In the present work, two experiments were conducted in which people's eye movements were recorded while they interpreted graphs presenting health-related information. The aim of these experiments was to investigate if and how graph literacy affects the processes underlying comprehension for such kind of graphs. Graphs were designed in such a way that identifying the specific variables depicted and their numerical values was crucial for accurate interpretations, while direct translations of visual features into conceptual relations led to erroneous interpretations. As will be described below, this enabled to investigate the impact of the inclusion of misleading features in graphs, and at the same time pinpoint differences in graph comprehension processes linked to graph literacy.

Top-down influences of prior knowledge on graph comprehension: The impact of graph literacy

Graph literacy refers to one's ability to obtain meaning from graphically presented information, and includes general knowledge about making inferences from different graphic formats (Freedman & Shah, 2002; Galesic & Garcia-Retamero, 2011b; Glazer, 2011; Shah & Freedman, 2011). Graph literacy can include mental representations stored in long-term memory that contain knowledge about the properties of different kinds of displays and procedures for interpreting them (i.e., *graph schemas*; Maichle, 1994; Peebles & Cheng, 2001, 2003; Pinker, 1990; Ratwani & Trafton, 2008; Simkin & Hastie, 1987). Individuals with higher graph literacy can have more complete schemas, which contribute to recognizing specific types of graphs, identifying the most relevant features in each graph, and making accurate interpretations of the information depicted.

Although assessing viewers' level of graph literacy is paramount to anticipate potential difficulties in graph comprehension and to tailor displays to the needs of different viewers, efforts to develop scales addressing this construct have been scarce in the context of health risk communication. As pointed out by Galesic and Garcia-Retamero (2011b), only a few document literacy questions have investigated specific aspects of graph comprehension, but most of the items included are relatively complex (e.g., Kutner, Greenberg, Jin, & Paulsen, 2006; Tuijnman, 2000). Some tests have also been developed in the context of the literature concerned with the effect of different instructional methods on the acquisition of graphical skills in students, including the 36-item Graph Interpretation Test developed by Kramarski and Mevarech (2003) and the 26-item Test of Graphing in Science (TOGS; McKenzie & Padilla, 1986). However, both tests include questions that require relatively advanced skills, focus mainly on line graphs, and are too long to be used in clinical settings. In contrast, efforts to develop instruments to measure health numeracy have been widespread, including both subjective measures (Fagerlin, Ubel, Smith, & Zikmund-Fisher, 2007) and objective measures (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012; Lipkus et al., 2001; Peters et al., 2006; Schapira et al., 2012; Schwartz et al., 1997; Weller et al., 2013), varying in length and difficulty. Of note, items evaluating graph comprehension have in some cases been included in numeracy scales (Schapira et al., 2012).

Galesic and Garcia-Retamero (2011b) recently developed a scale which assesses both basic and more advanced graph comprehension, includes examples of different types of graphs, and is embedded in the health domain. This scale was pretested in the laboratory and evaluated on nationally representative samples in the United States and Germany. The construction of the scale was grounded on the division of graph comprehension skills in three levels proposed by Friel, Curcio and Bright (2001), namely (1) the ability to *read the data*, that is, to find specific information in the graph, which corresponds to the more elementary level, (2) the ability to *read between the data*, that is, to find relationships in the data as shown on the graph, which corresponds to an intermediate level, and (3) the ability to *read beyond the data*, or make inferences and predictions from the data, which corresponds to an advanced level. On the basis of pretest results, 13 items were selected to be included in the refined version of the scale. The scale showed satisfactory levels of reliability (Cronbach's alpha was .74 in Germany and .79 in the United States). Average correlations of the total score with education levels were .29 in Germany and .54 in the United States. Correlations with graph comprehension items from existing literacy questionnaires were .32 in Germany and .50 in the United States, indicating satisfactory convergent validity. This scale was administered in all experiments reported in the present work, and a copy of the full scale can be found in the Appendix.

It is important to note that graph literacy constitutes only one of the different types of prior knowledge that can affect graph comprehension. In addition to data interpretation and graph literacy skills, Shah et al. (2005) highlighted the role of viewers' *knowledge about content* (i.e., the availability of mental representations of the specific content depicted), and individual differences in *visuospatial abilities* and *working memory* (Hegarty & Waller, 2005). Concerning the former type, Carpenter and Shah (1998; see also Freedman & Shah, 2002) noted that less experienced graph viewers can be more likely to rely on prior knowledge concerning typical relations among variables to interpret graphs. Other authors have also noted that familiarity with the content can affect the ease with which different individuals make inferences, even among expert scientists (Roth & Bowen, 2003). Expertise in specific domains (e.g., economics) can also contribute to establish more direct associations between different visual patterns and concepts (Tabachneck-Schijf, Leonardo, & Simon, 1997). As content knowledge can affect graph interpretations independently of graph literacy (Freedman & Shah, 2002; Shah et al., 2005; Shah & Freedman, 2011), participants' level of knowledge concerning relevant clinical conditions was also evaluated in one of the experiments that will be reported. Finally, the relations between visuospatial abilities, working memory and graph interpretations have received less attention in past work, and will not be directly addressed here. Interested readers are referred to work by Feeney, Adams, Webber, and Ewbank (2004), Garcia-Rodríguez, Summers, and Duxbury (2011), Kellen, Chan, and Fang (2013), and Penna, Agus, Peró-Cebollero, Guàrdia-Olmos, and Pessa (2012).

Overview of the Doctoral Thesis

As advanced above, the overarching goal of the present dissertation was to achieve a theoretically-grounded understanding of how the manipulation of different design features of graphs affect the comprehension of health-related statistics in individuals with low and high graph literacy. Six experiments are reported which examine the difficulties and errors that are more prominent among individuals with varying levels of graph literacy, as well as the underlying cognitive processes.

The first two experiments were conducted in a laboratory setting, and focused on examining the efficacy of icon arrays to improve understanding of treatment risk reduction in individuals with low and high graph literacy. In particular, **Chapter II** describes an experiment investigating the effect of icon arrays to reduce denominator neglect, a common judgment bias (Denes-Raj, Epstein, & Cole, 1995; Reyna & Brainerd, 2008). Additionally, the experiment examined whether framing risk information in positive vs. negative terms (i.e., chances of surviving vs. chances of dying, respectively; see e.g., McNeil, Pauker, Sox, & Tversky, 1982) could affect the magnitude of denominator neglect, as previous studies investigating the effect of this bias on understanding of treatment risk reduction had always presented information framed in negative terms (Garcia-Retamero et al., 2010; Garcia-Retamero & Galesic, 2009). We reasoned that an attentional bias for negative information (see, e.g., Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001) could have amplified the effect of denominator neglect observed in previous research. Results revealed that participants showed denominator neglect when only numerical information was provided, and that denominator neglect held consistently, independently of whether risk information was presented in positive or negative terms. In line with previous research, icon arrays helped people to take into account denominators in their estimates of treatment risk reduction, increasing accuracy of risk understanding. However, the effectiveness of icon arrays was moderated by individual differences in graph literacy. Additionally, the provision of icon arrays resulted in a significant increase in subjective confidence only for participants with high graph literacy, indicating that greater confidence was matched by higher accuracy of risk understanding. In sum, the findings of this experiment highlighted that, even though icon arrays may improve understanding by disentangling nested or overlapping classes (Reyna & Brainerd, 2008; Reyna et al., 2009) and by bringing attention to the number of people at risk (Ancker et al., 2006; Stone et al., 2003), a certain level of graph literacy can be necessary to fully benefit from such visual displays.

Chapter III sought to examine the effectiveness of icon arrays including different types of dynamic features. The aim of this experiment was to determine how to enhance risk comprehension among less graph literate individuals, and at the same time to better understand the cognitive processes underlying superior performance. Performance with five different types of dynamic icon arrays was compared to that with a static set of icon arrays. The dynamic icon arrays were designed to promote specific graph comprehension processes (i.e., attention and encoding of the visual pattern; identification of the referents of the different regions of icon arrays; Carpenter & Shah, 1998), as well as a more active processing of the information (e.g., Natter & Berry, 2005) to encourage the generalization of the risk information depicted. In addition to accuracy and confidence, we measured participants' evaluations of the different types of displays. Results showed that the only manipulation that contributed to significantly improve risk understanding was the inclusion of a reflective question (i.e., an estimate concerning the number of people harmed after taking a drug), followed by visual feedback provided interactively through icon arrays. This suggested that viewers may benefit from icon arrays to a larger extent when encouraged to engage in a more active, elaborative, processing of information, leading to richer, better integrated representations which better support subsequent task performance (Cokely & Kelley, 2009). Additionally, most of the dynamic icon arrays were associated with increases in graph evaluation ratings, both for participants with low and high graph literacy. This finding provided new evidence for the lack of correspondence existing between people's subjective evaluations and their performance.

Chapter IV turned to examine how graph literacy affects interpretations and decisions made on the basis of a different type of graph, namely bar graphs. The study was conducted online using Amazon's Mechanical Turk, which provides access to more diverse samples than laboratory-based experiments. Notably, graphs were designed to contain conflicts between information conveyed by spatial features (e.g., heights of bars) and information conveyed by features linked to arbitrary conventions (e.g., axes labels or scale values). This enabled to examine the extent to which graph literacy affects the use of mappings grounded in people's real world experience to interpret graphs (i.e., assuming that a higher bars represents larger quantities), within the framework of

Carpenter and Shah's (1998) model of graph comprehension. Two distinct types of conflicts were distinguished, namely (1) *scale-spatial conflicts* (i.e., involving numerical scales), and (2) *textual-spatial conflicts* (i.e., involving titles and axes labels). While the impact of manipulations linked to scales has often been emphasized in the literature (Gillan et al., 1998; Kosslyn, 2006), the influence of the message conveyed in textual elements has received less attention. However, viewers who fail to incorporate information in titles and labels clarifying the type and nature of the information depicted, will likely misinterpret some types of graphs (e.g., those representing information about percentage change rates or people with negative diagnoses). This experiment also investigated the effect of the orientation of graphs (i.e., vertical vs. horizontal). While some studies have examined how the orientation of icon arrays (McCaffery et al., 2012; Price, Cameron, & Butow, 2007) and of bar graphs (Feldman-Stewart et al., 2007, 2000) affects interpretations and preferences (Schapira et al., 2001), such efforts have rarely been theoretically driven. In the present work, two alternative predictions concerning the effect of orientation on interpretations were proposed and tested.

Results revealed that individuals with low graph literacy more often relied on spatial-to-conceptual mappings and neglected important information in conventional features. Additionally, manipulating the orientation of graphs only affected performance for individuals with high graph literacy, when graphs contained essential information in scales. However, the methodology employed did not enable to determine whether the observed differences in performance were driven by differences in the allocation of attention to different conventional features, or instead by differences in conceptual understanding and mental operations on elements of graphs. This question was addressed in the next chapter.

The next two experiments (**Chapter V**) were conducted in the laboratory in Germany, and involved recording participants' eye movements while they processed graphs. The experimental design was similar to that employed in the previous chapter, but focused on vertically oriented graphs. Line graphs were also included as stimuli to increase the generalizability of our findings, as such kind of graphs are also used to communicate health-relevant information, including changes in the number of patients surviving after different treatments over time (Armstrong, FitzGerald, Schwartz, & Ubel, 2001; Armstrong, Schwartz, FitzGerald, Putt, & Ubel, 2002; Lipkus & Hollands, 1999; Mazur & Hickam, 1990, 1993; Mazur & Merz, 1993). As participants were relatively well educated, four more difficult graph literacy items were included to achieve a better discrimination (see Appendix). Additionally, the second experiment reported in this chapter measured participants' domain-specific knowledge using a computerized version of the Minimum Medical Knowledge questionnaire (Bachmann et al., 2007), as well as other variables that could constitute potential confounding factors of the effect of graph literacy (e.g., knowledge that graphs can be misleading and careless responding).

Results revealed that lower graph literacy was associated with less time spent viewing numerical scales on *x* or *y* axes. Differences in viewing times, in turn, mediated the link between graph literacy and interpretations. In contrast, time spent viewing relevant conventional features in graphs involving textual conflicts did not predict accuracy of understanding and was not related to graph literacy. These findings suggested that graph literacy affects people's tendency to strategically direct attention to and encode some conventional features, thus expanding previous research on perceptual and cognitive processes in graph comprehension (Carpenter & Shah, 1998; Kosslyn, 1989; Lohse, 1993; Pinker, 1990; Shah & Carpenter, 1995; Simkin & Hastie, 1987). Skilled individuals were more able to recognize and focus on task-relevant information (see also Haider & Frensch, 1996, 1999), suggesting that graph literacy can affect *which* information is processed, and not only *how* the information is processed. At the same time, these findings indicated that methods to direct attention to essential information in conventional features could enhance performance for graphs containing essential information in scales, while specific training might be required at a conceptual level for graphs containing essential information in textual elements.

Finally, **Chapter VI** turned to examine the impact of more basic pattern perception processes specific to bar graphs, which can also have important consequences for the graphical communication of health-related information. Recent research has revealed that when people are shown a bar graph representing a mean, they often believe that data points located within bars are more likely to be part of the underlying distribution than equidistant points outside bars (Newman & Scholl, 2012). An experiment was conducted to examine the generalizability of the *within-the-bar bias* in the medical domain, as well as the moderating effect of graph literacy. Results revealed that this bias led participants to prefer to modify their blood glucose levels, even when the information provided gave them no justifiable reason to do so. Interestingly, individuals with higher levels of graph literacy showed the largest biases. These findings were interpreted in terms of differences in the extent to which participants with low and high graph literacy focused on textual-based information vs. on graphs. Graph literate individuals were also found to benefit more from the inclusion of bidirectional error bars in graphs, suggesting that the success of debiasing efforts may be contingent on the viewer's level of graph literacy. In sum, this experiment revealed that ensuring that bar graphs comply with principles of good graph design is most likely necessary, but not sufficient, to promote accurate comprehension and informed decision making.

The dissertation concludes with **Chapter VII**, which reviews and integrates all reported findings, discusses theoretical and practical implications, and provides general conclusions, along with a series of recommendations for avenues for future research.

CHAPTER II.

INDIVIDUAL DIFFERENCES IN GRAPH LITERACY: OVERCOMING DENOMINATOR NEGLECT IN RISK COMPREHENSION

Individual Differences in Graph Literacy:
Overcoming Denominator Neglect in Risk Comprehension

Abstract

Graph literacy is an often neglected skill that influences decision-making performance. We conducted an experiment to investigate whether individual differences in graph literacy affect the extent to which people benefit from visual aids (icon arrays) designed to reduce a common judgment bias (i.e., denominator neglect—a focus on numerators in ratios while neglecting denominators). Results indicated that icon arrays more often increased risk comprehension accuracy and confidence among participants with high graph literacy as compared to those with low graph literacy. Results held regardless of how the health message was framed (chances of dying vs. chances of surviving). Findings contribute to our understanding of the ways in which individual differences in cognitive abilities interact with the comprehension of different risk representation formats. Theoretical, methodological, and prescriptive implications of the results are discussed (e.g., the effective communication of quantitative medical data).

Published as:

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2012). Individual differences in graph literacy: Overcoming denominator neglect in risk comprehension. *Journal of Behavioral Decision Making*, *25*, 390–401.

Introduction

The popular saying "a picture is worth a thousand words" reflects the widespread belief that pictures and graphical displays can facilitate the communication and comprehension of complicated information. In modern societies there is a growing need for such simplification. For instance, research has documented many ways in which doctors and patients struggle to grasp numerical concepts that are prerequisites for the accurate evaluation and communication of risks (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Peters et al., 2006; Schwartz, Woloshin, Black, & Welch, 1997). Fortunately, tools such as visual displays—including line plots or bar charts—can help overcome some of these difficulties in professionals and the public alike (Ancker, Senathirajah, Kukafka, & Starren, 2006; Fuller, Dudley, & Blacktop, 2002; Lipkus, 2007; Lipkus & Hollands, 1999). However, graphs are not equally useful for all individuals (Ancker et al., 2006; Garcia-Retamero & Galesic, 2010b; Lipkus, 2007). Recent research has shown that people differ substantially in their ability to understand graphically presented information, or *graph literacy* (Galesic & Garcia-Retamero, 2011b). In this paper we address the question of how individual differences in graph literacy influence the efficacy of visual displays.

Individuals with high graph literacy have been found to make more elaborate inferences when viewing graphical displays as compared to less graph-literate individuals. For instance, highly graph-literate individuals extract information of a higher level of complexity when viewing line graphs than do individuals with low graph literacy. They are also more likely to direct their attention to typical line graph information (e.g., quantitative trend information; Maichle, 1994). When viewing bar graphs, individuals with high graph literacy are more capable of making main effect inferences on the basis of the data represented than are less graph-literate individuals (Shah & Freedman, 2011). Moreover, novice graph viewers often neglect the relevance of important elements of graphs (Mazur & Hickam, 1993) and interpret graphs incorrectly as compared to experienced graph viewers (Shah & Hoeffner, 2002). Differences in individual-level skills such as numeracy (i.e., the ability to process basic probability and numerical concepts; Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012; Galesic & Garcia-Retamero, 2010; Peters et al., 2006; Reyna, Nelson, Han, & Dieckmann, 2009) also affect people's reactions to different graphic representation formats (Wright, Whitwell, Takeichi, Hankins, & Marteau, 2009), the extent to which visual aids are useful in the assessment of treatment risk reduction (Galesic, Garcia-Retamero & Gigerenzer, 2009;

Garcia-Retamero & Galesic, 2010b), and recall of numerical information (Garcia-Retamero & Galesic, 2011). Other studies have documented a range of other factors, such as age (Garcia-Retamero, Galesic, & Gigerenzer, 2010), or people's familiarity with the specific content depicted (Shah, 2001; Shah & Hoeffner, 2002), that tend to be related to graph literacy. For example, familiarity can affect the ease with which different individuals make inferences based on graphs, even amongst expert scientists (Roth & Bowen, 2003).

Although a growing body of evidence has documented differences in people's ability to understand graphs, relatively less research has examined how graph literacy influences the efficacy of risk communication interventions. . Visual displays such as line plots or bar charts facilitate the communication of information by enabling the representation of quantitative information in spatial locations. Thus, visual displays facilitate inferences about conceptual relations in the data to be made on the basis of spatial relations (Gattis, 2002, 2004; Gattis & Holyoak, 1996; Kosslyn, 2006; Tversky, 2001). However, it is unclear how, when, and why differences in graph literacy affect the efficacy of risk communication tools such as visual aids When people are faced with such tools, graph literacy could interact with fundamental skills needed for competent decision making and reasoning (e.g., consistency in risk perception; Bruine de Bruin, Parker, & Fischhoff, 2007; Parker & Fischhoff, 2005). Here, we report on an investigation of the extent to which individual differences in graph literacy influence the effectiveness of graphs designed to reduce a common bias in judgment and decision making, namely denominator neglect.

Denominator neglect refers to people's tendency to pay too much attention to numerators in ratios (i.e., the number of times a target event has happened) and insufficient attention to denominators (i.e., the overall opportunities for it to happen; Denes-Raj, Epstein, & Cole, 1995; Reyna, 2004; Reyna & Brainerd, 2008). This tendency has been observed in the health domain in numerous studies, leading people, for example, to judge cancer as riskier when it is described as killing 1,286 out of 10,000 people than as killing 24.14 out of 100 people (Yamagishi, 1997), or to judge 36,500 people dying of cancer every year as riskier than 100 dying every day (Bonner & Newell, 2008). An example of denominator neglect in a medical context would be focusing on the number of treated and non-treated patients who die, without considering the overall number of treated and non-treated patients, in judging whether a treatment was effective.

Consequently, people often assess treatment risk reduction inaccurately (Garcia-Retamero & Galesic, 2009).

Denominator neglect can be particularly problematic when people are required to judge the effectiveness of a treatment using information from unequally sized groups of treated and non-treated patients, which is a common situation in medical practice (e.g., Grossarth-Maticek & Ziegler, 2008; Lichtenberg, Levinson, Sharshevsky, Feldman, & Lachman, 2008). In particular, in a loss frame (i.e., number of patients who died), when the overall number of patients who receive a treatment is smaller than the number of those who do not receive it (e.g., 100 and 800, respectively), people tend to overestimate risk reduction (Garcia-Retamero & Dhami, 2011; Garcia-Retamero & Galesic, 2009; Garcia-Retamero et al., 2010). In this situation, people would overestimate treatment risk reduction because they would take into account the absolute numbers of treated and non-treated patients who die (e.g., 2 and 80, respectively) rather than the proportion of treated and non-treated patients who die (e.g., 2 out of 100 and 80 out of 800, respectively).

Icon arrays (i.e., graphical representations consisting of a number of circles or other icons symbolizing individuals who are affected by some risk; Ancker et al., 2006; Edwards, Elwyn, & Mulley, 2002; Paling, 2003) are an effective method for eliminating denominator neglect and increasing the accuracy of people's risk estimates (Garcia-Retamero & Galesic, 2009; Garcia-Retamero et al., 2010).It has been suggested that icon arrays improve the accuracy of quantitative reasoning by disentangling classes which are overlapping in ratios, making part-to-whole relations visually available (e.g., Reyna, 1991; Reyna & Brainerd, 2008; see also Ancker et al., 2006). However, in line with the literature reviewed examining the comprehension of visual displays such as line plots or bar charts (e.g., Maichle, 1994; Shah & Freedman, 2011), it is possible that a certain level of graph literacy is required to associate the visual patterns contained in icon arrays with meaningful interpretations of the data represented (i.e., risk reduction information). This would imply that individual differences in graph literacy could affect the effectiveness of icon arrays in improving the accuracy of quantitative reasoning. Investigating this issue was the main aim of this paper.

We tested the hypothesis (H_1) that the effectiveness of icon arrays in reducing denominator neglect would be larger for individuals with high graph literacy than for individuals with low graph literacy. We further hypothesized (H_2) that highly graph-literate participants would report more confidence in their estimates when icon arrays are provided as compared to less graph-literate participants. We measured individual differences in graph literacy using a scale developed by Galesic and Garcia-Retamero (2011b). This scale covers four frequently used graph types—i.e., line plots, bar charts, pies and icon arrays—and has been designed to measure graph comprehension in the medical domain.

Another factor that plays an important role in risk comprehension is the structure and the content of the message (i.e., *message framing*). Previous research has demonstrated that the presentation of information in a negative vs. a positive frame can have a large impact on judgment and decision making (Edwards, Elwyn, Covey, Matthews, & Pill, 2001; Garcia-Retamero & Cokely, 2011; Garcia-Retamero & Galesic, 2010a; Levin, Schneider, & Gaeth, 1998; Rothman & Salovey, 1997). For instance, studies in medical contexts have shown that the likelihood for people to engage in illness-detecting behaviors is larger when messages are framed in terms of potential losses, while gain-framed messages are more likely to lead to prevention behaviors (Banks et al., 1995; Gerend & Shepherd, 2007; Rivers, Salovey, Pizarro, Pizarro, & Schneider, 2005; Rothman, Martino, Bedell, Detweiler, & Salovey, 1999; Toll et al., 2010). Other studies have documented how manipulating the way in which risks associated with different treatments are framed (i.e. chances of surviving vs. chances of dying) affects evaluations and preferences for these treatments (Haward, Murphy, & Lorenz, 2008; Marteau, 1989; McNeil, Pauker, Sox, & Tversky, 1982; Wilson, Kaplan, & Schneiderman, 1987).

In studies investigating the effect of denominator neglect on perceptions of treatment risk reduction, information has always been framed in negative terms (Garcia-Retamero & Galesic, 2009; Garcia-Retamero et al., 2010). To the best of our knowledge, no study has yet analyzed whether denominator neglect equally affects the accuracy of people's estimates of risk reduction when information is presented in either negative or positive terms (i.e., chances of dying and surviving, respectively). This issue is of interest as research has demonstrated the generalized existence of an attentional bias toward negative information (Baumeister, Bratslavsky, Finkenauer, & Vohs, 2001). This bias has been documented in studies analyzing people's visual search of faces (Hansen & Hansen, 1988; Öhman, Lundqvist, & Esteves, 2001), color-naming latencies in the emotional Stroop task (Pratto & John, 1991), and event-related brain potentials associated to negative vs. positive stimuli (Smith, Cacioppo, Larsen, & Chartrand, 2003). A common finding in these studies is that negative stimuli elicit attention more automatically than positive stimuli. Therefore, in tasks investigating understanding of treatment risk reduction, the presentation of information in negative terms (i.e., chances of dying), rather than positive terms (i.e., chances of surviving) could exacerbate the effect of denominator neglect. The attentional bias for negative information might lead people to focus on the absolute numbers in the numerators (i.e., the number of people that die) and overlook the denominator. Investigating this issue was an additional aim of this paper. In particular, we aimed to determine whether an attentional bias for negative information could amplify the effect of denominator neglect. If this is the case, denominator neglect should be larger when information is presented in terms of chances of dying, as compared to when it is presented in terms of chances of surviving.

To test the hypotheses stated so far, we conducted a study where we analyzed participants' understanding of medical risk reductions after treatment. Participants with different levels of graph literacy were presented with scenarios involving equally effective treatments but differing in the overall number of treated and non-treated patients. In some conditions, the number of patients who did receive a treatment was equal to those who did not; in other conditions, it was smaller or larger. Some participants were provided with icon arrays alongside numerical information about risk reduction, whereas others received numerical information only. To test the hypotheses that differences in graph comprehension skills affect both the effectiveness of icon arrays in reducing denominator neglect (H_1) and people's confidence in their estimates (H_2), we compared the accuracy of risk reduction estimates and the self-reported confidence in these estimates in participants with high vs. low graph literacy scores. Additionally, to analyze whether message framing could affect risk understanding we provided half of the participants with the information for the medical scenarios in terms of chances of dying, while the other half received the information in terms of chances of surviving.

Method

Participants

Participants were 168 undergraduate students from the University of Granada, Spain (16% women, median age of 20 years, range 18–28). They were recruited through the University's online recruitment pool and through advertisements made during lectures, and participated in exchange of course credit. A paper-and-pencil questionnaire was completed by participants in group sessions ranging from two to twelve participants. The sessions were always conducted under the supervision of one of the researchers, in order to ensure that questionnaires were completed individually. The tasks relevant for the present study took between 15–20 minutes to complete. Afterwards, participants completed other unrelated tasks for an additional 40 minutes. Participants were randomly assigned to the different experimental groups.

Stimuli and Procedure

Participants were presented with four medical scenarios describing the usefulness of hypothetical new drugs for reducing cholesterol that also decreased the risk of dying from a heart attack. The order of the four scenarios was randomized. Participants read and evaluated information about the risks and subsequently completed a graph literacy scale.

Measurement of graph literacy: Graph literacy scores were collected using the instrument developed by Galesic and Garcia-Retamero (2011b). This scale consists of 13 items and includes items reflecting three levels of graphical comprehension traditionally outlined in the literature (see Friel, Curcio, & Bright, 2001): (1) the ability to *read the data*, that is, to find specific information in the graph, which corresponds to the more elementary level (for instance, the ability to read off the height of a particular bar within a bar chart, or the number of icons of a particular type in an icon array); (2) the ability to *read between the data*, that is, to find relationships in the data as shown on the graph, which corresponds to an intermediate level (for instance, the ability to read off the difference between two bars or sets of icons); and (3) the ability to *read beyond the data*, or make inferences and predictions from the data, which corresponds to an advanced level (for example, the ability to project a future trend from a line chart, or to understand the importance of attending to scale ranges and scale labels when comparing two charts). The scale contains four items assessing the ability to *read the data*, four items assessing the ability to *read between the data*, and five items assessing the ability to *read beyond the data.* For examples of items, see Figure 1.

Additionally, the scale is designed to cover four frequently used graph types—line plots, bar charts, pies, and icon arrays—and includes items dealing with the communication of medical risks, treatment efficiency, and prevalence of diseases. In sum, the scale measures both basic graph-reading skills and more advanced graph comprehension, for different types of graphs. The psychometric properties of this scale have been assessed in a survey conducted on probabilistically representative national samples of people from Germany and the United States, demonstrating satisfactory levels of internal consistency (Cronbach alpha of .74 in Germany and .79 in the United States; .70 in the sample of the current study) and convergent validity (the average correlation of the total score with graph comprehension items from existing literacy questionnaires was .44; for further details on the psychometric properties of the scale see Galesic & Garcia-Retamero, 2011b).

We split participants into two groups according to the median graph literacy score for the total sample (i.e., 10). The group of participants with low graph literacy included those who obtained nine or fewer correct responses (n = 68), while the group of participants with high graph literacy included those who obtained 10 or more correct responses (n = 100). Participants with low graph literacy answered on average 7.8 items correctly (*SD* = 2.0), while participants with high graph literacy answered on average 10.9 items correctly (*SD* = .8).

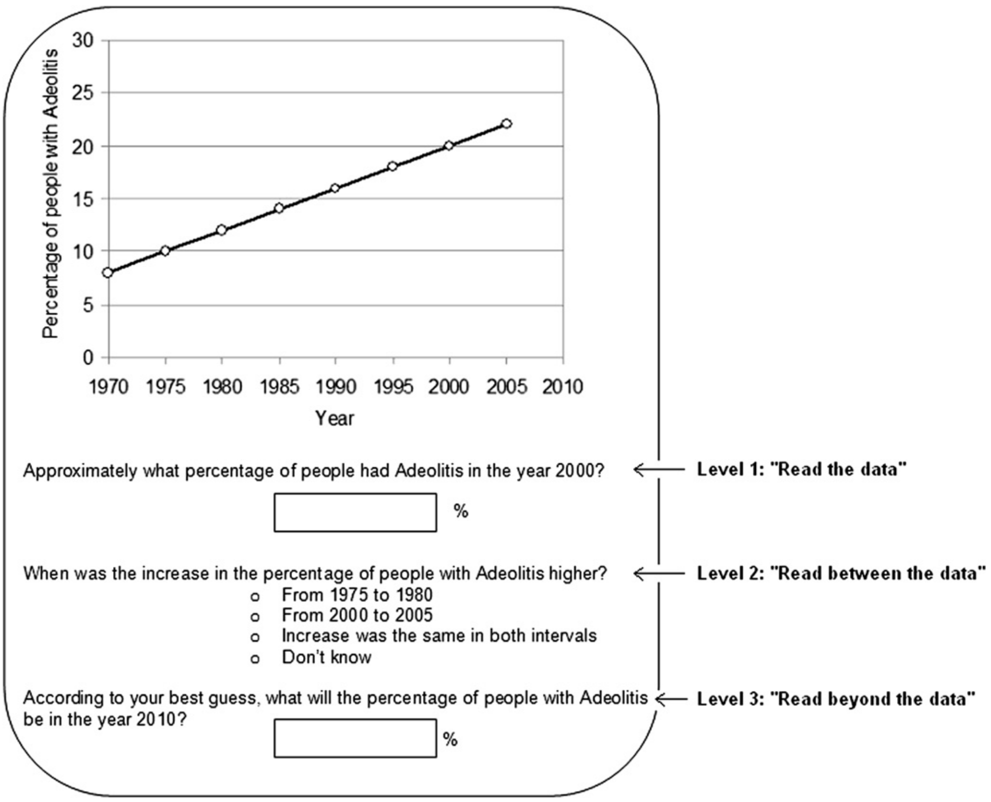


Figure 1. Examples of items measuring the three abilities of graph comprehension. M. Galesic & R. Garcia-Retamero, Medical Decision Making, 31, 444–457, Copyright © 2011 by Society for Medical Decision Making. Reprinted by Permission of SAGE Publications

Information about medical risks: An example of the information presented in the medical scenarios for the negative framed message condition is as follows (the original material was in Spanish)**:**

"A new drug that reduces cholesterol, Benofreno, *decreases* the chances of *dying* after a heart attack for people with high cholesterol. Here are the results of a study of 900 people with high cholesterol: 80 out of 800 people who *did not take* the drug *died* after a heart attack, compared to 2 out of 100 people who *took* the drug."

In the positive framed message condition, the information was presented as follows:

"A new drug that reduces cholesterol, Benofreno, *increases* the chances of *surviving* after a heart attack for people with high cholesterol. Here are the results of a study of 900 people with high cholesterol: 10 out of 100 people who *took* the drug *survived* after a heart attack, compared to 16 out of 800 people who *did not take* the drug."

The rest of the drugs were named Cenofreno, Denofreno and Genofreno, respectively.

Design

Three independent variables were manipulated in the study. First, *message frame* was manipulated between-subjects by providing half of the participants with the information for all medical scenarios in terms of chances of dying, while the other half received the information in terms of chances of surviving. Second, the overall numbers of treated and non-treated patients (i.e., *the sizes of the denominators*) were manipulated within-subjects, and were set to be either 800-800, 100-800, 800-100, 100-100, where the first and second quantities reflect the overall number of patients who did and did not take the drug, respectively. The sizes of the numerators—i.e., the number of treated and non-treated patients who died (survived) in the negative (positive) framed message condition—varied within conditions depending on the size of the denominator. In particular, the treatment always had an 80% relative risk reduction or increase in survival rate (i.e., from 10% to 2% in terms of chances of dying, or from 2% to 10% in terms of chances of surviving; see Table 1)

Table 1. Number of treated and non-treated patients who die (negative framing; top panel) or survive (positive framing; bottom panel) after a heart attack for all denominator sizes.

	Treated patients		*Non-treated patients*	
Denominator Sizes	Patients who died	Population size	Patients who died	Population size
800-800	16	800	80	800
800-100	16	800	10	100
100-800	2	100	80	800
100-100	2	100	10	100

	Treated patients		*Non-treated patients*	
Denominator Sizes	Patients who survived	Population size	Patients who survived	Population size
800-800	80	800	16	800
800-100	80	800	2	100
100-800	10	100	16	800
100-100	10	100	2	100

Note: Risk reduction/ increase in survival rate is 80% in all conditions

Finally, the presentation of *visual aids* was manipulated between-subjects by providing half of the participants with two icon arrays in addition to the numerical information for each medical scenario. These icon arrays presented the risk of dying of a heart attack/surviving after a heart attack when the drug was and was not taken. All icon arrays contained either 800 or 100 circles depending on the overall number of patients who did and did not take the drug. The patients who died (in the negative framed message condition) or survived (in the positive framed message condition) were represented with black circles at the end of the array (see Figure 2 for an example, original material was in Spanish).

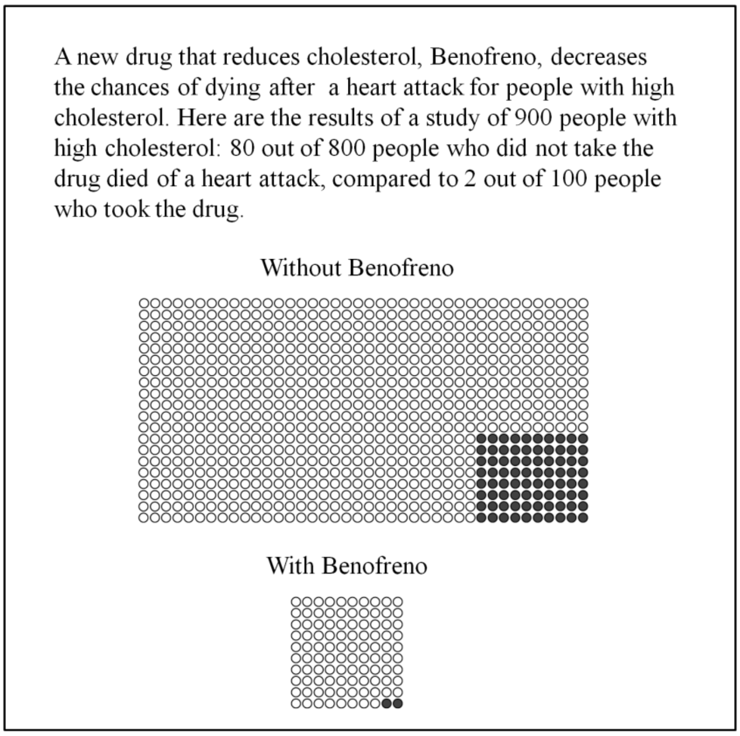


Figure 2. Numerical information about risk reduction and icon arrays that participants received. This example presents the information provided in the negative framing, 800-100 condition

As dependent variables we measured participants' accuracy of risk understanding after they read the information provided for each medical scenario, as well as their subjective confidence in the estimates given. In order to measure accuracy of risk understanding, we followed the procedure used by Schwartz, Woloshin, Black, and Welch (1997). First, participants were asked how many of 1,000 patients with high cholesterol might die of/survive after a heart attack (for negative and positive message frame, respectively) if they do not take the drug. Second, they were asked how many of 1,000 patients with high cholesterol might die of/survive after a heart attack if they did take the drug (for negative and positive message frame, respectively). By subtracting the second from the first answer and dividing it by the first, we calculated the estimated relative risk reduction for the negative message frame. For the positive message frame, we subtracted the first from the second answer and then divided it by the second, in order to calculate the estimated increase in survival rate. Participants were then classified depending on whether their estimates were accurate, lower, or higher than the exact value (i.e., 80%). Estimates were treated as correct only when they were exactly right.

Participants' degree of confidence in their estimates was measured on a scale of 1 to 10, were 1 represented "not at all confident" and 10 represented "very confident."

In sum, our experimental design can be summarized as a 2 (high- vs. low-graph literacy; between-subjects) × 2 (positive vs. negative message frame; between-subjects) × 4 (size of denominators 800-800, 100-800, 800-100, 100-100; within-subjects) × 2 (absence vs. presence of icon arrays; between-subjects) factorial design. To assess the effect of these factors on risk perceptions and confidence we conducted analyses of variance (ANOVAs). Note that for the ANOVAs on risk perceptions that will be described below we only considered whether scores were accurate (1) or inaccurate (0). Thus, the dependent variable accuracy of risk understanding is dichotomous .We used the Bonferroni correction for post hoc analyses.

Results

First, we aimed to determine whether participants showed denominator neglect in their estimates of treatment risk reduction, and to analyze differences in denominator neglect as a function of message frame. To this end, we conducted a 2 × 4 ANOVA with message frame as between-subjects factor and sizes of denominators as a within-subjects factor on the percentage of participants whose estimates of risk reduction were accurate in the numerical condition only (i.e., when participants did not receive icon arrays). We followed Lunney (1970; see also Cleary & Angel, 1984), who showed that ANOVAs can be used to obtain conservative results for large samples of a dichotomous dependent variable. The analysis only revealed a main effect of sizes of the denominators, $F(3, 216) = 6.45$, $p = .001$, $\eta_p^2 = .082$. Thus, consistent with previous research, when no icons were provided and the sizes of the denominators were equal (i.e., in the 800-800 and 100-100 conditions), participants' estimates were significantly more accurate than when the denominators were different (i.e., 800-100 and 100-800 conditions). The ANOVA did not reveal a main effect of message frame, or an interaction involving message frame implying that the percentage of accurate estimates did not reliably vary as a function of this variable ($F < 1$). Thus, we did not find support for the notion that an attentional bias might exacerbate the effect of denominator neglect when information is framed negatively.[1]

[1] An anonymous reviewer suggested the possibility of treating sizes of denominators as a factor with two levels (i.e., same denominators: 800-800 and 100-100 conditions vs. different denominators: 800-100 and 100-800 conditions), instead of as a factor with four levels. Two additional ANOVAs conducted following

These results suggest that participants tended to pay too much attention to numerators and insufficient attention to denominators in both message framing conditions (see Figure 3). Accordingly, in the *negative* framed message condition, when the number of treated patients was lower than the number of those who did not receive the treatment (i.e., in the 100-800 condition), 52% of participants' estimates were *higher* than the exact value compared to 13% and 35% that were lower and accurate, respectively. Denominator neglect accounts for this result, given that the absolute number of patients who received the treatment and died is much lower than the absolute number of patients who did not receive the treatment and died (e.g., 2 and 80, respectively). Focusing on the absolute numbers in the numerators would lead participants to believe that the treatment had a larger effect than it actually did. Instead, in the *positive* framed message condition there was a tendency to underestimate risk reduction: 48% of participants' estimates were *lower* than the exact value, compared to 9% and 43% that were higher and accurate, respectively. Here, focusing on the absolute numbers in the numerators would lead participants to believe that the treatment had a smaller effect than it actually did.

Instead, when the number of treated patients was higher than the number of patients who did not receive the treatment (i.e., in the 800-100 condition), there was a tendency to underestimate risk reduction in the *negative* framed message condition: 61% of participants' estimates were *lower* than the exact value, compared to 3% and 36% that were higher and accurate, respectively. Instead, in the *positive* framed message condition, 67% of the estimates were *higher* than the exact value, compared to 5% and 28% that were lower and accurate, respectively.

For the 100-800 condition, participants in the negative message framed condition were significantly more likely to overestimate risk reduction, while participants in the positive message framed condition were more likely to underestimate risk reduction, $X^2(1, N = 51) = 21.40$, $p = .001$. Instead, for the 800-100 condition participants in the negative message framed condition were more likely to underestimate risk reduction, while participants in the positive message framed condition were more likely to overestimate risk reduction, $X^2(1, N = 54) = 42.55$, $p = .001$.

---

this approach yielded converging results. Specifically, two 2 × 2 × 2 ANOVAs with message frame as a between-subjects factor, and sizes of denominators (same vs. different) and either sizes of non-treated patients (800 vs. 100; first ANOVA) or sizes of treated patients (800 vs. 100; second ANOVA) as within-subjects factors revealed only a significant main effect of sizes of denominators ($p < .001$) and no main effect or interaction involving message frame or the sizes of non-treated or of treated patients (*Fs* < 1).

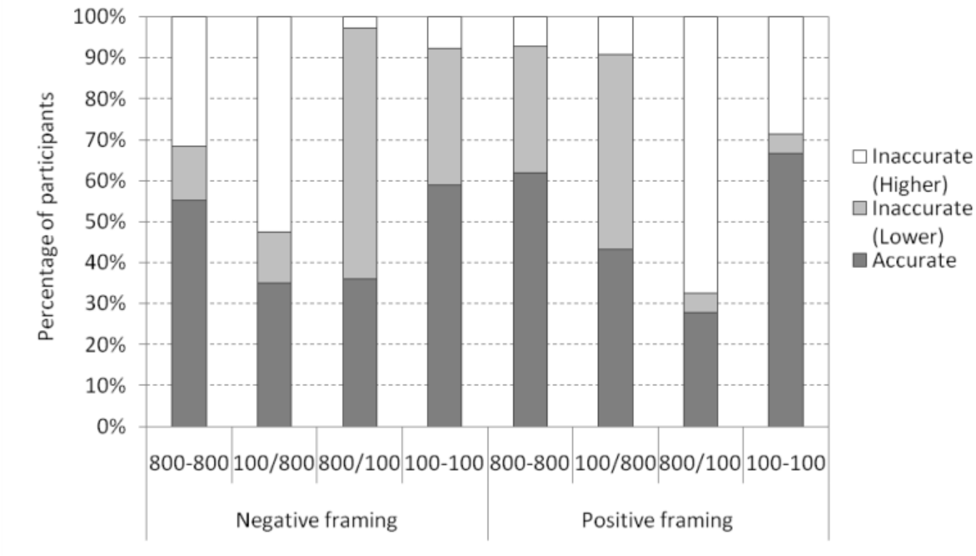


Figure 3. Percentage of participants whose estimates of risk reduction were accurate, lower, or higher than the exact value in the numerical condition only, as a function of the sizes of the denominators and message framing

Next we tested whether icon arrays were more effective in reducing denominator neglect in participants with high graph literacy than in those with low graph literacy (H_1). To this end, we conducted a 2 × 2 × 4 ANOVA with icon arrays and graph literacy as between-subjects factors and sizes of denominators as a within-subjects factor on the percentage of participants whose estimates of risk reduction were accurate. Note that in contrast to the first ANOVA reported above, the current analysis included data for participants who did and did not receive icon arrays. Additionally, we excluded message framing as a factor, given the absence of any previous significant effect on denominator neglect. This analysis revealed a main effect of sizes of denominators, $F(3, 411) = 2.85$, $p = .037$, $\eta_p^2 = .020$, a main effect of icon arrays, $F(1, 137) = 26.62$, $p = .001$, $\eta_p^2 = .163$, and an interaction between icon arrays and sizes of denominators, $F(3, 411) = 5.74$, $p = .001$, $\eta_p^2 = .040$. These results indicate that icon arrays helped people to take into account both the overall number of treated and non-treated patients in their estimations of treatment risk reduction. When the sizes of the denominators were different and icon arrays were presented alongside numerical information, the percentage of correct estimates increased from 42% to 73%, and from 34% to 81%, for the 100-800 and 800-100 conditions, respectively ($p < .001$). Instead, when the sizes of denominators were equal, the increase in accuracy when icon arrays were provided was not significant (p > .1).

The ANOVA also revealed a significant main effect of graph literacy, $F(1, 137) = 6.40$, $p = .013$, $\eta_p^2 = .045$, and an interaction between icon arrays and graph literacy, $F(1, 137) = 4.81$, $p = .030$, $\eta_p^2 = .034$.[2] When icon arrays were not provided, 48 % of the participants with low graph literacy provided correct estimates, compared to 64 % when icon arrays were provided. For participants with high graph literacy, the percentage of correct estimates instead raised from 51 % to 87 %. In sum, icon arrays helped all participants (i.e., with high and low graph literacy) to take into account the overall number of treated and non-treated patients, thus reducing denominator neglect. However, in line with H_1, the overall increase in accuracy of risk understanding when icon arrays were provided was significantly larger for participants with high graph literacy that for those with low graph literacy (see Figure 4).[3]

Finally, to test whether differences in graph literacy affected participants' confidence in their estimates of treatment risk reduction (H_2), we conducted a 2 × 2 × 4 ANOVA with icon arrays and graph literacy as between-subjects factors and sizes of denominators as a within-subjects factor on participants' confidence ratings. This analysis revealed a significant main effect of graph literacy, $F(1, 164) = 7.61$, $p = .006$, $\eta_p^2 = .044$, and icon arrays, $F(1, 164) = 6.67$, $p = .011$, $\eta_p^2 = .039$, and an interaction between graph literacy and icon arrays, $F(1, 164) = 5.59$, $p = .019$, $\eta_p^2 = .033$. As Figure 5 shows, the provision of icon arrays alongside the numerical information resulted in a significant increase in confidence for participants with high graph literacy, but not for those with low graph literacy, supporting H_2. The mean confidence reported by participants with high

[2] The results of an ANOVA including all factors manipulated yielded converging results. Specifically, a 2 × 2 × 2 × 4 ANOVA with icon arrays, graph literacy, and message frame as between-subjects factors, and sizes of denominators as a within-subjects factor revealed a main effect of icon arrays ($p = .001$) and of graph literacy ($p = .014$), a marginally significant effect of sizes of denominators ($p = .062$), and interactions between icon arrays and sizes of denominators ($p = .001$) and between icon arrays and graph literacy ($p = .040$). The analysis did not yield a main effect of message frame or any interaction involving this factor.

[3] The dichotomization of a continuous measure by median split can have negative consequences such as the loss of statistical power. Therefore, we also performed logistic regressions using the full range of graph literacy scores as a predictor and accuracy as a dependent variable, for each of the four sizes of denominators conditions. In all cases a test of the model versus a model with intercept only was statistically significant when icon arrays were provided ($X^2 = 3.91 – 12.53$, $df = 1$, $p = .001 –.048$, OR = 1.26 – 1.59) but not when only numerical information was provided ($p > .20$ for all tests). These results are in line with those obtained in the ANOVA reported, where graph literacy was dichotomized via a median split.

graph literacy when only numerical information was provided was 6.9 (SE = .3), while it increased to 8.4 (SE = .3) when icon arrays were also presented. Instead, the mean confidence reported by participants with low graph literacy when only numerical information was provided was 6.8 (SE = .3), and 6.8 when icon arrays were provided (SE = .3).

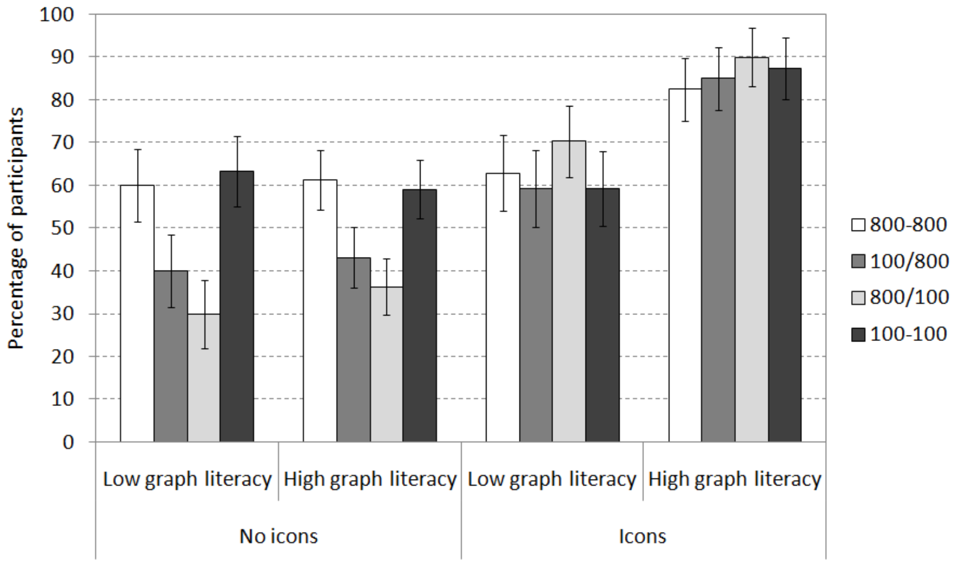


Figure 4. Percentage of participants whose estimates of risk reduction were accurate, as a function of graph literacy, icon arrays and sizes of the denominators. Error bars represent one standard error

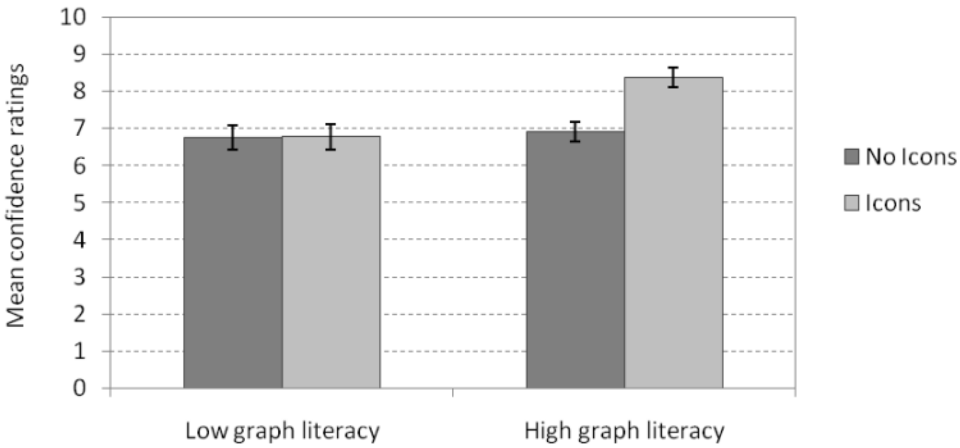


Figure 5. Mean confidence in risk reduction estimates, as a function of graph literacy and icon arrays. Error bars represent one standard error of the mean

Discussion

A precise understanding of numerical information is essential to accurate judgment and decision-making in many contexts. However, numerical concepts can be subject to biases and errors that undermine judgment and decision making. In this paper, we sought to document some of the ways that individual differences in graph literacy can affect the extent to which people benefit from visual aids designed to overcome a common judgment bias (denominator neglect).

Consistent with previous research (e.g., Garcia-Retamero & Galesic, 2009; Garcia-Retamero et al., 2010), we demonstrated that people show denominator neglect when judging the effectiveness of a treatment using information from unequally sized groups of treated and non-treated patients. These results support the idea that ratio concepts are particularly hard to understand (Bonato, Fabbri, Umiltà, & Zorzi, 2007; Ni & Zhou, 2005), and that people tend to behave as if they only compare magnitudes across numerators, thereby neglecting the denominators (Denes-Raj et al., 1995; Reyna, 2004; Reyna & Brainerd, 2008). We also found that icon arrays can help to reduce the effect of denominator neglect. This finding supports the idea that graphical displays are an effective method to reduce judgment biases that can help people to make decisions based on an accurate understanding of risk information. Thus, we support and extend our own and others' previous findings indicating that visual aids often facilitate risk communication in the health domain. Specifically, icon arrays symbolizing individuals who are affected by some risk using circles (Galesic, Garcia-Retamero, & Gigerenzer, 2009; Garcia-Retamero et al., 2010), squares (Zikmund-Fisher et al., 2008) or human figures (Paling, 2003) have been shown to make medical risks easier to interpret. Grids with squares representing visually the specificity or sensitivity of a medical test can also help people to correctly update posttest probabilities (Lloyd & Reyna, 2001). Visual aids can also improve understanding of risks associated with different medical treatments, screenings, and life-styles (Ancker et al., 2006; Galesic et al., 2009; Garcia-Retamero & Galesic, 2010b; Lipkus, 2007), promote consideration of beneficial treatments that have side effects (Waters, Weinstein, Colditz, & Emmons, 2007) and eliminate errors induced by anecdotal narratives (Fagerlin, Wang, & Ubel, 2005).

Additionally, we found that the increase in the percentage of participants providing accurate estimates when icon arrays were provided was larger when the sizes of the denominators were different, than when they were equal. Considering the information in the denominators is crucial to calculate risk reduction accurately when the sizes of denominators differ (Garcia-Retamero & Galesic, 2009; Garcia-Retamero et al., 2010). Thus, interventions such as icon arrays that help people to take denominators into account can be particularly effective in such cases.

Our results extended previous research in at least two other notable ways. First, we demonstrated that individual differences in graph literacy can moderate the effectiveness of icon arrays in increasing accuracy of risk understanding. Icon arrays helped to reduce denominator neglect both for participants with high and low graph literacy. However, the increase in risk comprehension was larger for highly graph-literate participants than for less graph-literate participants. Second, we established that individual differences in graph literacy can affect people's confidence in their estimates of risk reduction. The provision of icon arrays resulted in a significant increase in highly graph-literate participants' self-reported confidence in their own estimates, while less graph-literate participants' confidence was not reliably affected. Finally, our results also provided evidence inconsistent with the suggestion that an attentional bias might exacerbate the effect of denominator neglect when information is framed negatively, given that the percentage of accurate responses did not vary as a function of the type of framing provided. These results suggest that denominator neglect can hold consistently, independent of whether risk information is presented in positive or negative terms.

Why do visual aids improve accuracy in risk understanding? Potential explanatory mechanisms and open questions for future research

Theoretical frameworks such as the Fuzzy Trace Theory (e.g., Brainerd & Reyna, 1990; Reyna & Brainerd, 1995; Reyna, et al., 2009) have been used to provide an explanation of the potential mechanisms underlying the effects of icon arrays on denominator neglect. According to previous data, the problems that people face to understand ratio concepts stem from the fact that the references of classes overlap, which leads to class-inclusion errors (Brainerd & Reyna, 1990; Reyna, 1991). Thus, denominator neglect would be produced by people's tendency to focus on the target classes in numerators, thereby neglecting classes in denominators. Previous research suggests that manipulations that contribute to disentangle classes (e.g., icon arrays) can help to reduce biases such as denominator neglect as a function of shifting processing from more verbatim based to gist based representation of set structures (Brainerd & Reyna, 1990; Reyna et al., 2009).

A related explanation of the power of icon arrays to reduce denominator neglect was put forward by Stone et al. (2003; see also Ancker et al., 2006). Stone and colleagues suggested that Yamagishi's (1997) findings illustrating denominator neglect can be explained in terms of the saliency of foreground information (e.g., number of people harmed, or subset) vs. background information (e.g., the number of people at risk, or superordinate set). Thus, according to Stone et al., the numerical presentation of information can lead to a focus on foreground information, while graphical formats displaying both foreground and background information—pie charts and stacked bar graphs—contribute to bringing people's attention to the background too.

Our findings are compatible with the hypotheses that (1) icon arrays contribute to disentangle classes and (2) icon arrays bring people's attention to background information. As a consequence these kinds of displays are likely to help people to overcome denominator neglect in a variety of common situations. The mechanisms outlined by Brainerd and Reyna (1990; see also Reyna et al., 2009) and by Stone et al. (2003) anticipate the effectiveness of icon arrays in reducing denominator neglect observed in our study. However, the interactions between icon arrays and graph literacy obtained here (for both accuracy and confidence) suggest that (i) the power of icon arrays to increase the accuracy of risk reduction estimates is larger for highly graph-literate individuals than for less graph-literate ones, and that (ii) the subjective perceptions of individuals with low graph literacy (i.e., confidence ratings) are not necessarily influenced by the presence of icon arrays. These findings are compatible with our hypothesis that a certain level of graph literacy can be necessary in order to associate the visual patterns contained in icon arrays with meaningful interpretations of the data represented (i.e., risk reduction information).

Future work should directly aim to trace attentional and cognitive processes underlying the effect of icon arrays in individuals with different levels of graph literacy. The use of process tracing methodologies such as eye-tracking or verbal protocol analysis would provide a more nuanced understanding of the time course and operations involved for participants of varying skill levels. For example, the analysis of eye movement data would assess the proportional fixation times on the circles in the icon arrays representing number of patients who died (foreground information) vs. on circles representing number of people at risk (background information). This data might reveal differences in the saliency of foreground vs. background information in icon arrays for different viewers. Additionally, process tracing would enable the testing and refining of higher fidelity cognitive process theories that explain or, more importantly, predict how various kinds of displays are processed by different viewers. This would expand previous research that has documented graph comprehension processes (e.g., encoding of the visual pattern, translation of visual features into conceptual relations) mainly in homogeneous viewers, focusing on displays which include features such as axes or scales (e.g. line plots or bar charts; Carpenter & Shah, 1998; Lohse, 1993, Pinker, 1990).

From a translational or applied standpoint, process tracing is an essential step in efforts to facilitate the development of training methods for individuals with low graph literacy. These methods could be based in part on the processes that highly graph-literate individuals follow to understand these kinds of displays. That is, process-tracing studies allow for ‘reverse engineering’ of superior performance by revealing encoding or search strategies of successful individuals that may confer benefits to those participants who do not yet use such strategies (Cokely, Kelley, & Gilchrist, 2006; Cokely & Kelley, 2009). Moreover, an understanding of the encoding and search processes of low performing individuals may also provide clues for the design of environments that facilitate more appropriate search and representation. Nevertheless, many open questions remain concerning when and for whom simple differences in encoding or representational strategies (e.g., gist based representation) would be sufficient for improved performance. Ongoing research is investigating these issues (Okan, Galesic, & Garcia-Retamero, 2010).

Theoretical and practical implications

Taken together, our findings show that visual aids do not necessarily facilitate risk comprehension to the same extent for everyone. Our results emphasize the importance of considering the fit between (i) persons, (ii) cognitive processes, and (iii) task environments when designing interventions such as visual aids. Individual differences in graph literacy moderate the effect of such visual aids, affecting the accuracy of risk judgments. Similarly, Parker and Fischhoff (2005) identified a set of tasks that capture four basic skills required by competent decision makers. One of these skills (*belief assessment*) refers to people’s ability to judge the probability of occurrence of events. Our data indicate that the accurateness of this kind of judgment may be affected by variations in graph literacy, suggesting that graph literacy is usefully characterized as a cognitive skill that influences competent decision making.

A growing body of research has documented a variety of individual differences that influence decision making performance. These include domain general decision making skills such as those identified by Parker and Fischhoff (see also Bruine de Bruin et al.,2007; Finucane, Mertz, Slovic, & Schmidt, 2005) or skills such as numeracy (Peters & Levin, 2008; Peters et al., 2006). Other individual differences that can have an effect on the quality of risky judgment and decision making include decision making styles (Baron, 2000; Campitelli & Labollita, 2010; Frederick, 2005; Shiloh, Salton, & Sharabi, 2002), specific expertise (Ericsson, Prietula, & Cokely, 2007; Garcia-Retamero & Dhami, 2009; Shanteau, 1992), and domain general cognitive abilities (Cokely & Kelley, 2009; Del Missier, Mäntylä, & Bruine de Bruin, 2010, 2012; Stanovich & West, 2000, 2008). Research indicates that general decision-making skills have significant relations amongst them and with other measures of cognitive abilities and styles (Bruine de Bruin et al., 2007; Del Missier et al., 2010; Parker & Fischhoff, 2005). Future research should aim to achieve a more precise specification of the relations between graph literacy, the set of individual differences outlined above, and the cognitive processes that mediate differences.

The current findings also provide additional support for the predictive validity of the scale developed by Galesic and Garcia-Retamero (2011b; see also Garcia-Retamero & Galesic, 2010b). The evidence obtained so far suggests that this scale can be useful to predict performance in tasks involving not only icon arrays, but also other kinds of graphs. Future work should examine the extent to which graph literacy moderates performance in tasks involving displays such as pie charts or line plots.

Future research should aim to enhance (1) the generalizability, and (2) the ecological validity of the present study. Concerning the first point, it should be noted that the instrument and the materials used focused on the medical domain. Thus, the extent to which graph literacy can influence risky decision-making performance in other important domains, such as finance or politics, is not yet well documented or understood. However, ongoing studies do seem to suggest that graph literacy will have some predictive validity across diverse domains. Concerning ecological validity, the fact that our experiments were not conducted in a clinical setting prompts us to suggest some caution regarding immediate prescriptive applications of our findings for medical practice. Research has shown that the effect of manipulations observed in the lab, such as framing of information in the context of risk communication, may not be generalizable to clinical practice (Edwards et al., 2001). Thus, future research should aim to provide more converging evidence on the effect of graph literacy in the ecology of interest (e.g., in the clinic).

Conclusion

In the present article we have demonstrated that individual differences in graph literacy can moderate the magnitude of the effect of visual risk communication interventions (i.e., icon arrays). This finding is relevant to modern societies, where graphical displays are increasingly being used and recommended for the communication of risks to the public (Ancker et al., 2006; Fuller et al., 2002; Lipkus, 2007). However, these graphs are rarely designed on the basis of a systematic set of principles of good graph construction. As a consequence, data represented in graphs is frequently misinterpreted by viewers (Beattie & Jones, 2002; Cooper, Schriger, Wallace, Mikulich, & Wilkes, 2003). This can lead to substantial alterations in viewers' preferences and decisions (Arunachalam, Pei, & Steinbart, 2002). Principles of good graph design have been put forward in some cases (Cleveland, 1994; Cleveland & McGill, 1984; Kosslyn, 2006; Tufte, 2001), but a consensus does not always exist regarding the adequacy or effective implementation of these principles. For instance, Kosslyn (2006) argues that it is adequate to use bar charts in which part of the y axis has been removed, provided that marcs are used to indicate discontinuities. In contrast, other authors suggest that the scale in the y axis of bar charts should start at 0, in order for the proportion between the lengths of the bars to reflect the proportion between the quantitative data (e.g., Cleveland, 1994).

There is a need for a unified and usable set of standards for guiding graph designers' work. We suspect the timing is right to work toward refining the available guidelines that have been developed across several fields (e.g., cognitive psychology, mathematics, human factors). Resulting standards could play an important role in inoculating professionals, policy makers, and the general public against potentially distorted and misleading communications. As highlighted by the current data, graph design standards should allow for custom-tailored designs that are sensitive to the various needs and abilities of diverse individuals. Additionally, the scope of the effect of graph literacy on a range of different judgment and decision making tasks should be investigated, along with relations to other decision making competences. To an important extent, such goals hinge on our ability to achieve a more unified understanding of the theoretical and practical relevance of the construct of graph literacy, including documentation of both construct and predictive validity of the instruments used to measure it. This work holds the promise of important theoretical and translational benefits such as improved decision making in medicine, business, and potentially many other domains involving the communication of risk.

Acknowledgments

This study is part of the projects, “How to Improve Understanding of Risks about Health (PSI2008-02019),” “Helping Doctors and Their Patients Make Decisions about Health,” (PSI2011-22954) and “Decision Making and Risk Behavior” (PSI2009-12217), funded by the Ministerio de Ciencia e Innovación (Spain). We would also like to thank three anomymous reviewers for their valuable comments on this work.

CHAPTER III.

ENHANCING ACCURACY OF RISK COMPREHENSION USING DYNAMIC ICON ARRAYS

Enhancing Accuracy of Risk Comprehension Using Dynamic Icon Arrays

Abstract

Icon arrays have been found to improve risk understanding and reduce biases in risk perception across a wide range of studies. However, recent research has shown that individuals with low graph literacy benefit from icon arrays to a lesser extent than those with high graph literacy. In an experiment, we examined the effectiveness of five different types of dynamic icon arrays designed to improve understanding of treatment risk reduction. The different types of displays were designed to promote specific cognitive processes involved in the comprehension of icon arrays. Results showed that, compared to a set of static icon arrays, performance improved only for a display including a reflective question (i.e., an estimate about the information displayed), followed by accuracy feedback provided visually. This improvement in performance was achieved even among less graph literate individuals. This suggests that encouraging more active, elaborative processing of information may be essential for enhancing the beneficial effects of icon arrays. Findings contribute to our understanding of the processes underlying superior performance and have prescriptive implications for graphical communication of medical information.

Submitted as:

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (submitted). Enhancing accuracy of risk comprehension using dynamic icon arrays.

Introduction

Informed medical decision making requires understanding information about risks, benefits, and drawbacks of different treatments. However, patients often struggle with numerical concepts that are prerequisites for the accurate evaluation and communication of risks (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Lipkus, Samsa, & Rimer, 2001; Peters et al., 2006; Schwartz, Woloshin, Black, & Welch, 1997). Research indicates that visual aids can be highly effective decision support tools that can reach vulnerable populations (Garcia-Retamero & Cokely, in press). In particular, icon arrays (i.e., graphical representations consisting of a number of circles or other icons symbolizing individuals who are affected by some risk; Ancker, Senathirajah, Kukafka, & Starren, 2006; Edwards, Elwyn, Covey, Matthews, & Pill, 2001; Paling, 2003) have been found to improve risk understanding and reduce biases in risk perception across a wide range of studies (Galesic, Garcia-Retamero, & Gigerenzer, 2009; Garcia-Retamero & Galesic, 2009; Lloyd & Reyna, 2001; Zikmund-Fisher et al., 2008). In some cases, icon arrays enhance risk comprehension to a larger extent than other visual displays, including bar graphs and pie charts (Hawley et al., 2008; Waters, Weinstein, Colditz, & Emmons, 2007). Moreover, according to theories on elemental perceptual tasks in graph comprehension, the perceptual features involved in processing icon arrays are among those that are easiest to process (Price, Cameron, & Butow, 2007).

Unfortunately, icon arrays are not equally beneficial for everyone. People in the general population differ substantially in their *graph literacy*—the ability to understand graphically presented information (Freedman & Shah, 2002; Galesic & Garcia-Retamero, 2011b; Kutner, Greenberg, Jin, & Paulsen, 2006), and individuals with low graph literacy have been found to profit from icon arrays to a lesser extent than more graph literate individuals (Garcia-Retamero & Galesic, 2010; Okan, Garcia-Retamero, Cokely, & Maldonado, 2012). Recent studies have started to shed some light on the underlying cognitive processes that account for differences in performance with bar graphs and line graphs (Okan, Galesic, & Garcia-Retamero, 2013; Woller-Carter, Okan, Cokely, & Garcia-Retamero, 2012). However, the reasons underlying the reduced efficacy of icon arrays among individuals with low graph literacy are not well understood. It is also unclear if and how the manipulation of specific design features of icon arrays can improve risk comprehension among such individuals. Accordingly, in this paper we sought to examine the effectiveness of icon arrays including different types of dynamic features, designed to promote specific processes involved in the comprehension of icon arrays. Our aim was to determine how to enhance risk comprehension among less graph literate individuals, and at the same time to better understand the cognitive processes underlying superior performance.

Prominent graph-comprehension models have identified three global processes involved in making inferences from graphical displays, such as line or bar graphs (e.g., Carpenter & Shah, 1998; Kosslyn, 1989; Lohse, 1993; Pinker, 1990; Shah & Carpenter, 1995; Simkin & Hastie, 1987). The first is *encoding the visual pattern* to identify the principal features in graphs. Encoding the visual pattern involves making different visual judgments of the elements (e.g., judgments of position along a scale, length, or angle; Cleveland & McGill, 1986). The second process is the *translation of the identified visual features into conceptual relations*. For example, variations in the size of spatial features (e.g., bars of different heights) can be used to indicate variations in the quantity of the variables represented. The third process involves *determining the referents of the concepts identified* by associating them with the specific variables shown in the graph and their numerical values (Carpenter & Shah, 1998; Shah & Carpenter, 1995). To the best of our knowledge, such processes have not yet been studied in the context of the comprehension of icon arrays. We believe they provide a useful framework for investigating the cognitive dynamics that may be at play when individuals with varying levels of graph literacy view icon arrays.

Theoretically, the limited ability of less graph literate individuals to benefit from icon arrays could arise from at least three different mechanisms. One possibility is that differences in performance between individuals with low and high graph literacy arise primarily from differences in attention and encoding of the visual pattern (i.e., first process involved in graph comprehension). Individuals with low graph literacy may be less comfortable with graphs and so they may be less likely to attend to the icon arrays, as compared to highly graph literate individuals. This effect may be magnified when relevant information can be extracted both from the text and from graphs, as less graph literate individuals may spend more time focusing on the numerical and text-based information, largely avoiding the icon arrays and their benefits (see also Gaissmaier et al., 2012). Additionally, even when such individuals attend to the icon arrays, they may allocate insufficient attention to some regions of the arrays (see Keller & Siegrist, 2009, for a discussion concerning the saliency of different regions). If less graph literate individuals do not allocate sufficient attention to icon arrays or to parts of them, this could reduce the potential of such displays to enhance quantitative reasoning among these individuals. Accordingly, a display that directs viewers' attention to the different regions of the arrays should enhance comprehension.

A second possibility is that impaired performance among less graph literate individuals reflects difficulties with identifying the referents in icon arrays (i.e., third process involved in graph comprehension; Carpenter & Shah, 1998).[4] For line graphs and bar charts, this process involves identifying and inferring information from elements determined by arbitrary graph conventions, including titles of graphs, axis labels, legends, and numerical values on the scales, and integrating this information with that extracted in the first two processes. For icon arrays, the identification of the referents in some cases will also involve integrating information in accompanying numerical scales or labels (e.g., Feldman-Stewart, Brundage, & Zotov, 2007; Zikmund-Fisher et al., 2008). However, as specific quantities in icon arrays can also be inferred by counting individual icons (Hess, Visschers, & Siegrist, 2011), such displays are often presented without accompanying scales or labels. In such cases, identifying the referents may require integrating information contained in accompanying numerical and text-based information. If individuals with low graph literacy have difficulties engaging in such integration processes, then the addition of descriptive labels accompanying each region of the array could improve performance. This addition may be particularly useful if labels are presented in a sequential manner, following the presentation of each region of the array. Such presentation format should contribute to direct attention to the elements that enable participants to identify the referents in icon arrays, and at the same time promote the integration of information in such elements with the information represented in the different regions of the arrays.

A third possibility is that less graph literate individuals benefit less from static icon arrays because they fail to engage in an *active* processing of the information depicted. That is, even if they engage successfully in all the graph comprehension processes outlined above, they may process the information in a more passive fashion, engaging to a lesser extent in relational and elaborative processing. Research indicates that the link

[4] Translations of visual features into conceptual relations (i.e., second process) are often nonarbitrary and cognitively constrained, as certain associations (e.g., "higher equals more") emerge consistently in adults and children with no graphing experience (Gattis, 2002, 2004). This entails that spatial features often can convey meaning independent of viewers' level of graph literacy (Gattis, 2002, Okan, Garcia-Retamero, Galesic, & Cokely, 2012; Tversky, 2001). In contrast, identifying the specific variables and numerical values associated (i.e., third process) can require specific graph-related knowledge (Okan et al., 2013a).

between domain-general cognitive abilities and superior risky decision making can be fully mediated by differences in meaning-based elaborative processing (Cokely & Kelley, 2009). Individuals with low graph literacy may be less likely to engage in this kind of active processing, and as a consequence may have more surface-level representations of the problem (see Kintsch, 1988, 2004, for related arguments). Such representations would make it difficult to generalize the risk information, beyond the specific task or set of icon arrays (e.g., understanding that the relative risk reduction associated with a treatment is 80%). In contrast, more graph literate individuals may be more likely to engage in an elaborative processing of icon arrays, leading to richer, better integrated representations in long-term memory, which better support subsequent task performance (Cokely, Kelley, & Gilchrist, 2006; Cokely & Kelley, 2009; Ericsson & Kintsch, 1995; Mitchum & Kelley, 2010; Vigneau, Caissie, & Bors, 2005).

If individuals with low graph literacy are less likely to engage in an active processing of the information depicted, then their performance could be enhanced by the use of displays that promote such kind of processing. One means to do so is to encourage people to construct their own representations, instead of having them interpret presented representations (Cox, 1999; Stern, Aprea, & Ebner, 2003). For instance, requiring participants to fill in or to circle the appropriate number of squares in an array (Cosmides & Tooby, 1996), or to portray the size of a risk on a bar chart (Natter & Berry, 2005) has been shown to significantly improve understanding. However, active constructions of icon arrays may sometimes not be correctly completed (Brase, 2009), or may lead people to devote a large amount of attention and resources on understanding what the task requires (Zikmund-Fisher, Dickson, & Witteman, 2011). To have a beneficial effect, displays should be constructed correctly (Cox, 1999).

An alternative means to promote an active processing is to require people to answer a *reflective* question concerning the information depicted (e.g., 'How many out of 100 people who take this medication will experience one or more side effects?'), prior to the completion of the task (Natter & Berry, 2005). Reflective questions require the generation of a solution that is not explicitly represented in the display. Hence, such questions can encourage a more elaborative processing which may induce the steps required to solve the task at hand, and to apply that knowledge to new tasks (Lee & Hutchison, 1998; see Jacoby, 1978, for related arguments concerning the effect on retention of constructing vs. remembering a solution). In the context of the present research, this may be accomplished by presenting a second set of icon arrays, and posing a reflective question concerning this set of arrays, to be answered on the basis of the information depicted in the first set. This could stimulate an active processing, while avoiding the potentially confusing requirement of portraying a given risk on the icon array itself (see, e.g., Zikmund-Fisher et al., 2011).

Two additional aims of the present work were to examine the effect of the different types of dynamic icon arrays on subjective confidence, as well as on subjective preferences (e.g., perceived usefulness of the displays; Zikmund-Fisher et al., 2012). Concerning confidence, previous research has shown that static icon arrays were associated with higher levels of confidence in risk reduction estimates among participants with high graph literacy, but not among those with low graph literacy (Okan et al., 2012a). This indicates that greater confidence was matched by higher accuracy of risk understanding, consistent with findings indicating that effective representations can enable effective monitoring even among participants with considerable differences in cognitive abilities (Mitchum & Kelley, 2010; see also Garcia-Retamero & Dhami, 2011). Of note, however, knowledge and confidence judgments are not necessarily linked as correlations are often only modest (e.g., Parker, Bruine de Bruin, Yoong, & Willis, 2012). Poor correspondence between confidence and knowledge can lead to worse real-world decision outcomes, including inappropriate risky behavior (Bruine de Bruin, Parker, & Fischhoff, 2007; Parker et al., 2012). Therefore, we sought to examine whether any variations in accuracy of risk understanding linked to individual differences in graph literacy or to the type of display, are also reflected in variations in confidence.

Finally, concerning subjective preferences, a number of studies have shown that graph evaluations are not always aligned with performance (Feldman-Stewart, Kocovski, McConnell, Brundage, & Mackillop, 2000; McCaffery et al., 2012; Micallef, Dragicevic, & Fekete, 2012; Waters, Weinstein, Colditz, & Emmons, 2006). To illustrate, shaded icons can be preferred to unshaded ones, even though shading can in some cases be detrimental for response times (Price et al., 2007). Thus, we also sought to examine how individuals with low and high graph literacy evaluate different types of dynamic icon arrays, and whether variations in evaluations are matched with variations in performance.

Experiment

We conducted an experiment including different medical scenarios depicting the usefulness of different treatments (i.e., hypothetical drugs). Past research examining people’s understanding of risk reduction after treatment has often used samples of treated and non-treated patients of the same size (Fagerlin, Ubel, Smith, & Zikmund-Fisher, 2007; Galesic et al., 2009). However, this is not representative of the type of information that people normally encounter, as groups of treated and non treated patients are often unequally sized in medical practice (e.g., Grossarth-Maticek & Ziegler, 2008; Lichtenberg, Levinson, Sharshevsky, Feldman, & Lachman, 2008). In such cases, understanding treatment risk reduction can become particularly challenging, given that people often pay too much attention to numerators in ratios (i.e., the number of times a target event has happened) and insufficient attention to denominators (i.e., the overall number of opportunities for it to happen; Denes-Raj, Epstein, & Cole, 1995; Reyna, 2004; Reyna & Brainerd, 2008). In the current study we investigated how the different design features of icon arrays outlined above affect people's risk understanding both in medical scenarios involving unequal and equal denominators (see also Garcia-Retamero and Galesic, 2009; Okan et al., 2012a). We manipulated the *type of icon arrays* between-subjects, and compared performance with different types of dynamic icon arrays to that with a static set of icon arrays.

Method

Participants.

Participants were 458 undergraduate students recruited from the University of Granada (307 female), with a mean age of 20.9 years (*SD* = 4.7, range 17–60 years). All participants signed an informed consent form to participate in the experiment, and received course credit.

Materials and design.

Medical risk scenarios. Participants were presented with four medical scenarios describing the usefulness of hypothetical new drugs for reducing cholesterol that also decreased the risk of dying from a heart attack. The scenarios involved equally effective treatments but differed in the overall number of treated and non-treated patients (i.e., sizes of denominators). These were set to be either, 800–800, 100–800, 800–100, or 100–100, where the first and second quantities reflected the overall number of patients who did and did not take the drug, respectively.[5] The sizes of the numerators (i.e., the number of treated and non-treated patients who died) varied depending on the size of the denominator. The treatment always had an 80% relative risk reduction (i.e., from 10% to 2%). Appendix A shows the combinations of numerators for all denominator sizes. An example of the information presented in the medical scenarios is as follows (the original material was in Spanish): “A new drug that reduces cholesterol, Benofreno, decreases the chances of dying after a heart attack for people with high cholesterol. Here are the results of a study of 900 people with high cholesterol: 80 out of 800 people who did not take the drug died after a heart attack, compared to 2 out of 100 people who took the drug.” The rest of the drugs were named Denofreno, Cenofreno, and Fenofreno, respectively. This information was always presented on the top part of the screen.

Icon arrays. We generated one static and five types of dynamic icon arrays. The icon arrays were presented below the text including numerical information for each medical scenario, and depicted the risk of dying of a heart attack when the drug was and was not taken. For each set of icon arrays, the proportion of non-treated people who died was depicted in one array (*top icon array*), and the proportion of treated people who died was depicted in another array presented immediately below (*bottom icon array*). The group of icons representing the overall number of people at risk (i.e., *background* of the array) was represented with white circles, and the group of icons representing the people who died (i.e., *foreground* of the array) was represented with black circles at the end of the array (see Figure 1d).

Participants in the control (*static*) condition viewed all the information simultaneously (i.e., numerical information and the two icon arrays). In all dynamic conditions, the numerical information was presented first, and different elements were incorporated one at a time. The top icon array appeared before the bottom icon array, and the background of each array appeared before the foreground. The elements of icon arrays appeared following user clicks on *Continue*. In the first type of dynamic icon array (the *sequential* condition), the background and the foreground were displayed in a sequential manner, following user clicks. This type of display was developed to encourage the allocation of attention and encoding of all regions of icon arrays (i.e., first process involved in graph comprehension). The second type of icon array (the *labeling* condition) was equivalent to the previous one, with the exception that an explanatory label appeared

[5] Participants in some conditions viewed one set of icon arrays corresponding to each scenario, while others viewed two sets of arrays for each scenario, as will be described below.

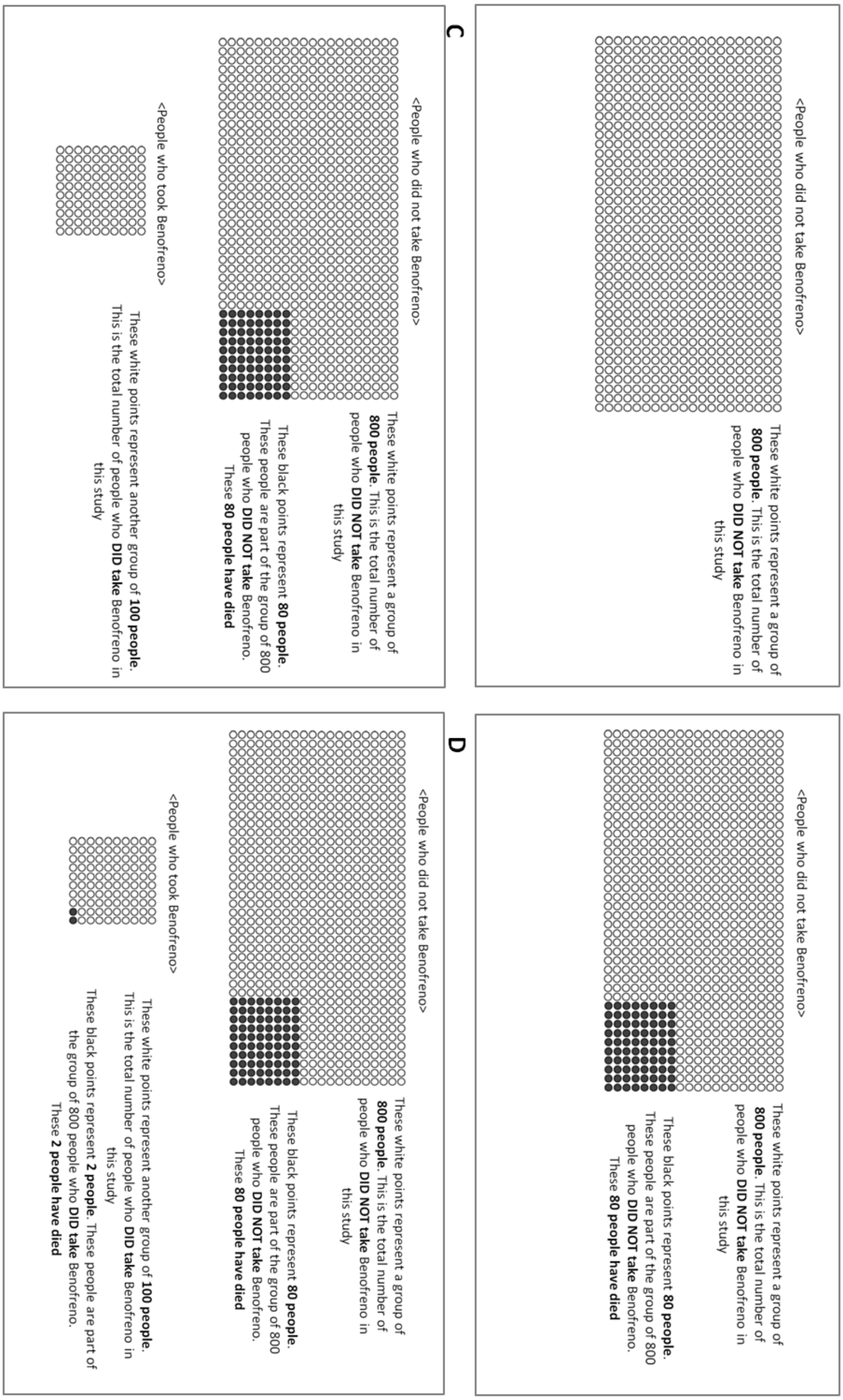


Figure 1. Example of presentation sequence for the labeling condition.

next to each region indicating what it represented. Labels appeared on the right side of each part of the array, one second after that part was displayed (see Figure 1). This type of display was developed to encourage the allocation of attention to elements that enable one to identify referents of icon arrays, as well as to promote the integration of such information with that depicted in icon arrays (i.e., third process involved in graph comprehension).

The third type of display (the *transfer estimate* condition) included two sets of icon arrays for each medical scenario, resulting in a total of eight sets of arrays. For each scenario, the second set was displayed following the first one, and showed the results of a second study conducted with the same drug. Treatment risk reduction was also 80%, but both denominators and numerators were halved (see Appendix A). For both the top and the bottom icon array of the second set, participants were required to answer a reflective question concerning the number of people who died before this information was presented visually. In particular, participants were told: 'please estimate the number of people who would die in the group of people who did not take/took the drug.' As advanced above, this display was constructed to promote an active processing of the risk information, leading to a generalization of the risk information depicted.

Two more types of icon arrays were developed to exclude potential confounds of any effect observed for the transfer estimate condition. In particular, one display (the *labeling reproduce* condition) was identical to the labeling condition, with the exception that participants did not only view the labels but were also required to restate the information shown (e.g., 'How many people died in the group of people who did not take the drug?). This question differs from that included in the transfer estimate condition in that it does not require participants to estimate information which has not yet been displayed. If transfer estimate affects performance merely because it requires participants to answer a question concerning the information depicted (and not because it requires them to provide a reflective question involving an active estimate), a similar effect should be observed for labeling reproduce and transfer estimate. Finally, the *transfer reproduce* condition was equal to transfer estimate, with the exception that participants were not required to answer questions concerning the second set of icon arrays. If transfer estimate affects performance because it includes two sets of icon arrays instead of one, then a similar effect should be observed for transfer reproduce. Figure 2 shows a schematic summary of all the types of dynamic icon arrays.

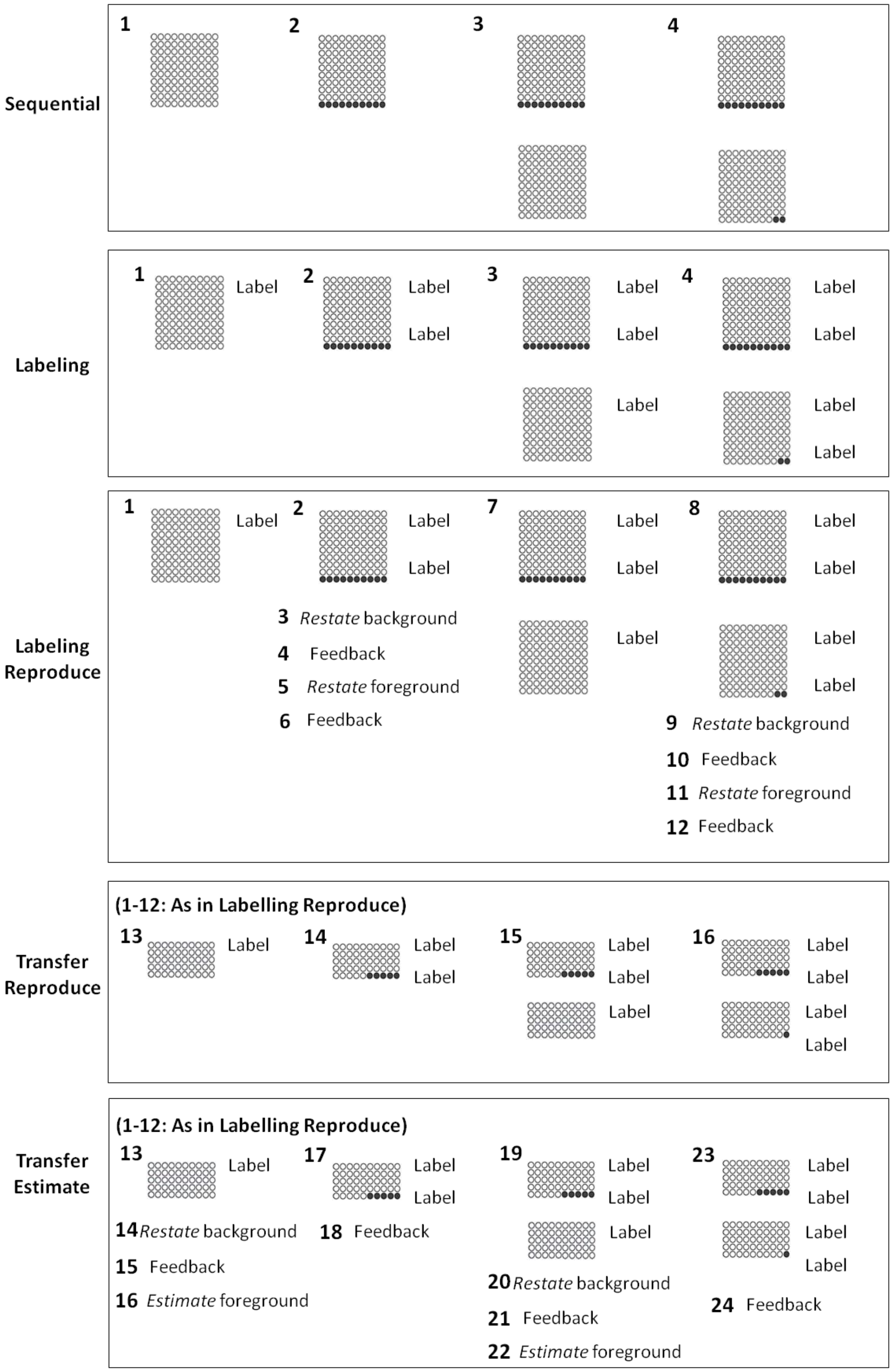


Figure 2. Schematic summary of the different types of dynamic icon arrays.

In the conditions requiring that participants answer questions concerning the information depicted in the icon arrays (i.e., labeling reproduce, transfer reproduce, and transfer estimate), questions were presented on the bottom part of the screen while the icon array was visible, and participants were instructed to type in a response. Following a participant's response, the text of the question was replaced by feedback indicating whether the response was correct or incorrect. Additionally, the correct value was highlighted in red on the label which contained the information enquired about.[6] Appendix B includes the full instructions provided to participants, the text of explanatory labels presented alongside icon arrays, of the questions asked in each case, and of the feedback provided.

Outcome measures. As dependent variables, we measured (1) accuracy of risk understanding after reading the information provided for each medical scenario, (2) subjective confidence in estimates, and (3) evaluations of the icon arrays. In order to measure accuracy of risk understanding, we followed the procedure used by Schwartz, Woloshin, Black and Welch (1997). First, participants were asked how many of 1000 patients with high cholesterol might die of a heart attack if they do not take the drug. Second, they were asked how many of 1000 patients with high cholesterol might die of a heart attack if they did take the drug. By subtracting the second from the first answer and dividing it by the first, we calculated the estimated relative risk reduction. Estimates were treated as correct only when they were exactly right (i.e., 80%; see, e.g., Garcia-Retamero & Dhami, 2011; Garcia-Retamero, Galesic, & Gigerenzer, 2010 for a similar procedure). Participants were then classified depending on whether their estimates were correct in both scenarios with different denominators (100–800 and 800–100), and in both scenarios with the same denominators (100–100 and 800–800).

Participants' degree of confidence in their estimates was measured on a scale from 1 to 10, were 1 represented "not at all confident" and 10 represented "very confident." For each participant, confidence ratings were averaged across scenarios with different denominators and with the same denominators. Finally, icon array evaluations were measured using the procedure reported by Zikmund-Fischer et al. (2012). Specifically, following the presentation of all medical scenarios, three items were included asking participants to rate how well the graphs described the risk of dying of a heart attack, how helpful the graphs were, and whether the participant would like to seek information in this

[6] In the transfer estimate condition, the foreground and accompanying label appeared *prior* to the feedback for the reflective question, as these elements had not been presented previously (see Figure 2).

type of graph. All items included a response scale of 1–10. These three ratings were combined based on the average rating, following Zikmund-Fischer et al. (2012). Cronbach's alpha for the 3 graph evaluation items was .78.

Measurement of graph literacy. Graph literacy was measured using the scale developed by Galesic and Garcia-Retamero (2011b). This scale consists of 13 items dealing with the communication of medical risks, treatment efficiency, and prevalence of diseases, and covers four frequently used graph types—line plots, bar charts, pies, and icon arrays. We split participants into two groups according to the median graph literacy score for the total sample (i.e., 10 of the total 13). This enabled us to examine the effect of the different types of icon arrays, separately for individuals with low and high graph literacy (see Zikmund-Fisher et al., 2012, for a similar procedure). The group of participants with low graph literacy included those who obtained 10 or fewer correct responses (n = 285, mean score 8.5, SD = 1.5), while the group of participants with high graph literacy included those who obtained 11 or more (n = 173, mean score 11.5, SD = .6).

Measurement of numeracy. As understanding treatment risk reduction involves comparing and transforming different proportions, we controlled for the effect of numeracy (i.e., the ability to understand and manipulate different numerical expressions of probability; Lipkus et al., 2001; Peters, 2012). To this end, we used the 11 items included in the general and expanded numeracy scales developed by Lipkus et al. (2001), as well as the Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012).

Procedure.

The questionnaire was administered in the laboratory of the University of Granada, and all materials were implemented as an electronic survey in Unipark (www.unipark.de). In addition to the experiment presented here, the session included other unrelated tasks concerning medical risks, and took approximately 50 minutes to complete. The tasks relevant for the present study took between 25 and 30 minutes. First, participants signed a consent form and the instructions were presented on the screen. They were then presented with the medical scenarios. Time to read the scenarios and to answer the questions was unlimited. Finally, they completed: (1) the Graph Literacy Scale, (2) the Berlin Numeracy Test, (3) the Lipkus numeracy test, and (4) some demographic questions. Participants were assigned randomly to the different experimental conditions.

Results

Do graph literacy and type of icon arrays affect risk understanding, for scenarios with equal and different denominators? To answer this question we conducted two logistic regressions predicting accuracy of risk understanding. Graph literacy and type of icon arrays were entered as predictors. The models were a strong predictor of accuracy, both for scenarios with equal, $X^2(6) = 45.85$, $p = .001$, Nagelkerke $R^2 = .13$, and different denominators, $X^2(6) = 46.77$, $p = .001$, Nagelkerke $R^2 = .13$. As expected, people with high graph literacy were significantly more likely than those with low graph literacy to provide correct responses (OR = 1.34, $p = .001$, for equal denominators, and OR = 1.36, $p = .001$, for different denominators), and type of icon arrays significantly predicted performance ($p = .028$ and $p = .020$ for equal and different denominators, respectively). Including the interaction term of graph literacy and type of icon array in a separate step did not improve predictions of accuracy either for scenarios with equal denominators, $X^2(5) = .92$, $p = .969$, Nagelkerke R^2 change = .04, or different denominators, $X^2(5) = 4.29$, $p = .508$, Nagelkerke R^2 change = .003. Additionally, McNemar's test of paired proportions revealed that the percentage of participants who provided correct estimates in scenarios with equal denominators was larger than the percentage of those who provided correct estimates in scenarios with different denominators, for all icon array conditions (ps < .002).[7]

How does type of icon array affect risk understanding, for individuals with low vs. high graph literacy? Table 1 shows the percentage of participants with low and high graph literacy who provided correct responses for the different types of icon arrays, as well as results of logistic regressions predicting accuracy in both groups of participants. In all cases, the static condition served as the reference category to evaluate the effect of type of icon array. The transfer estimate condition resulted in a large increase

[7] Consistent with these analyses, a 2 × 2 × 6 mixed ANOVA with size of denominator (equal vs. different) as within-subjects factor and graph literacy and type of icon array as between-subjects factors revealed a main effect of size of denominator, $F(1, 446) = 66.85$, $p = .001$, $\eta_p^2 = .13$, of graph literacy, $F(1, 446) = 31.91$, $p = .001$, $\eta_p^2 = .07$, and of type of icon array, $F(5, 446) = 3.13$, $p = .009$, $\eta_p^2 = .03$. Interactions between these factors were not significant. Additionally, in a hierarchical logistic regression, graph literacy continued to account for unique variance after controlling for numeracy, both for scenarios with equal denominators, $X^2(2) = 14.02$, $p = .001$, Nagelkerke R^2 change = .04, and with different denominators, $X^2(2) = 20.95$, $p = .001$, Nagelkerke R^2 change = .06.

Table 1. Percentage of correct responses in scenarios with equal and different denominators as a function of type of icon array, for participants with low graph literacy (top panel) and high graph literacy (bottom panel).

Low graph literacy						
	Equal denominators (R^2 = .08)			Different denominators (R^2 = .08)		
	% correct	OR (95% CI)	*P*	% correct	OR (95% CI)	*P*
Static	**41.5**	—	—	**24.5**	—	—
Sequential	**34.6**	.72 (.32, 1.61)	.42	**19.2**	.71 (.28, 1.83)	.48
Labeling	**50.0**	1.73 (.74, 4.06)	.20	**32.5**	1.8 (.71, 4.58)	.22
Labeling reproduce	**40.0**	1.06 (.44, 2.58)	.90	**28.6**	1.39 (.52, 3.70)	.52
Transfer reproduce	**45.1**	1.23 (.56, 2.73)	.61	**27.5**	1.24 (.51, 3.02)	.64
Transfer estimate	**59.3**	2.21 (1.00, 4.87)	.05	**44.4**	2.64 (1.14, 6.13)	.023

High graph literacy						
	Equal denominators (R^2 = .22)			Different denominators (R^2 = .26)		
	% correct	OR (95% CI)	*P*	% correct	OR (95% CI)	*P*
Static	**54.8**	—	—	**45.2**	—	—
Sequential	**70.0**	2.56 (.81, 8.07)	.11	**60.0**	2.6 (.83, 8.13)	.10
Labeling	**76.0**	2.67 (.77, 9.20)	.12	**52.0**	1.3 (.41, 4.16)	.65
Labeling reproduce	**56.7**	1.41 (.46, 4.32)	.55	**40.0**	1.04 (.33, 3.27)	.95
Transfer reproduce	**64.5**	1.73 (.57, 5.22)	.33	**58.1**	2.1 (.69, 6.41)	.19
Transfer estimate	**84.6**	6.49 (1.67, 25.23)	.007	**76.9**	6.49 (1.83, 23.06)	.004

Note: Nagelkerke R^2 is reported. Results are reported for logistic regression models controlling for numeracy.

in performance relative to the static condition, both for scenarios with equal denominators (OR = 2.21 and OR = 6.49, for participants with low and high graph literacy, respectively) and with different denominators (OR = 2.64 and OR = 6.49, for participants with low and high graph literacy). This indicates that the requirement to engage in an active processing of information was beneficial in all cases, even though the increase in odds was larger among more graph literate participants. The labeling condition was also associated overall with an increase in performance relative to the static condition; however this increase did not reach conventional levels of significance (see Table 1). Importantly, the labeling reproduce and transfer reproduce conditions did not result in a significant increase in performance. It is also worth noting that trends for the sequential condition were in opposite directions for participants with low vs. high graph literacy (i.e., a reduction in correct responses for the former, and an increase for the latter), even though differences were not statistically significant.

Do graph literacy and type of icon arrays affect subjective confidence, for scenarios with equal and different denominators? We conducted two linear regressions predicting confidence ratings, for scenarios with equal and with different denominators. As with accuracy, graph literacy, and type of icon arrays were entered as predictors. The full model predicted confidence for equal, $R^2 = .15$, $F(6, 451) = 12.71$, $p = .001$, and for different denominators, $R^2 = .16$, $F(6, 451) = 14.60$, $p = .001$. In line with the results for accuracy, higher graph literacy was associated with higher levels of subjective confidence ($t = 6.87$, p = .001, $\beta = .30$ for equal denominators, and $t = 7.44$, p = .001, $\beta = .32$, for different denominators), and type of icon arrays significantly predicted confidence (*p*s < .001). Including the interaction term of graph literacy and type of icon array in a separate step did not improve predictions of accuracy (R^2 change = .002, $p > .9$, and R^2 change = .003, $p > .8$ for equal vs. different denominators, respectively).[8]

[8] In line with these analyses, a 2 × 2 × 6 mixed ANOVA with size of denominator (equal vs. different) as within-subjects factor and graph literacy and type of icon array as between- subjects factors on confidence ratings revealed a main effect of size of denominator, $F(1, 446) = 30.11$, $p = .001$, $\eta_p^2 = .06$, of graph literacy, $F(1, 446) = 51.20$, $p = .001$, $\eta_p^2 = .10$, and of type of icon array, $F(5, 446) = 6.25$, $p = .001$, $\eta_p^2 = .07$. Additionally, an interaction between size of denominator and graph literacy was observed, $F(1, 446) = 5.41$, $p = .020$, $\eta_p^2 = .01$. The differences in confidence as a function of denominator size were larger for participants with low graph literacy than for those with high graph literacy; however they reached statistical significance for both groups of participants ($p < .001$ and $p = .045$, respectively).

Table 2. Mean subjective confidence in scenarios with equal and different denominators as a function of type of icon array, for participants with low graph literacy (top panel) and high graph literacy (bottom panel).

	Equal denominators			Different denominators		
Low graph literacy						
	Equal denominators (R^2 = .11)			Different denominators (R^2 = .11)		
	M (SEM)	*β*	*P*	M (SEM)	*β*	*P*
Static	**5.95 (.34)**	—	—	**5.47 (.33)**	—	—
Sequential	**6.59 (.26)**	.10	.16	**5.92 (.27)**	.07	.35
Labeling	**6.56 (.38)**	.12	.08	**5.79 (.38)**	.08	.27
Labeling reproduce	**6.13 (.38)**	.04	.53	**5.80 (.40)**	.06	.36
Transfer reproduce	**6.89 (.33)**	.17	.02	**6.20 (.34)**	.13	.08
Transfer estimate	**7.78 (.26)**	.32	.001	**7.45 (.30)**	.33	.001
High graph literacy						
	Equal denominators (R^2 = .14)			Different denominators (R^2 = .18)		
	M (SEM)	*β*	*P*	M (SEM)	*β*	*P*
Static	**7.11 (.43)**	—	—	**6.85 (.44)**	—	—
Sequential	**7.78 (.36)**	.16	.09	**7.38 (.42)**	.13	.16
Labeling	**8.46 (.29)**	.24	.009	**7.92 (.41)**	.17	.06
Labeling reproduce	**7.60 (.32)**	.12	.18	**7.48 (.30)**	.15	.10
Transfer reproduce	**7.95 (.38)**	.18	.056	**7.79 (.43)**	.18	.04
Transfer estimate	**8.60 (.30)**	.30	.001	**8.75 (.29)**	.35	.001

Note: Results are reported for regression models controlling for numeracy.

How does type of icon array affect subjective confidence, for individuals with low vs. high graph literacy? Table 2 shows mean confidence ratings among participants with low and high graph literacy for the different types of icon arrays, and results of linear regressions predicting confidence ratings. As can be seen in Table 2, subjective confidence was in most cases significantly higher in the transfer reproduce and transfer estimate conditions. For participants with high graph literacy, this was also the case for the labeling condition. These results indicate that some types of dynamic icon arrays that were not associated with a significant increase in accuracy were nevertheless associated with an increase in confidence.

How are graph evaluations affected by graph literacy and type of icon arrays? As planned, the three graph evaluation ratings were averaged to obtain mean ratings for each type of icon array. As participants evaluated graphs after they had viewed all medical scenarios, mean evaluations reflected both scenarios with equal and with different denominators. A linear regression model predicting mean evaluations was built, including graph literacy and type of icon arrays as predictors. The full model predicted evaluations, $R^2 = .08$, F (6, 451) = 6.14, $p = .001$. Interestingly, graph literacy was not a significant predictor ($t = 1.53$, p = .13, $\beta = .07$); however, including the interaction term of graph literacy and type of icon array in a separate step improved predictions of accuracy (R^2 change = .05, $p < .01$).

Table 3. Mean graph evaluation ratings as a function of type of icon array, for participants with low graph literacy and high graph literacy.

	Low graph literacy ($R^2 = .06$)			High graph literacy ($R^2 = .24$)		
	M (SEM)	β	*P*	M (SEM)	β	*P*
Static	**5.48 (.30)**	—	—	**4.94 (.43)**	—	—
Sequential	**6.29 (.25)**	.16	.04	**5.36 (.42)**	.07	.41
Labeling	**5.93 (.32)**	.10	.17	**8.05 (.26)**	.50	.001
Labeling reproduce	**6.51 (.30)**	.19	.009	**6.76 (.33)**	.31	.001
Transfer reproduce	**6.29 (.26)**	.17	.026	**7.35 (.36)**	.42	.001
Transfer estimate	**6.65 (.27)**	.24	.001	**6.55 (.32)**	.26	.003

Note: Results are reported for regression models controlling for numeracy.

How does type of icon array affect graph evaluations, for individuals with low vs. high graph literacy? Table 3 shows mean graph evaluation ratings among participants with low and high graph literacy for the different types of icon arrays, and results of linear regressions predicting evaluations. Among participants with low graph literacy, ratings were significantly higher in all conditions except the labeling condition, as compared to the static condition (see Table 3). Among participants with high graph literacy, ratings were significantly higher in all conditions except the sequential condition. These results indicate that increases in graph evaluations did not reflect increases in accuracy.

Discussion

In this paper, we sought to examine the effectiveness of the manipulation of different types of dynamic features in icon arrays designed to improve understanding of treatment risk reduction. Building on previous research (Garcia-Retamero & Cokely, 2011; Garcia-Retamero & Galesic, 2010; Okan et al., 2012a), we also sought to document the influence of individual differences in graph literacy on the efficacy of the different types of dynamic displays. Our results revealed that the only type of dynamic display that contributed to significantly increase accuracy of risk understanding over and above a set of static icon arrays was one including a reflective question (i.e., an estimate concerning the number of people harmed after taking a drug), followed by feedback provided visually. This display was associated with improved performance both among individuals with low and with high graph literacy, across medical scenarios including equally and unequally sized groups of treated and non-treated patients. Additionally, graph literacy predicted accuracy of risk understanding with the different types of icon arrays, as well as people's self-reported confidence in their risk estimates. However, increases in confidence and in perceived usefulness of the displays were observed for some types of dynamic icon arrays that did not improve performance.

The present findings shed light on the processes that give rise to some of the cognitive performance benefits associated with icon arrays. In particular, they suggest that it is unlikely that the limited ability of less graph literate individuals to benefit from static icon arrays is attributable solely to insufficient attention and encoding of the visual pattern in icon arrays (i.e., first process involved in graph comprehension; Carpenter & Shah, 1998) or to a failure to identify the referents in icon arrays (i.e., third process involved in graph comprehension). Dynamic icon arrays designed to support the first process (by directing viewers' attention to the different regions of the arrays in a sequential manner) or the third process (by including explanatory labels appearing sequentially next to each region indicating what it represented) did not result in a significant increase in performance, as compared to static arrays. These findings do not rule out the possibility that individuals with low graph literacy engage to a lesser extent in these processes when viewing static icon arrays, as compared to more graph literate ones (see Okan et al., 2013a, for evidence of differences in the processes underlying the comprehension of bar graphs and line graphs). However, they suggest that such processes may be contributing to, but cannot fully account for, the impaired performance exhibited by less graph literate individuals when presented with icon arrays.

The increase in accuracy of risk understanding observed in the condition involving a reflective question (transfer estimate condition) suggests that more active, elaborative processing of the risk information is an essential component of the beneficial effects of icon arrays (c.f., Natter & Berry, 2005). These findings accord with results showing that elaborative processing is strongly related to superior risky decision making (Cokely & Kelley, 2009). In our study, active processing was encouraged by the inclusion of a reflective question that required the generation of a solution not yet displayed. In contrast, displays which required people to restate the information depicted or which included a second set of arrays (but not a reflective question), did not contribute to significantly improve performance. This suggests that it is unlikely that the beneficial effect of the transfer estimate condition was merely due to the fact that people were instructed to answer a question concerning the information depicted (and not necessarily a reflective question), or to the fact that they viewed two sets of icon arrays instead of one. Additionally, the current results show that displays that encourage active processing of the information can improve performance not only among less graph literate individuals, but also among those who are more graph literate.[9]

Our findings also have implications for the graphical communication of medical information, as they suggest ways to enhance risk comprehension even among less graph literate individuals. The needs of such individuals should be taken into account not only by doctors communicating information to patients, but also by scientific experts who are responsible for developing communication materials for the general public (Bruine de Bruin & Bostrom, 2013). At the same time, the current results converge with recent studies showing that interactive and animated design features in graphical displays do not necessarily contribute to improve risk comprehension (Zikmund-Fisher et al., 2012, 2011;

[9] The need for viewers to generate information to comprehend graphs has also been discussed by Trickett and Trafton (2006), who showed that this is often achieved by means of mental spatial transformations on the information explicitly represented in the graphs. Although it seems unlikely that the types of displays and tasks employed in our study involved the need to perform spatial transformations, our results converge to highlight that, for some types of displays, an active generation of information may be required to fully comprehend the information depicted.

but see also Ancker, Weber, & Kukafka, 2011b). The novelty of some displays or their potential to divide people's attention reinforces the notion that sometimes less can be more when presenting information (Peters, Hibbard, Slovic, & Dieckmann, 2007; Zikmund-Fisher et al., 2012, 2011). For instance, Zikmund-Fischer et al. (2011) recently showed that providing people with numerical information and requiring them to depict this information interactively in icon arrays did not have a beneficial effect. Similarly, our results suggest that asking people to restate or reproduce information already provided in icon arrays or in accompanying labels is not likely to help them understand the information about risk.

Our results also revealed that many of the dynamic icon arrays were evaluated more positively than the static ones both by participants with low and high graph literacy, even though only one of them (i.e., transfer estimate) resulted in a reliable increase in performance. Additionally, participants with high graph literacy did not evaluate icon arrays more positively overall than those with low graph literacy. These findings are in line with research showing that the design features that viewers like or find most useful are not necessarily those that contribute to enhance performance (e.g., Feldman-Stewart et al., 2000; McCaffery et al., 2012; Micallef et al., 2012; but see also Zikmund-Fisher et al., 2012). In addition, we found that some dynamic displays that were not associated with significant increases in accuracy of risk understanding were nevertheless associated with increases in self-reported confidence. Thus, our results suggest that it is not advisable to rely solely on patients' subjective confidence or perceptions of the extent to which different types of icon arrays are useful (see Ancker et al., 2006, for a discussion). Finally, our finding that both accuracy and confidence were higher when medical scenarios included groups of treated and non-treated patients of the same size suggests that communicating risk information with unequally sized groups should be avoided where possible, even if information is provided both numerically and using visual aids (see also Ancker et al., 2006; Garcia-Retamero & Dhami, 2011; Garcia-Retamero & Galesic, 2009; Paling, 2003).

As with most studies, our study has some limitations and our results point to a number of key questions for future research. First, future research should seek to isolate the effect of each of the different design features manipulated in the present study. Here, the different types of icon arrays were constructed by introducing new design features (e.g., the addition of explanatory labels) built on top of previously introduced ones (e.g., the sequential presentation of the different regions of the arrays). This design was chosen in order to maximize the potential effect of our manipulations; however, it does not enable us to estimate the specific effect of the different design features employed in isolation. Accordingly, it is hard to determine the degree to which encouraging the allocation of attention to different regions of the arrays is a necessary prerequisite for increasing accuracy of risk understanding. The pattern of results observed for the labeling condition (i.e., an increase in performance, albeit not statistically significant) may be interpreted as evidence that the beneficial effect of the transfer estimate condition was not due exclusively to the inclusion of a reflective question, but to some extent also to the presence of explanatory labels supporting the identification of the referents in icon arrays. Future research should explore this issue, in order to determine whether the effects of the different design features can indeed be additive.

Second, future research could examine the effectiveness of simpler icon arrays that may also encourage a more active, elaborative processing of risk information. A promising means might be to remove numerical information from the text (see Micallef et al., 2012), as forcing people to explore the icon arrays might prompt them to process the depicted information in a more active fashion. Studies examining text comprehension have shown that placing some impediments in a text (e.g., deleting propositions that mark the role of each paragraph) can in some cases enhance learning by encouraging a deeper, more active processing of the text (McNamara, E. Kintsch, Songer, & W. Kintsch, 1996; see Kintsch 2004, for a review). Finally, future research could aim to enhance the ecological validity of the present study. Because our experiments were not conducted with patients in a clinical setting our participants might have been less motivated to think about the risk information provided than would patients trying to understand the risk reduction of a treatment. To the extent that this is correct, our results suggest that the estimated benefits might be relatively lower bound estimates of the potential benefits of well-designed dynamic icon arrays.

In sum, the present work reveals that the power of icon arrays to enhance the comprehension of quantitative information can be supplemented by including carefully designed features encouraging an active processing of the information. In our study this was achieved though self-administered reflective questions, followed by visual feedback. This kind of dynamic display could be included in regulated websites providing medical and health statistics tutorials, serving not only to improve the communication of risk among patients, but also for educational and training purposes. Such internet-based tools could contribute to empower consumers, equipping them with the knowledge and skills to make informed decisions. Future research could seek to examine whether such active processing can also be encouraged through simpler and quicker means (e.g., reflective questions posed by the doctor in the clinic). Our results also suggest that future research is needed on additional aids that may contribute to further enhance performance among less graph literate individuals.

Acknowledgments

This study has been funded by a grant awarded by the Informed Medical Decisions Foundation (United States) to Yasmina Okan (ref. # 0230-1), and by grants awarded by the Ministerio de Economía y Competitividad (Spain) to Rocio Garcia-Retamero (ref. # PSI2011-22954) and to Antonio Maldonado (ref. # PSI2009-12217). We are grateful to Mirta Galesic and Andrés Catena for their valuable comments and suggestions.

Appendix A

Number of treated and non-treated patients who die after a heart attack for all denominator sizes. Note: Values in brackets were only presented in the transfer reproduce and transfer estimate conditions, for the second set of icons included in these conditions. Risk reduction is 80% in all conditions.

	Treated patients		Non-treated patients	
Denominator Sizes	Patients who died	Population size	Patients who died	Population size
800–800	16	800	80	800
(400–400)	(8)	(400)	(40)	(400)
800–100	16	800	10	100
(400–50)	(8)	(400)	(5)	(50)
100–800	2	100	80	800
(50–400)	(1)	(50)	(40)	(400)
100–100	2	100	10	100
(50–50)	(1)	(50)	(5)	(50)

Appendix B

Task/ Instructions	Text
Background label	These white points represent a (another) group of XXX people. This is the total number of people who *did not (did) take* Benofreno in this study.
Foreground label	These black points represent XXX people. These people are part of the group of XXX people who *did not (did) take* Benofreno. These XXX people have died.
Restate background	Now please indicate the total number of people who *did not take(took)* Benofreno.
Restate foreground	Now please indicate the total number of people who died, in the group of people who *did not take (took)* Benofreno.
Introduction to second set of icon arrays (only in Transfer Reproduce and Transfer Estimate)	You will now see the results of a study conducted with Benofreno. This study was conducted with a group of people who had the same characteristics as the group of people who took part in the previous study, and was conducted in the same situation or context (same conditions). It is therefore expected that results will be along the lines of those shown for the first study.
Estimate foreground (only in Transfer Estimate)	Now please estimate the number of people who would die in the group of people who *did not take (took)* Benofreno in this second study. Note: Recall that this second study was conducted with a group of people that had the same characteristics as the group of people who took part in the previous study, and was conducted in the same situation or context (same conditions).
Feedback	Correct/Incorrect. You can see the correct response highlighted in red above.

Note: Text with italics indicates alternation in wording for the top vs. bottom icon arrays. XX indicates the total numbers of treated and non-treated patients and of those who died, and varied for the different medical scenarios (see Appendix A). For Transfer Reproduce and Transfer Estimate, the Foreground label for the second set of icon arrays stated only: “In this group XX people died,” to avoid repetition.

CHAPTER IV.

WHEN HIGHER BARS ARE NOT LARGER QUANTITIES: ON INDIVIDUAL DIFFERENCES IN THE USE OF SPATIAL INFORMATION IN GRAPH COMPREHENSION

When Higher Bars Are Not Larger Quantities: On Individual Differences in the Use of Spatial Information in Graph Comprehension

Abstract

Graphical displays use spatial relations to convey meaning, facilitating the communication of quantitative information. However, information conveyed by spatial features can conflict with that conveyed by features linked to arbitrary conventions (e.g., axes labels or scales), leading to misinterpretations. Here, we investigated the role of individual differences in graph literacy on the interpretation of health-related bar graphs containing such conflicts. Individuals with low graph literacy were more often biased by spatial-to-conceptual mappings grounded in their real world experience, neglecting information in titles of graphs, axes labels and scales. Implications for perspectives on embodied cognition and effective graphical design are discussed.

Published as:

Okan, Y., Garcia-Retamero, R., Galesic, M., & Cokely, E. T. (2012). When higher bars are not larger quantities: On individual differences in the use of spatial-to-conceptual mappings in graph comprehension. *Spatial Cognition and Computation, 12,* 195–218.

Introduction

Graphical displays represent quantitative information in spatial locations, often enabling better and faster comprehension as compared to numerical or text-based formats (Munzner et al., 2006; Tversky, 2001). The translation of spatial information into conceptual information in graphs—*the spatial-to-conceptual mapping*—is frequently rooted in our experience with the physical environment (Tversky, 2001, 2009). For example, if the content of a container reaches a higher level than the content of another, this typically means that the first one contains more substance. In graphs, the knowledge acquired in the physical world can serve as a basis to map spatial information onto information about more abstract quantities (Gattis, 2002) such as profits, people, and utilities (e.g., higher bars can reflect larger profits). Thus, graphs constitute external spatial representations that we can use to reason about non-spatial concepts, on the basis of knowledge acquired in interactions with our environment (Tversky, 2009; Wilson, 2002).

However, in some cases information conveyed by *spatial features* in graphs may conflict with information conveyed by features linked to arbitrary conventions such as axes labels and the range of scale values (e.g., the numerical values on the scale can be reversed so that higher bars mean less profit). In such cases, an overreliance on spatial-to-conceptual mappings can lead to misinterpretations of the data depicted. A correct interpretation would require considering information from *conventional features* such as the axes labels or the scale values, and overriding direct spatial-to-conceptual mappings (e.g., recognizing that a higher bar does not necessarily imply more profit).

In this paper, we address the issue of individual differences in *graph literacy* (i.e., the ability to understand graphically presented information; Galesic & Garcia-Retamero, 2011b), and its relation to one's reliance on spatial-to-conceptual mappings in graph comprehension. Specifically, we examine the extent to which graph literacy affects people's use of mappings grounded in their real world experience to interpret graphs presenting quantitative medical information (i.e., prevalence of different diseases or effects linked to different treatments). Additionally, we examine the effect of the orientation of such graphs (i.e., vertical vs. horizontal) on comprehension. Graphical displays containing health-related information have been shown to help people to overcome difficulties in the comprehension of risks and benefits of different medical treatments, screenings, and health behaviors (Ancker, Senathirajah, Kukafka, & Starren, 2006; Garcia-Retamero & Cokely, 2011; Garcia-Retamero & Galesic, 2010a; Lipkus, 2007). The investigation of individual differences in graph comprehension plays a key role in the development and customization of health-related decision-support systems and risk communication (Okan, Garcia-Retamero, Cokely, & Maldonado, 2012; see also Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012).

Processes involved in graph comprehension

Graph comprehension models have identified three types of processes that viewers must follow to extract information from graphical displays such as line or bar graphs (Carpenter & Shah, 1998; Lohse, 1993; Pinker, 1990). These processes are iterative and incremental (i.e., viewers must repeat the cycle of processes to comprehend the information represented; Carpenter & Shah, 1998).

The first process is encoding the visual pattern and identifying the principal features in the graphs. This involves making different visual judgments of the elements (e.g., judgments of position along a scale, slope, length, or angle; Cleveland & McGill, 1986; Simkin & Hastie, 1987). For instance, for the graph shown in Figure 1a the viewer should encode the different bars and make visual judgments concerning their height.

The second process is the translation of the identified visual features into conceptual relations (Carpenter & Shah, 1998; Kosslyn, 1989; Pinker, 1990). Visual features of graphical displays can convey meaning in a number of ways. For example, variations in the saliency of features in terms of size, color, or highlighting can be used to indicate variations in the quantity of the variables represented. Similarly, the spatial arrangement of different elements can be used to indicate relationships between the variables depicted (e.g., proximity in space of the elements frequently indicates proximity on properties such as time or value; Kosslyn, 2006; Tufte, 2001; Tversky, 2001). As noted, a crucial characteristic of the mappings that graph viewers establish between spatial features and conceptual relations is that these mappings are frequently grounded in their experience with the physical world (Tversky, 2001, 2009). For the graph shown in Figure 1a, the second process involved in graph comprehension would entail the mapping of spatial features (bars of different heights) onto the concept of quantity.

The third process outlined in graph comprehension models involves determining the referents of the concepts identified by associating them with the specific variables and their numerical values (Carpenter & Shah, 1998; Shah & Carpenter, 1995; see also Huestegge & Philipp, 2011 Ratwani, Trafton, & Boehm-Davis, 2008). This process entails identifying and inferring information from conventional features in graphs, including the title of the graph, axes labels, legends or numerical values on the scales. For instance, in line plots or bar graphs it is necessary to identify the variables represented on the *x* and the *y* axes and which are the values that these variables take. For the graph shown in Figure 1a, this third process would involve inferring information from the title, axes labels, and numerical scale. Crucially, this process would entail attending to textual information in the title and label for the dependent variable (i.e., “percentage of people without different allergies”), and deducing that in this graph higher bars represent lower quantities.

Individual differences in graph comprehension

Although a large body of research has investigated graph comprehension processes in the general population (Cleveland & McGill, 1986; Lohse, 1993; Simkin & Hastie, 1987), relatively less research has examined the factors that can moderate the extent to which different viewers engage in such processes. To illustrate, some authors have analyzed the impact of viewers’ graph-related knowledge (e.g., Shah & Carpenter, 1995; Shah & Freedman, 2011) and models of graph comprehension have been developed incorporating different aspects of viewers’ prior knowledge (e.g., Freedman & Shah, 2002). However, it is yet unclear how and when different kinds of prior knowledge affect the graph comprehension processes outlined above. The first aim of this investigation was to examine how *graph literacy* affects the extent to which viewers engage in these processes.

Graph literacy is a skill typically acquired through formal education, which can affect the comprehension of graphical representations of numerical information in important ways. For instance, individual differences in graph literacy can moderate the effectiveness of visual aids (Gaissmaier et al., 2012; Garcia-Retamero & Galesic, 2010b), affecting people’s decision-making performance, as well as subjective perceptions of this performance (Okan et al., 2012a). Graph literacy has also been shown to affect the likelihood that viewers generate different inferences from data in graphical displays. When viewing bar graphs, individuals with high graph literacy are more capable of providing descriptions of main effects than are less graph-literate individuals (Shah & Freedman, 2011). When viewing more complex displays such as weather maps, expert viewers spend more time exploring task-relevant information than novice viewers, and show superior performance in making inferences from such displays (Canham & Hegarty, 2010).

Here we examined the influence of graph literacy on people's tendency to interpret graphs on the basis of mappings between spatial features and conceptual relations (i.e., the second process involved in graph comprehension). When spatial features do not readily evoke a conceptual relation, viewers lacking specific graph-related knowledge can have difficulties interpreting graphs accurately (Shah, Freedman, & Veriki, 2005; Shah & Hoeffner, 2002). However, as noted above spatial-to-conceptual mappings in graphs are often rooted in one's experiences with the physical world. In such cases, viewers with low graph literacy can apply the knowledge acquired in their environment to translate spatial features to concepts in graphs. Instead, the identification of referents of concepts (i.e., third process involved in graph comprehension) is guided by specific graph-related knowledge and experience. This knowledge could affect viewers' ability to integrate information from axes labels or the numerical scale with the corresponding lines or bars in the chart. Additionally, graph-related knowledge could direct subsequent cycles of encoding and interpretation (Carpenter & Shah, 1998; Pinker, 1990), directing attention to labels or scales which contain information required to answer questions about the data.

Taking the roles of experience into account we predicted that individuals with low graph literacy would be more likely than those with high graph literacy to rely primarily on spatial-to-conceptual mappings grounded in their experience with the environment to interpret graphs, often failing to incorporate information from the title, axes labels, or numerical scales. We further predicted that differences between individuals with low and high graph literacy would be clear in graphs containing a *conflict* between information conveyed by *spatial features* (e.g., bar heights) and information conveyed by *conventional features.* Therefore, we hypothesized that low graph literacy would be related to an overreliance on spatial-to-conceptual mappings when interpreting graphs containing such conflicts. As a consequence, participants with low graph literacy would more frequently misinterpret the data depicted than highly graph literate participants (H_1). We further hypothesized that the overreliance on spatial-to-conceptual mappings would more often lead individuals with low graph literacy to make non-normative decisions as compared to highly graph literate individuals (H_2).

Properties of graphical displays: The effect of orientation on comprehension

Other important factors that can affect graph comprehension processes are the properties of the graphical displays. For instance, variations in the perspective of bar graphs (i.e., two-dimensional vs. three-dimensional) can impact accuracy in the judgment of bar heights (Fischer, 2000; Zacks, Levy, Tversky, & Schiano, 1998). Bar graphs can also vary in their *orientation* (i.e., vertical vs. horizontal) and this can affect viewers' speed in judging the quantities represented (Fischer, Dewulf, & Hill, 2005). However, it is currently unknown which orientation is best suited to enhance comprehension when a conflict exists between information conveyed by spatial features and by conventional features.

The second aim of this paper was to assess how the comprehension of data in bar graphs is affected by their orientation: vertical or horizontal. There are at least two ways in which a change in the orientation of a bar graph can affect comprehension processes. First, changes in comprehension processes can be triggered by a change in the orientation of *bars*. When bars are oriented vertically viewers may be more likely to rely primarily on spatial-to-conceptual mappings than when bars are oriented horizontally. The rationale is that the association between the spatial position of a substance and its quantity in the physical world is more robust along the vertical dimension than along the horizontal dimension (Tversky, 2001, 2009). Indeed, people associate spatial position along the horizontal dimension with numerical magnitudes (Dehaene, Bossini, & Giraux, 1993) or temporal sequences (Gevers, Reynvoet, & Fias, 2003; Tversky, Kugelmass, & Winter, 1991). However, the directionality of this representation seems to be rooted in reading habits for words and numbers, rather than to rest on a natural correspondence (Shaki, Fischer, & Petrusic, 2009; Tversky et al., 1991).

In contrast, a universal correspondence exists along the vertical dimension between upward positions with larger quantities, and lower positions with smaller quantities (Lakoff & Johnson, 1980; Tversky et al., 1991). This correspondence is grounded in the physical environment, where increasing the quantity of any substance typically increases its vertical extent. Thus, when trying to understand graphs viewers may be more likely to establish the mapping "higher bars = more" as compared to the mapping "horizontally longer bars= more. As a consequence, when bars are vertical viewers may rely to a larger extent on spatial-to-conceptual mappings grounded in their real world experience to interpret graphs.

A second way in which a change in the orientation of a bar graph can affect comprehension processes is linked to the change in the orientation of conventional features. When a graph is rotated the *numerical scale* necessarily varies its orientation (instead, the orientation of *axes labels* can be kept constant). When the scale is oriented horizontally viewers may be more likely to incorporate the information it contains in their interpretation of the graph, as compared to when it is oriented vertically. The rationale is that such information may be easier to read and integrate if it is displayed horizontally (i.e., matching Westeners' reading habits) than vertically.

Graph orientation and types of conflict between spatial and conventional features

As mentioned above, in the present research we examined how people interpret graphs containing a conflict between information conveyed by *spatial features* (i.e., bar heights) and information conveyed by *conventional features*. Such a conflict can occur when spatial features of the graph convey different meaning than (1) textual information in the title and axes labels (*textual-spatial conflict*) or (2) numerical values on the scale (*scale-spatial conflict*). The effect of a change in the orientation of a bar graph on comprehension can be linked to the type of conflict existing in the graph. Taking this into account, we generated two alternative hypotheses.

If comprehension is affected by the change in the orientation of *bars*, the stronger reliance on spatial-to-conceptual mappings in vertical than in horizontal bar graphs should occur for both (1) graphs containing a textual-spatial conflict and for (2) graphs containing a scale-spatial conflict. That is, no matter what type of conflict a graph contains viewers should be more likely to rely primarily on spatial-to-conceptual mappings and thus to misinterpret information more often when graphs are oriented vertically than when they are oriented horizontally (H_{3a}). Additionally, for both types of conflict the overreliance on spatial-to-conceptual mappings should lead viewers to make non-normative decisions more often when graphs are oriented vertically than when they are oriented horizontally (H_{4a}).

Alternatively, comprehension may not be affected by the change in the orientation of bars, and may be affected instead by the change in the orientation of conventional features (i.e., the *scale*). In such case, performance on graphs containing a textual-spatial conflict should not be affected significantly by orientation. The rationale is that in such graphs essential information can be extracted from the title (which does not change its orientation when the graph is rotated) and from the axes labels (which can be displayed in the same orientation even if the graph is rotated). Therefore, a larger tendency to rely primarily on spatial-to-conceptual mappings for vertical than for horizontal bar graphs should be observed only when a scale-spatial conflict exists (H_{3b}). Additionally, for graphs containing such conflict viewers should make non-normative decisions more often when graphs are oriented vertically than when they are oriented horizontally (H_{4b}).

To test our hypotheses, we conducted an experiment in which participants with different levels of graph literacy were presented with bar graphs displaying quantitative medical information. .These graphs were constructed in such a way that following spatial-to-conceptual mappings—according to which higher or longer bars imply more quantity and lower or shorter bars imply less quantity—would lead to erroneous interpretations of the data and to non-normative decisions. Half of the graphs contained essential information in the title and axes labels (i.e., a textual-spatial conflict), while the other half contained essential information in numerical values on the scale (i.e., a scale-spatial conflict). Some participants were provided with vertically oriented bar graphs, while others received horizontally oriented graphs. For each graph, participants answered a question designed to evaluate their interpretation of the information represented and made a decision on the basis of this information.

Method

Participants

Participants were recruited via Amazon's Mechanical Turk. Mechanical Turk provides access to a paid internet participant panel that can be used for conducting behavioral research. The magnitude of effects obtained using this platform have been found to be equivalent to those obtained using traditional subject pools in laboratory-based experiments (Paolacci, Chandler, & Ipeirotis, 2010; Sprouse, 2011; see also Cokely et al., 2012; Cokely, Ghazal, Galesic, Garcia-Retamero, & Schulz, 2013; Feltz & Cokely, 2011).

The online study was hosted on the web survey platform Unipark (www.unipark.de) and participants were redirected to this website after clicking on a link provided in the Human Intelligence Task (HIT) forum on Mechanical Turk (www.mturk.com). Upon completion of the study participants were required to enter a self-generated user code both in Unipark and in the Mechanical Turk HIT, in order to verify participation. A total of 251 residents of the United States completed the study. Of those, 68 participants were randomly assigned to a control condition where the question used to evaluate the interpretation of the graphs was modified, as will be described below. Results did not vary as a function of this modification in the question and so for simplicity these data are not reported. The final sample included 182 participants (54% women, median age of 34 years, range 18–68).

The mean completion time was 15.6 minutes (*SD* = 5.6). Most participants (98%) completed the study in 30 minutes or less. Duration was correlated with age [$r(180) = .297$, $p = .001$]. We excluded a young participant who took 44 minutes to complete the study, as we suspected that he or she was not focused on the tasks. Participants were randomly assigned to the different experimental conditions through a random trigger with a uniform distribution generated in Unipark (n = 63 on average). All participants consented to participation through an online consent form at the beginning of the study.

Design and Materials

Participants were presented with four bar graphs depicting quantitative medical information (i.e., prevalence of different diseases or effects linked to different treatments). Each graph contained five data points, a title, and the corresponding labels for both axes. We manipulated the *type of conflict* in the graphs within-subjects and the *orientation* of graphs between-subjects.

To manipulate the *type of conflict* we constructed two different sets of bar graphs. In *Graphs A and B*, essential information was included in the title and in the textual label for the dependent variable. Therefore these graphs contained a textual-spatial conflict. Specifically, Graph A presented data about percentages of people *without* different types of allergy (see Figure 1a). Participants were asked about the type of allergy affecting the largest percentage of people. To answer this question correctly, participants had to attend to the title and the label for the dependent variable in order to infer that the usual spatial-to-conceptual mapping was reversed (i.e., they had to infer that higher bars represented lower values). Graph B presented data about the *change* in the percentage of people with different types of cancer during the previous year (see Figure 1b). Participants were asked about the type of cancer that affected the smallest percentage of people during the previous year. To answer this question correctly, participants had to attend to the title and the label for the dependent variable in order to infer that the height of bars did not correspond to quantities in terms of *absolute* percentages (i.e., they had to infer that the required information was not reported).

For *Graphs C and D* essential information was provided in the numerical scale for the dependent variable. Therefore these graphs contained a scale-spatial conflict. Graph C presented data about the percentage of people with different types of influenza and the numerical scale was reversed (i.e., values increased from top to bottom for vertically oriented graphs, and from right to left for horizontally oriented graphs; see Figure 1c). Participants were asked about the type of influenza affecting the largest percentage of people. To answer this question correctly participants had to attend to the scale in order to infer that the usual spatial-to-conceptual correspondence between height and quantity was reversed. Finally, Graph D presented data about the percentage change in patients' body weight associated with different treatments. This final graph contained both positive and negative values; however, the zero baseline was not indicated and positive (negative) values were not represented by bars above (below) the baseline (see Figure 1d; see also Kosslyn, 2006). Participants were asked about the treatment that resulted in the smallest change in the patients' body weight. To answer this question correctly, participants had to attend to the scale in order to infer that the height of bars did not correspond to the magnitude in percentage change.

To analyze the effect of the *orientation*, we constructed three different versions of each graph: (1) a vertical graph where the bars and the label for the y axis were oriented vertically (see Figure 1d); (2) a vertical graph where the bars were oriented vertically and the label on the y axis was oriented horizontally (see Figure 1e); and (3) a horizontal graph where the pattern in the vertical graphs was rotated 90 degrees clockwise (see Figure 1f). We will refer to these graphs as *vertical standard*, *vertical with horizontal text*, and *horizontal* graphs, respectively. The second vertical condition was included to control for the potential effect of the change in the orientation of the label for the dependent variable when the graph is rotated. This control is relevant for graphs containing a textual-spatial conflict, where essential information is provided both in this label and in the title. In the vertical standard condition this label is displayed vertically, while the title is displayed horizontally. In contrast, in the horizontal condition both elements are displayed horizontally, and this may increase the likelihood that information contained in them is integrated. In the vertical with horizontal text condition, both elements are also displayed horizontally. Thus, for graphs containing a textual-spatial conflict, any differences in performance between the horizontal and the vertical with horizontal text conditions can be attributed to the change in the orientation of bars.

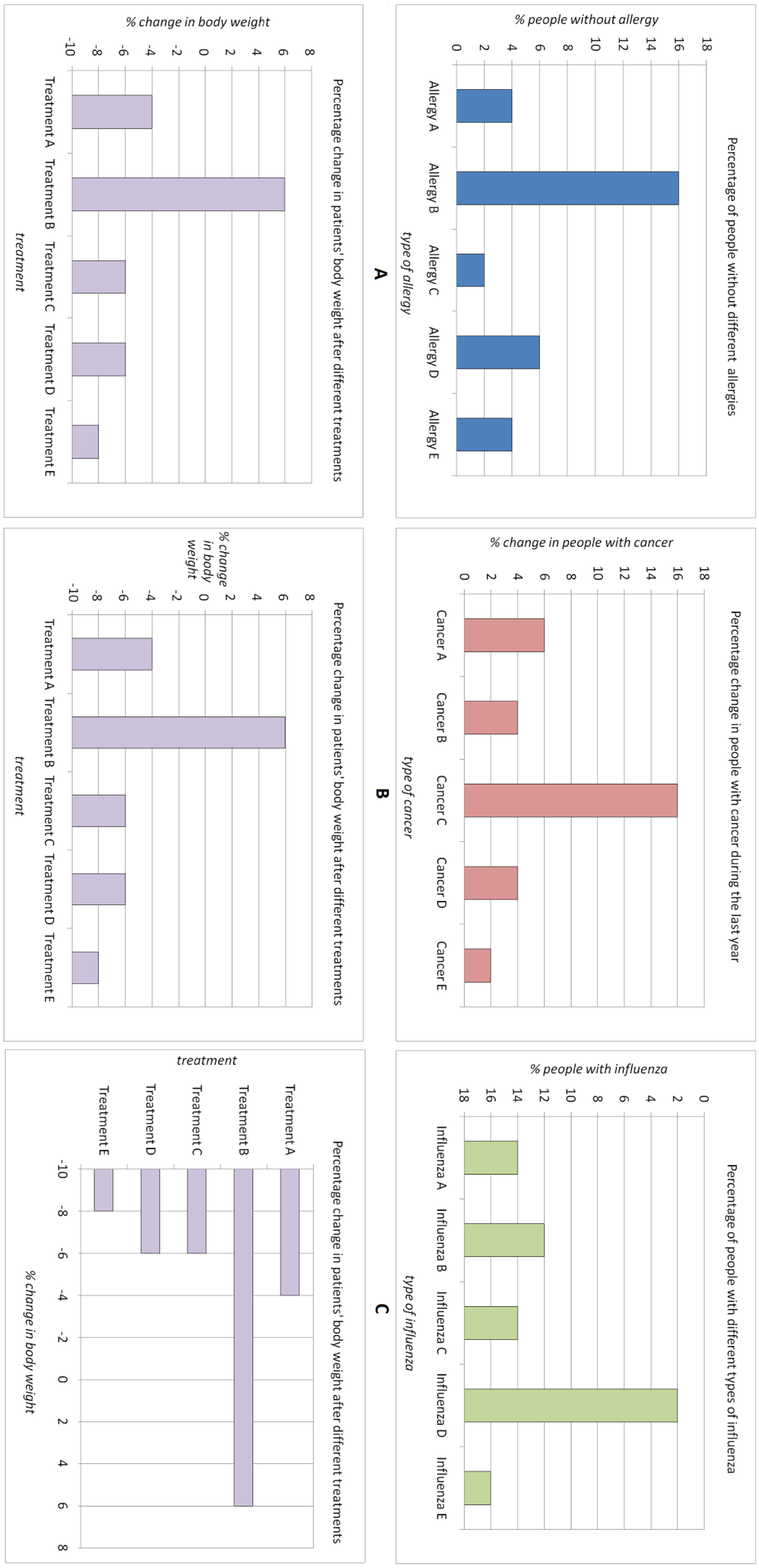


Figure 1. (A) Graph A with a *textual-spatial conflict*. Reverse information is reported in the graph; (B) Graph B with a *textual-spatial conflict*. The required information is not reported in the graph; (C) Graph C with a *scale-spatial conflict*. The numerical scale is reversed; (D) Graph D with a *scale-spatial conflict*. The numerical scale includes positive and negative values; (E) Graph D in the *vertical with horizontal text* orientation; (F) Graph D in the *horizontal* orientation. Data shown in all cases is fictional

In summary, we constructed two sets of bar graphs that differed in the type of conflict they contained. We further constructed three orientations for each bar graph. Each participant was presented with two graphs containing a textual-spatial conflict (Graphs A and B) and two graphs containing a scale-spatial conflict (Graphs C and D) in one of the three possible orientations.[10] Vertically (horizontally) oriented graphs were constructed to ensure that an exclusive reliance on the mappings high = more and low = less (horizontally longer = more and horizontally shorter = less) would lead to an incorrect interpretation of the data presented. Thus, for all graphs a correct interpretation required the integration of essential information presented in conventional features.

As dependent variables we measured participants' (1) *interpretation* of the information presented and (2) *decisions* made on the basis of the information presented in each graph. Both interpretations and decisions were measured using a multiple-choice item for each graph. For both dependent variables, the options provided for this item included (i) the correct response; (ii) an incorrect response corresponding to the mappings described above; and (iii) three other incorrect responses.

For the dependent variable *interpretation*, the item was designed to evaluate accuracy in understanding the data. For instance, for the graph providing information about the percentage of people *without* different types of allergies (Graph A, see Figure 1a) the question was "What type of allergy affected the largest percentage of people?" The correct response was "Allergy C," which was represented by the lowest bar (vertical orientation) or horizontally shorter bar (horizontal orientation). The incorrect response corresponding to the spatial mapping would be "Allergy B," which was represented by the highest bar (vertical orientation) or horizontally longer bar (horizontal orientation). The other three options (i.e., Allergy A, D, and E) were coded as other type of incorrect responses.

The multiple-choice item for the dependent variable *decision* was designed to assess participants' preference among different hypothetical treatments on the basis of the data (e.g., "There are different treatments to prevent each allergy. If you had to take one treatment, which one would you prefer?"). For graphs presenting information related with the

[10] As an additional control we included a condition in which participants were presented with the *vertical standard* graphs, but the question used to evaluate their interpretation of the graphs was phrased differently. Instead of being asked about types of disease affecting the *largest/smallest* percentage of people, participants in this condition were asked about the *most/least frequent* type of disease. This condition was included with the aim of determining whether a similar tendency to misinterpret graphs would be observed when the spatial-to-conceptual mapping entails a more abstract concept. Results in this condition did not differ from those in those in the vertical standard graph condition with the question phrased in terms of largest/smallest percentage. For simplicity, we do not report these results.

prevalence of different types of disease, the correct option was considered to be choosing the treatment that prevents the most frequent disease (Graphs A and C), or not having a preference when information about prevalence was not provided (Graph B). For the graph presenting effects linked to different treatments (Graph D) the correct option was considered to be choosing the treatment leading to the particular goal indicated in the question (i.e., the smallest possible change in body weight).

As our hypotheses were concerned with people's tendency to rely on mappings from spatial features of bars onto quantities of variables, for the analyses reported below we focused on the mean number of items where the incorrect response corresponding to the mapping was provided, both for interpretations and decisions. For each individual we computed the total number of incorrect responses corresponding to the mapping for graphs containing a spatial-scale conflict and for graphs containing a spatial-textual conflict. Thus, in each case the range of possible scores was 0 to 2, where 0 indicated that the participant had not given the response corresponding to the mapping in any graph of the group, and 2 indicated that he or she had given this response in all graphs of the group.

Procedure

Participants were presented with four bar graphs. The order of the graphs was randomized. For each graph, participants answered a question designed to evaluate their interpretation of the information represented and made a decision on the basis of this information. Subsequently they completed a graph literacy scale and a numeracy scale. Finally, participants completed a series of demographic questions and were debriefed.

Measurement of graph literacy: Graph literacy was measured using the instrument developed by Galesic and Garcia-Retamero (2011b; see also Garcia-Retamero & Galesic, 2010b). This scale consists of 13 items and measures three levels of graphical comprehension (Friel, Curcio, & Bright, 2001): (1) the ability to *read the data*, that is, to find specific information in the graph, which corresponds to the more elementary level (for instance, the ability to read off the height of a particular bar within a bar chart); (2) the ability to *read between the data*, that is, to find relationships in the data as shown on the graph, which corresponds to an intermediate level (for instance, the ability to read off the difference between two bars); and (3) the ability to *read beyond the data*, or make inferences and predictions from the data, which corresponds to an advanced level of graph comprehension (for example, the ability to project a future trend from a line chart).

Additionally, the scale is designed to cover four frequently used graph types—line plots, bar charts, pies, and icon arrays—and includes items dealing with the communication of medical risks, treatment efficiency, and prevalence of diseases. In sum, the scale measures both basic graph-reading skills and more advanced graph comprehension, for different types of graphs. The psychometric properties of this scale have been assessed in a survey conducted on probabilistically representative national samples of people from Germany and the United States, demonstrating satisfactory levels of internal consistency (Cronbach alpha of .74 in Germany and .79 in the United States) and convergent validity (the average correlation of the total score with graph comprehension items from existing literacy questionnaires was .44; for further details on the psychometric properties of the scale see Galesic & Garcia-Retamero, 2011b).

We split participants into two groups according to the median graph literacy score for the total sample (i.e., 11). Thus, the group of participants with low graph literacy included those who obtained 11 or fewer correct responses (n = 104), whereas the group of participants with high graph literacy included those who obtained 12 or more correct responses (n =78). Participants with low graph literacy answered on average 9.6 items correctly (*SD* =1.7), while participants with high graph literacy answered on average 12.5 items correctly (*SD* = .5).

Measurement of numeracy: In the experiment, we also assessed participants' numerical skills (i.e., the ability to use basic probability and numerical concepts; Lipkus, Samsa, & Rimer, 2001; see also Cokely et al., 2012). Participants' numeracy was measured using the three items in the general numeracy scale by Lipkus et al. (2001), based on the items developed by Schwartz, Woloshin, Black, and Welch (1997). Thus, the range of possible scores was from 0 to 3. An example of an item is "Imagine that we rolled a fair, six-sided die 1,000 times. Out of 1,000 rolls, how many times do you think the die would come up even (2, 4, or 6)?"

Results

First, we examined proportions of correct and incorrect responses for all graphs. The average proportion of correct responses for interpretations was 38.3% (*SE* = 6.7), while the average proportion of incorrect responses corresponding to the spatial-to-conceptual mapping (*mapping responses*) was 58.8% (*SE* = 6.2). As expected, proportions of incorrect responses not corresponding to the mapping were low (2.9% on average; *SE* = 1.7), indicating that the majority of participants who misinterpreted the graphs did so on the basis of direct spatial-to-conceptual mappings. Similarly, for decisions the average proportion of correct responses was 41.0% (*SE* = 7.6), while average proportions of incorrect responses corresponding and not corresponding to the mapping were 53.7% (*SE* = 4.9) and 5.3% (*SE* = 3.0), respectively. As proportions of incorrect responses not corresponding to the mapping were low, subsequent analyses focus on the total number of mapping responses computed for each participant, as planned.

We next conducted 3 × 2 × 2 analyses of variance (ANOVAs) with orientation (vertical standard vs. vertical with horizontal text vs. horizontal) and graph literacy (high vs. low) as between-subjects factors, and type of conflict (scale-spatial conflict vs. textual-spatial conflict) as within-subjects factor, on the total number of mapping responses for interpretations and decisions. We used the Bonferroni correction for post hoc analyses. The analyses revealed a significant main effect of graph literacy for interpretations, F (1, 176) = 31.26, p = .001, η_p^2 = .151, and decisions, F (1, 176) = 10.31, p = .002, η_p^2 = .055. These main effects were qualified by a reliable interaction between type of conflict and orientation, F (2, 176) = 6.12, p = .003, η_p^2 = .065 for interpretations, and, F (2, 176) = 6.30, p = .002, η_p^2 = .067 for decisions, and between graph literacy, type of conflict, and orientation for interpretations, F (2, 176) = 5.13, p = .007, η_p^2 = .055 (see Figure 2).[11]

Overall, the mean number of items in which a mapping response was provided was higher for participants with low graph literacy than for participants with high graph literacy, both for interpretations (M = 1.38, SE = .06 vs. M = .89, SE = .07, respectively) and decisions (M = 1.19, SE = .06 vs. M = .91, SE = .07, respectively).[12] These results are in line with our hypotheses suggesting that individuals with low graph literacy would be more likely than those with high graph literacy to rely primarily on spatial-to-conceptual mappings when interpreting graphs (H_1) and making decisions (H_2), often neglecting information in conventional features. However, the interaction obtained for interpretations between graph literacy, type of conflict, and orientation indicates that this tendency did not hold in all cases. No reliable differences were found between participants with high and low graph literacy for vertical graphs containing a scale-spatial conflict (p > .1 for pairwise comparisons).

[11] The inclusion of numeracy scores as a covariate did not influence the pattern of results. Additionally, analyses including graph literacy as a covariate instead of as a factor yielded converging results.

[12] In linear regressions, graph literacy scores were found to significantly predict the total number of mapping responses both for interpretations, β = -.24, t = 5.64, p = .001, R^2 = .15, and for decisions, β = -.16, t = 3.63, p = .001, R^2 = .07. Additionally, a hierarchical regression model predicting the total number of mapping responses was constructed, including numeracy and graph literacy scores. After controlling for the effect of numeracy, graph literacy continued to account for unique variance both for interpretations, F (1, 179) = 11.3, p = .001, R^2_{change} = .050, and for decisions, F (1, 179) = 6.3, p = .013, R^2_{change} = .032.

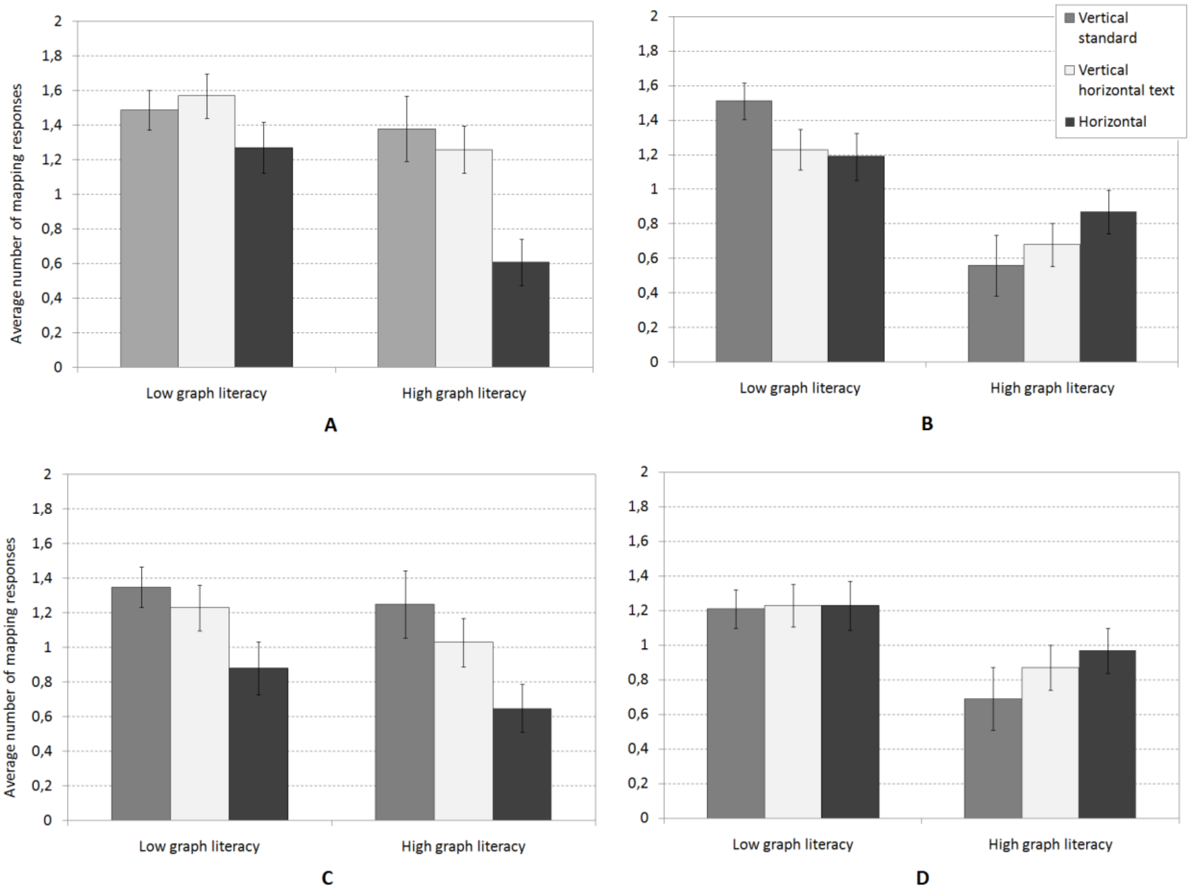


Figure 2. Mean number of mapping responses, as a function of graph literacy and orientation (A) for graphs containing a *scale-spatial conflict*, for *interpretations;* (B) for graphs containing a *textual-spatial conflict*, for *interpretations*; (C) for graphs containing a *scale-spatial conflict*, for *decisions*; (D) for graphs containing a *textual-spatial conflict*, for *decisions*. Error bars represent one standard error of the mean.

Concerning the effect of orientation, results indicated that the number of mapping responses varied as a function of orientation only for graphs containing a scale-spatial conflict. For graphs containing a textual-spatial conflict, the mean number of mapping responses for interpretations or decisions did not vary as a function of orientation (p > .1 for all pairwise comparisons). In contrast, for graphs containing a scale-spatial conflict the mean number of mapping responses was lower for the horizontal condition than for the vertical conditions, both for interpretations and for decisions (ps < .05). This result is inconsistent with the hypotheses that graphs containing vertical *bars* would lead to a larger reliance on spatial-to-conceptual mappings to interpret graphs (H_{3a}) and to make decisions (H_{4a}) than graphs containing horizontal bars, regardless of the type of conflict. Instead, results accord with the alternative hypotheses that viewers would be less likely to rely primarily on spatial-to-conceptual mappings to interpret graphs (H_{3b}) and to make decisions (H_{4b}) when the *scale* is displayed horizontally than when it is displayed vertically. Interestingly, the interaction observed for interpretations between graph literacy, type of conflict, and orientation suggests that graph literacy moderated the effect of the change in orientation of the scale. For graphs containing a scale-spatial conflict the difference in the number of mapping responses between the horizontal and the vertical orientations was significant for individuals with high graph literacy (ps < .01), but not for those with low graph literacy (ps > .1). A similar tendency was observed for decisions, but it was not significant.

Discussion

Graphical displays are powerful tools that can facilitate the communication and comprehension of quantitative information. A key to the success of graphs is that they exploit the human ability to think about abstract concepts in spatial terms. Through our direct experience with the physical world we acquire associations between spatial and conceptual aspects (e.g., the higher a pile of elements, the larger its quantity), which we frequently use as a basis to infer meaning from graphs (Tversky, 2001, 2009). However, on some occasions information conveyed by spatial features in graphs can conflict with information conveyed by conventional features such as the axes labels or the range of scale values. In such cases an overreliance on spatial-to-conceptual mappings grounded in the physical world can lead people to misinterpret the data depicted. In this study, we demonstrated that people's reliance on direct spatial-to-conceptual mappings varied as a function of graph literacy.

Our results revealed the existence of a strong tendency for people to make erroneous inferences about data presented in bar graphs containing a conflict between information conveyed by spatial features and information conveyed by conventional features. For a graph presenting the percentage of people *without* different types of allergy more than 40% of the participants incorrectly inferred that the most prevalent allergy was the one represented by the largest bar. For other graphs the tendency was even more dramatic with over 70% of the participants misinterpreting the data depicted. These findings support the notion that people frequently rely on spatial-to-conceptual mappings grounded in their real-world experience to interpret graphs (Tversky, 2001, 2009). Notably, our results suggest that people may frequently rely on these mappings even when this leads to erroneous inferences and non-normative decisions.

Crucially, our results also demonstrated that the tendency to rely primarily on spatial-to-conceptual mappings to interpret graphs and make decisions was stronger among less graph-literate individuals than among highly graph-literate individuals. Individuals with low graph literacy more often neglected important information in the title of the graphs, axes labels, and the numerical scales. These findings contribute to our understanding of some of the ways in which viewers' graph-related knowledge interacts with the characteristics of displays in graph comprehension. In particular, our data indicate that individuals with limited graph-related knowledge may often interpret graphs on the basis of direct translations of visuospatial features into conceptual information, grounded in their real world experience. As a consequence of their limited knowledge concerning graphic conventions, these individuals can be less likely to incorporate information from conventional features such as the axes labels or numerical values on the scales in their interpretation of graphs. Interestingly, our data also revealed that highly graph literate individuals may in some instances be as likely as less graph literate individuals to show errors linked to an overreliance on spatial-to-conceptual mappings. This was the case for vertical graphs containing essential information in the scale. One possible explanation of this finding is that participants did not engage in a thorough encoding of all elements in the graphs. This might have led even highly graph literate participants to fail to identify information in the scales. Whether this is in fact the case should be investigated in future research.

Our results also revealed that manipulating the orientation of bar graphs affected participants' tendency to show incorrect responses corresponding to direct spatial-to-conceptual mappings. However, this was only the case for graphs containing essential information in the scale and this effect of orientation was moderated by graph literacy. In particular, when such graphs were oriented horizontally highly graph literate participants were less likely to show incorrect responses corresponding to mappings than when they were oriented vertically. Instead, among less graph literate participants, the number of incorrect responses corresponding to mappings did not reliably vary as a function of orientation. This result is inconsistent with the hypothesis that viewers would be more likely to rely on spatial-to-conceptual mappings for graphs containing vertical *bars* than for graphs containing horizontal *bars.* This prediction was based on the assumption of a strong association existing in the physical world between quantity and position along the vertical dimension (Lakoff & Johnson, 1980; Tversky et al., 1991). If vertical bars prompted to a larger extent an association between height and quantity, a larger number of responses corresponding to spatial-to-conceptual mappings should have been observed for all kinds of vertical graphs, as compared to horizontal graphs.

Instead, the interaction observed between orientation and type of conflict is consistent with the hypothesis that the orientation of *conventional features* (i.e., the scale) affects the likelihood that such features will be incorporated in viewers' interpretations. In horizontal graphs numbers in the scale are oriented in a way that matches Westeners' reading habits (i.e., along the horizontal dimension), and this might facilitate the task of reading and integrating the values shown. Additionally, horizontal graphs are less prevalent than vertical graphs (Kosslyn, 2006; Tversky, 2001). Thus, the horizontal orientation might also motivate viewers to engage in a more thorough exploration of graphs. The finding that the change in orientation only reliably affected performance of highly graph literate individuals further supports the notion that higher levels of knowledge of graphic conventions help override the reliance on spatial-to-conceptual mappings grounded in the physical world. When provided with a format that may encourage a more thorough exploration of graphs and that can facilitate reading values on the scale, highly graph literate individuals showed a larger tendency to incorporate this information in their interpretations, while the performance of less graph literate individuals remained unaffected.

Taken together, our findings show that individual differences in graph literacy can be linked to differences in the likelihood to engage in the general processes outlined in theoretical graph comprehension models (i.e., encoding of the visual pattern, translating visual features into conceptual relations, and determining the referents of the concepts; Carpenter & Shah, 1998; Lohse, 1993; Pinker, 1990). When spatial features can be readily translated into information about quantities through knowledge acquired in the environment, less graph literate individuals show a bias toward basing their interpretations of graphs primarily on such translations (i.e., to rely to a larger extent on the second process involved in graph comprehension). It is not clear whether the larger tendency of less graph literate individuals to neglect information in conventional features is driven mainly by difficulties to integrate information contained in such features with the corresponding bars on the chart (i.e., third process involved in graph comprehension) or if this mainly reflects a failure to sufficiently encode and elaborate on the relevant features. In any case, these processes are interrelated, as integration processes entail subsequent cycles of encoding of conventional features and data points (Carpenter & Shah, 1998; Huestegge & Philipp, 2011; Shah & Carpenter, 1995). A more precise account of the differences in the time course of underlying cognitive dynamics is beyond the scope of the current investigation and methods. Ongoing research using cognitive process tracing methodologies (i.e., eye tracking; reaction time analysis) is investigating these issues (Okan, Galesic, & Garcia-Retamero, 2013; Woller-Carter, Okan, Cokely, & Garcia-Retamero, 2011).

Notably, our results highlight that associations acquired through experience with the physical world can constitute a basis for the translation of spatial information onto information about more abstract quantities in graphs. This is consistent with the theory that off-line cognition is affected by our interactions with the physical world (Fischer & Zwaan, 2008; Zwaan & Taylor, 2006). That is, our findings converge with the perspective on embodied cognition suggesting that off-line cognitive activity—activity that is detached from direct physical inputs and outputs and that entails manipulating elements that are not directly present—is often rooted in knowledge acquired via interactions with our environment (Wilson, 2002; but for related arguments concerning the ecological grounding of cognition see Gigerenzer, Todd, & the ABC research group, 1999; Simon, 1996). Our interactions with the physical world shape the way in which we construct and infer information from external spatial representations such as graphs (Tversky, 2009). Notably, our findings point to individual differences in skill as a moderator of the extent to which people use embodied processes to interpret abstract information (see Madden & Zwaan, 2006; for related differences in elaborative encoding and abilities see Cokely & Kelley, 2009; Cokely, Kelley, & Gilchrist, 2006).

Our findings also contribute to our theoretical understanding of the mechanisms underlying graph comprehension in individuals of varying skill levels. However, it is important to acknowledge that the graph comprehension mechanisms outlined in the present paper may not generalize to different kinds of visual displays. Future research should aim to identify the mechanisms underlying comprehension for displays of varying complexity (see e.g., Canham & Hegarty, 2010; Trickett & Trafton, 2006) as well as to pinpoint the commonalities and differences in such mechanisms across displays.

The current investigation also has practical implications. First, it highlights common means by which graphical communication can be distorted, capitalizing on biases and causing judgment errors. Graphical displays are increasingly being used and recommended for the communication of medical risks to the public (Ancker et al., 2006; Lipkus, 2007). Our results suggest that caution should be taken to ensure that viewers of varying skill levels infer the correct meaning from graphs. Furthermore, they provide converging evidence on the effect of manipulations of values on the scales (e.g., variations in the range of values along the y axis) on viewers' judgments and decisions, which have been studied extensively in the literature of impression management with graphs (Arunachalam, Pei, & Steinbart, 2002; Pennington & Tuttle, 2009).

Second, our results suggest that some formats may be more prone to mislead viewers than others. We documented that participants more often made erroneous inferences when graphs containing a conflict between spatial features and values on the scale were oriented vertically. Vertical formats are more prevalent than horizontal formats (Kosslyn, 2006; Tversky, 2001) and experimental studies have demonstrated that vertical bar graphs can favor faster decision times concerning the quantities represented (Fischer et al., 2005). However, our findings suggest that horizontally oriented graphs can, in some cases, encourage the integration of important information contained in elements such as numerical scales, leading to an enhanced comprehension for some viewers.

As with all studies, the current work has a number of limitations. First, specific dispositions of the elements of the stimulus materials were created to foster high internal validity and allow clear theory evaluation. Accordingly, it is difficult to precisely estimate the ecological validity of these materials or the frequency with which related design features are present in medical or other graph-based communication. Nevertheless, research indicates that many graphs that are available to the public often do include misleading characteristics similar to those manipulated in the present study, such as improperly scaled axes (Beattie & Jones, 2002; Cooper, Schriger, Wallace, Mikulich, & Wilkes, 2003) or longer bars representing lower values (Kosslyn, 2006). It should also be noted that only four different sets of bar graphs were used as stimuli in the current study. As well, both the materials and the graph literacy instrument used focused on judgment and decision making in the medical domain. Thus, more research is needed before offering public policy implications. Future research should include more diverse and ecological materials along with a higher-fidelity examination of associated cognitive dynamics. Relevant research projects are currently underway in our laboratories with emphasis on comprehension processes in graphs used to communicate with the public and with professionals across a variety of domains (e.g., actual political, medical, and consumer communication; Woller-Carter et al., 2011).

In conclusion, we have demonstrated that associations acquired through experience with the physical world can constitute a robust basis for the translation of spatial information onto information about quantities depicted in graphs. Moreover, we have documented a link between embodiment and judgment bias. In the current experiment this bias led to robust medical judgment and decision making errors. However, it is important to note that in other environments such biases can also be adaptive and lead to superior performance (Gigerenzer & Brighton, 2009). Of note, the current study also showed that the observed judgment biases were moderated by individual differences. Individuals who were higher in graph literacy showed more flexible interpretations of graphs and were less likely to show judgment and decision errors. Ultimately, a precise theoretical understanding of the nature and causes of our judgment biases allows the anticipation of potential errors and development of improved educational interventions. Accordingly, the current findings provide theoretical links to fundamental embodied and ecological mechanisms that give rise to more and less effective graphical comprehension. Such findings can play a central role in the development of custom-tailored decision support systems built to inoculate professionals, policy makers, and the general public against potentially distorted and misleading communication.

Acknowledgments

This study is part of the projects, “How to Improve Understanding of Risks about Health (PSI2008-02019)” funded by the Ministerio de Ciencia e Innovación (Spain), and “Helping Doctors and Their Patients Make Decisions about Health (PSI2011-22954)” funded by the Ministerio de Economía y Competitividad (Spain).

CHAPTER V.

HOW PEOPLE WITH LOW AND HIGH GRAPH LITERACY PROCESS GRAPHS: EVIDENCE FROM EYE-TRACKING

How People with Low and High Graph Literacy Process Health Graphs: Evidence from Eye-Tracking

Abstract

Graphs facilitate the communication of important quantitative information needed to understand and manage risks. Yet people differ substantially in their graph literacy—the ability to understand graphically presented information. Although some features of graphs can be interpreted using spatial-to-conceptual mappings that are non-arbitrary and can be established by adults and children with no graphing experience (e.g., "higher bars equal larger quantities"), other features are linked to arbitrary graph conventions (e.g., axis labels and scales). In two experiments, we examined differences in the processes underlying the comprehension of graphs presenting medical information in individuals with low and high graph literacy. Participants' eye movements were recorded while they interpreted graphs in which information in conventional features (e.g., axis labels, values on scales) was incongruent with that conveyed by spatial features (e.g., heights of bars). Results revealed that participants with low graph literacy more often relied on misleading spatial-to-conceptual mappings and misinterpreted the data depicted. When graphs contained essential information in *y* and *x* axis scales, participants with low graph literacy spent less time viewing those regions compared to participants with high graph literacy. Differences in viewing times mediated the effect of graph literacy on accuracy of graph comprehension. These findings suggest that graph literacy affects people's tendency to direct attention to and encode some conventional features. This tendency, in turn, affects performance. Theoretical, methodological, and prescriptive implications for customization of decision-support systems are discussed.

Submitted as:

Okan, Y., Galesic, M., & Garcia-Retamero, R. (submitted). How people with low and high graph literacy process graphs: Evidence from eye tracking. *Journal of Behavioral Decision Making.*

Introduction

Graphical displays such as line plots, bar charts, and icon arrays can serve as highly valuable tools for overcoming difficulties in the comprehension of numerical concepts, enhancing accuracy in evaluation of risks (Ancker, Senathirajah, Kukafka, & Starren, 2006; Garcia-Retamero & Cokely, 2011; Lipkus, 2007). Unfortunately, graphs are not equally useful for all individuals, as people in the general population differ substantially in their ability to understand graphically presented information (Galesic & Garcia-Retamero, 2011b; Kutner, Greenberg, Jin, & Paulsen, 2006). These differences can affect the extent to which individuals benefit from visual displays (Gaissmaier et al., 2012; Garcia-Retamero & Galesic, 2010; Okan, Garcia-Retamero, Cokely, & Maldonado, 2012). Yet the processes underlying the effect of different displays in individuals with varying levels of graph literacy are not well understood. We used eye-tracking methodology to investigate how differences in *graph literacy* affect the processes underlying graph comprehension.

Graph literacy refers to one's ability to understand graphically presented information and includes general knowledge about making inferences from different graphic formats (Freedman & Shah, 2002; Shah & Freedman, 2011). That is, graph literacy constitutes one of the top-down influences that can affect graph interpretations, which can act in conjunction with other top-down influences such as the viewer's knowledge of the content being depicted and expectations about the data (Freedman & Shah, 2002; Novick, 2006; Shah, Freedman, & Vekiri, 2005). Like other types of literacy (e.g., prose and document literacy; Kutner et al., 2006), higher graph literacy has been found to be associated with higher educational levels (Galesic & Garcia-Retamero, 2011b), highlighting that developing this skill requires knowledge acquired through formal education and experience with graphs.

Graph literacy can include mental representations stored in long-term memory that contain knowledge about the properties of different kinds of displays and procedures for interpreting them (i.e., *graph schemas*; Maichle, 1994; Peebles & Cheng, 2001, 2003; Pinker, 1990; Ratwani & Trafton, 2008). Individuals with higher graph literacy may have more complete schemas, which can contribute to recognizing specific types of graphs, identifying the most relevant features in each graph, and making accurate interpretations of the information depicted. In line with this idea, individuals with high graph literacy have been found to direct their attention to the information that is most typical for the specific kind of graph viewed (e.g., quantitative trend for line graphs; Maichle, 1994) and to generate the relevant inferences that are supported by the type of graph (Shah & Freedman, 2011). It is not clear, however, how individual differences in graph literacy affect the processes involved in graph comprehension.

Prominent graph-comprehension models have identified three types of processes in which viewers engage when making inferences from graphical displays, such as line or bar graphs (e.g., Carpenter & Shah, 1998; Kosslyn, 1989; Lohse, 1993; Pinker, 1990; Simkin & Hastie, 1987). The first is *encoding the visual pattern* to identify the principal features in graphs (e.g., lines with different slopes), and it involves making different visual judgments of the elements (e.g., judgments of position along a scale, slope, length, or angle; Cleveland & McGill, 1986).

The second process is the *translation of the identified visual features into conceptual relations*. For example, variations in the size of spatial features (e.g., bars of different heights) can be used to indicate variations in the quantity of the variables represented. Spatial features are those contained in the pattern, including bars of different heights, or lines following an increasing or decreasing trend. There is evidence suggesting that these translations of spatial into conceptual information—*spatial-to-conceptual mappings*—are non-arbitrary and governed by general cognitive constraints, as certain mappings (e.g., "high equals more," "steeper equals faster") emerge consistently in adults and children with no graphing experience (Gattis, 2002, 2004). As shown by Gattis (2002, 2004) one of these constraints stems from the sensitivity to similarities of relational structure, which leads people to pair spatial elements with conceptual elements, (e.g., quantity and height), and spatial relations with conceptual relations (e.g., rate and slope). Another important constraint comes from viewers' experience with their physical environment (Tversky, 2001, 2009). To illustrate, in the physical world, larger quantities of substances typically reach higher positions along the vertical dimension (Lakoff & Johnson, 1980; Tversky, Kugelmass, & Winter, 1991). By applying this real-world experience to graphs, viewers can infer that higher data points represent larger values. Hence, often spatial features (e.g., bars of different heights) can convey meaning independent of viewers' level of graph literacy.

The third process involves *determining the referents of the concepts identified* by associating them with the specific variables shown in the graph and their numerical values (Carpenter & Shah, 1998; Shah & Carpenter, 1995). This process entails identifying and inferring information from *conventional features* in graphs, including the title of the graph, axis labels, legends, and numerical values on the scales, and integrating this information with that extracted in the first two processes. Conventional features are determined by arbitrary graph conventions and typically do not map onto viewers' experience with their environment

(Okan, Garcia-Retamero, Galesic, & Cokely, 2012). Viewers with low graph literacy may be less likely to have schemas incorporating arbitrary graph conventions and thus can be less prone to identify the relevant conventional features and to incorporate this information in their interpretations. In contrast, high graph literacy may help viewers to more readily identify and incorporate relevant conventional features in graphs. Additionally, graph-related knowledge could direct subsequent cycles of encoding and interpretation (Carpenter & Shah, 1998; Pinker, 1990), directing attention to labels or scales that contain information required to answer questions about the data.

The relevance of identifying and inferring information from conventional features can vary depending on specific properties of the graphical displays. If information conveyed by spatial features (e.g., bar heights) is congruent with that conveyed by conventional features, viewers could neglect conventional features and nevertheless reach correct interpretations by relying on spatial-to-conceptual mappings. However, if such congruency does not exist, identifying and inferring information from conventional features becomes critical to reach a correct interpretation. This can occur when spatial features of the graph convey a different meaning from textual information in the title and axis labels (*textual conflicts*) or numerical values on the scale (*scale conflicts*). For instance, a graph with a textual conflict might present the percentage of people *without* different types of allergy (as indicated in the title and axis label), implying that higher bars do not represent more prevalent allergies. In such cases, taking into account information in conventional features is crucial to override spatial-to-conceptual mappings and avoid misinterpretations (e.g., recognizing that higher bars do not necessarily imply larger quantities).

Differences in the tendency to neglect information in conventional features could arise from at least two mechanisms, which are linked to the third process outlined above. One possibility is that people with low and high graph literacy differ in the extent to which they attend to the relevant conventional features (i.e., less graph literate individuals might attend to such features to a lesser extent). This is in line with the information reduction framework proposed by Haider and Frensch (1996, 1999). This framework suggests that more skilled individuals acquire the ability to distinguish between task-relevant and task-redundant information and focus on the former, whereas less skilled individuals are not able to make such distinctions. This framework emphasizes the notion that skill acquisition leads to differences in attention allocation (i.e., differences in terms of *which* information is processed and not only *how* the information is processed; for related arguments concerning the effect of expertise on the use of relevant cues in judgment and decision making tasks, see Shanteau, 1992).

Consistent with the information reduction framework, eye-tracking studies investigating the effect of expertise and instruction in different domains, including meteorology (Canham & Hegarty, 2010), biological classification (Jarodzka, Scheiter, Gerjets, & van Gog, 2010), and driving (Di Stasi, Contreras, Cándido, Cañas, & Catena, 2011), have revealed that experts fixate to a larger extent on areas that are relevant for the task at hand. Additionally, research on problem solving and risky decision making has revealed that individuals who score high on cognitive-ability measures spend more time encoding the elements that are required to reach a solution (Sternberg, 1977) and engage to a larger extent in elaborative heuristic search (i.e., a more thorough exploration of the problem space), pointing to strategic differences in elaboration during encoding (Cokely & Kelley, 2009; Cokely, Kelley, & Gilchrist, 2006; see also Woller-Carter, Okan, Cokely, & Garcia-Retamero, 2012). This line of research also suggests the existence of a link between superior performance and knowledge-driven differences in attention allocation and encoding. Accordingly we expected that, for graphs containing the kinds of conflicts described above, differences should exist in the viewing time of the relevant conventional features between individuals with low and high graph literacy.[13]

An alternative possibility is that graph literacy may not affect attention allocation or encoding, and differences in interpretation could result primarily from differences in conceptual understanding about the meanings of elements of graphs, and mental operations on elements of graphs. That is, less graph literate individuals might attend to the relevant conventional features to the same extent as individuals with high graph literacy but may fail to incorporate this information at a conceptual level in their interpretations (for a related distinction in terms of perceptual and conceptual stages, see Haider & Frensch, 1999). If this is the case, no differences should be observed in the viewing time of relevant conventional features between people with high and low graph literacy.

To determine if and how graph literacy affects allocation of attention to conventional features, we conducted two experiments in which we recorded the eye movements of

[13] It should be noted that another prediction that derives from the information reduction framework is that higher skill will lead to an increase in the speed with which tasks are processed, as a consequence of the acquired ability to ignore task-redundant information. However, we do not expect this to be applicable to the tasks used in the current research, as none of the information depicted in the graphs can be considered redundant in this sense.

participants with low and high graph literacy while they interpreted line graphs and bar graphs displaying quantitative medical information (i.e., prevalence of different diseases or effects linked to different treatments). In both experiments we included a set of graphs constructed in such a way that following spatial-to-conceptual mappings grounded in experience with the environment would lead to erroneous interpretations. Additionally, in Experiment 2 we also included a set of graphs where following spatial-to-conceptual mappings would lead to correct interpretations.

Experiment 1

Experiment 1 included four graphs containing conflicts between spatial and conventional features. In two of the graphs essential information was included in the title and in the textual label for the *y* axis (textual conflict); the other two contained essential information in the numerical scale for the *y* axis (*y*-axis-scale conflict). Taking into account the roles of prior knowledge outlined above, we proposed three hypotheses. First, in line with recent findings reported by Okan et al. (2012b), we predicted that low graph literacy would be associated with a larger tendency to interpret graphs on the basis of spatial-to-conceptual mappings. For graphs containing conflicts, this should be reflected in a larger proportion of incorrect responses corresponding to mappings among participants with low graph literacy (e.g., they might often assume that the highest bar represents the highest value) (Hypothesis H1).

Second, in line with the information reduction framework proposed by Haider and Frensch (1996, 1999), we expected that individuals with low graph literacy would be less likely to recognize and attend to the conventional features that are essential to reach a correct interpretation, according to the conflict present in the graph. This should be reflected in relatively longer times spent viewing such features for participants with high graph literacy, as compared to those with low graph literacy (H2). Third, we expected that the longer viewing time of the relevant conventional features would mediate the relationship between graph literacy and the proportion of incorrect responses corresponding to mappings (H3).

Method

Participants.

A total of 52 participants were recruited from the respondent pool of the Max Planck Institute for Human Development in Berlin. Technical problems prevented recording the eye movements of 4 participants. Thus, the final sample consisted of 48 participants (50%

female), mean age of 25.7 years (*SD* = 3.3, range 19–34 years), 52% with up to high school education and 48% with at least some college. Participants were paid 10 euros for taking part in the study.

Materials.

Eye-tracking equipment. Participants' eye movements were recorded by a Tobii T120 Eye Tracker. In this system the eye-tracking cameras are integrated into a 17-in. thin film transistor monitor, allowing for unobtrusive recording of respondents' eye movements. The documentation of the T120 describes its accuracy to be within 0.5° with less than 0.3° drift over time and less than 1° as a result of head motion. It allows for head movement within a volume of 30 × 22 × 30 cm centered 70 cm from the camera. The sampling rate is 120 Hz. To define fixations we used the built-in fixation filter available in Tobii Studio (v. 2.0.3) with a fixation radius of 30 pixels on a screen with a resolution of 1,280 × 1,024 pixels. For all analyses we took into account fixations that lasted at least 100 ms, as this decreases noise in the data (Peebles & Cheng, 2003).

Stimuli. We constructed four graphs presenting medical information, such as prevalence of different diseases and effects linked to different treatments. In two of the graphs, essential information was included in the numerical scale for the *y* axis (graphs with *y*-axis-scale conflicts; see graphs G1 and G2 in Appendix A); the other two contained essential information in the title and in the textual label for the *y* axis (graphs with textual conflicts; see graphs G11 and G12 in Appendix A). To illustrate, one of the graphs involving a scale conflict was a line graph presenting data about the percentage of people with a fictitious disease. The numerical scale on the *y* axis was inverted (i.e., values increased from top to bottom; see graph G1 in Appendix A). Participants were asked to find the year in which the percentage of people with the disease was highest. To answer this question correctly, participants had to attend to the scale to infer that the usual spatial-to-conceptual correspondence between height and quantity was reversed. An example of the graphs involving textual conflicts is a bar graph presenting data about percentages of people *without* a fictitious disease in different clinics (see graph G11 in Appendix A). Participants were asked to identify the clinic in which the percentage of people with the disease was highest. To answer this question correctly, participants had to attend to the title and the label for the *y* axis to infer that the usual spatial-to-conceptual mapping was reversed (i.e., they had to infer that higher bars represented lower values). All materials were implemented as a Web survey using the platform Unipark (www.unipark.de).

Coding of eye fixations. For each graph we defined a set of areas of interest (AOIs) corresponding to the *conventional features* containing essential information to answer the question, according to the types of conflicts present (i.e., titles, labels for the *y* axes, and scales on the *y* axes). For each participant, we computed the total time spent viewing each of the AOIs, which constituted the sum of the duration of all fixations on the AOI. The number of fixations and the total time spent viewing each AOI were highly correlated (mean correlation = .95 across all variables computed). For the sake of simplicity we report only the results for total viewing times.

Measurement of graph literacy. Graph literacy was measured using the scale developed by Galesic and Garcia-Retamero (2011b; see also Garcia-Retamero & Galesic, 2010). This scale consists of 13 items dealing with the communication of medical risks, treatment efficiency, and prevalence of diseases, and covers four frequently used graph types—line plots, bar charts, pies, and icon arrays. Because the scale was designed for the general population, to achieve better differentiation of graph literacy in our somewhat better educated sample, we also included four more difficult items from other scales. In particular, we included one item from the Kramarski and Mevarech Graph Interpretation Test (Kramarski & Mevarech, 2003), two items assessing graph comprehension from the International Adult Literacy Survey (Tuijnman, 2000), and one item from the National Assessment of Adult Literacy (Kutner et al., 2006). The total score for each participant was computed by adding the score for these four items to the score obtained in the scale developed by Galesic and Garcia-Retamero (2011b).

For some analyses, we split participants into two groups according to the median graph literacy score for the total sample (i.e., 14.5 of the total 17). Thus, the group of participants with low graph literacy included those who obtained 14 or fewer correct responses (n = 24), while the group of participants with high graph literacy included those who obtained 15 or more correct responses (n =24). Participants with low graph literacy answered on average 12.5 (SD =1.6) items correctly, while participants with high graph literacy answered on average 16.2 (SD = 0.8) items correctly.

Measurement of numeracy. We also assessed participants' numeracy skills (i.e., the ability to understand and manipulate different numerical expressions of probability; Lipkus, Samsa, & Rimer, 2001; Peters, 2012). Participants' numeracy was measured using the 11 items included in the general and expanded numeracy scales developed by Lipkus et al. (2001).

Procedure.

The experiment took on average 23.2 min (*SD* = 5.7) to complete and included three sections. In the first section, participants signed a consent form that described the eye-tracking procedure and successfully completed a standardized calibration exercise. They were then presented with the four graphs depicting medical information. In the second section, participants completed Galesic and Garcia-Retamero's (2011b) graph literacy scale. In the third section, participants completed the four more difficult items measuring graph literacy described above and the numeracy scale, and they answered some demographic questions. As calibration can decrease in accuracy over time, respondents were recalibrated at the beginning of each new section. The study was approved by the Ethics Committee of the Max Planck Institute for Human Development.

Results

Does Graph Literacy Affect Interpretations of Graphs With Conflicts? The average proportion of correct responses to the questions across graphs was 56% (*SE* = 10.6), while the average proportion of incorrect responses corresponding to spatial-to-conceptual mappings (mapping responses, e.g., assuming that the highest value is the one represented by the highest bar) was 37% (*SE* = 9.5). As expected, the average proportion of incorrect responses that were not related to the mapping was low (7%; *SE* = 1.9), indicating that the majority of participants who misinterpreted the graphs did so on the basis of direct spatial-to-conceptual mappings. Thus, for the analyses presented below we computed for each participant the percentage of items in which she or he had provided the incorrect response corresponding to the spatial-to-conceptual mapping (mapping response), for each type of conflict (i.e., textual vs. y-axis-scale conflicts).

The average percentage of mapping responses among participants with low graph literacy was 42% (*SE* = 7.2) for y-axis-conflict graphs and 56% (*SE* = 6.9) for textual-conflict graphs. In contrast, participants with high graph literacy showed on average 27% (*SE* = 5.1) mapping responses for y-axis-conflict graphs and 23% (*SE* = 6.0) for textual-conflict graphs. A 2 × 3 analysis of variance (ANOVA) with graph literacy as between-subjects factor and type of conflict as within-subject factor on the average percentage of mapping responses revealed a main effect of graph literacy, $F(1,46) = 15.38$, $p = .001$, $\eta_p^2 = .25$, supporting H1. All other effects were not reliable (*F*s < 2, *p*s > .1). The average percentage of mapping responses did not reliably vary as a function of numeracy. The correlation of graph literacy with numeracy was .38.

Does Graph Literacy Affect the Viewing Time of Relevant Conventional Features? To answer this question we examined the distributions of time spent viewing the different AOIs. As distributions were skewed right, the analyses that will be presented below were conducted using log-transformed values. Table 1 shows raw and log-transformed mean viewing times for the different areas of the graph.

Table 1. Raw and log-Transformed Mean Times Spent Viewing the Relevant Areas of the Graphs and Total Viewing Times in Experiment 1, as a Function of Type of Conflict and Graph Literacy (SEM in Parentheses)

Area of graph	Graphs with *y*-axis-scale conflict		Graphs with textual conflict	
	Low graph literacy	High graph literacy	Low graph literacy	High graph literacy
y-axis scale				
Time	**2.59**	**3.10**	1.15	1.47
	(0.49)	**(0.46)**	(0.17)	(0.24)
Log time	**0.29**	**0.69**	-0.21	0.00
	(0.21)	**(0.19)**	(0.15)	(0.16)
Title & *y*-axis label				
Time	4.34	5.24	**4.96**	**6.44**
	(0.67)	(0.63)	**(0.65)**	**(0.64)**
Log time	1.10	1.38	**1.31**	**1.68**
	(0.15)	(0.12)	**(0.16)**	**(0.10)**
Graph total				
Time	14.98	15.43	12.97	15.49
	(1.46)	(1.39)	(1.05)	(1.39)
Log time	2.58	2.59	2.47	2.62
	(0.09)	(0.09)	(0.08)	(0.09)

Note: Relevant conventional features for each type of conflict are marked in bold.

As our hypothesis was concerned with the time spent viewing the relevant conventional features in each case, we performed a 2 × 3 ANOVA with graph literacy as between-subjects factor and type of conflict as within-subject factor on the mean log-transformed time spent viewing the relevant conventional features. This analysis revealed a main effect of graph literacy, $F(1,46) = 4.11$ $p = .048$, $\eta_p^2 = .08$, indicating that participants with high graph literacy spent more time fixating on the areas containing essential information in each graph, in line with H2. The analysis also revealed a main effect of type of conflict, $F(1,46) = 47.61$, $p = .001$, $\eta_p^2 = .51$, as the mean viewing time was higher for the textual elements (i.e., *y*-axis label and title) than for the *y*-axis scale.

In contrast, the total time that participants spent viewing the graphs did not vary as a function of graph literacy ($F < 1$, *p*s > .3). No significant differences were observed as a function of numeracy in the time spent viewing the relevant conventional features or in the total viewing time.

Is the Effect of Graph Literacy on Accuracy Mediated by Differences in the Viewing Time of Conventional Features? We conducted mediational analyses to determine if the influence of graph literacy on the tendency to show incorrect mapping responses was mediated by the time viewing spent conventional features, for graphs with *y*-axis and textual conflicts. In a regression equation, graph literacy scores significantly predicted the percentage of incorrect mapping responses for both *y*-axis-conflict graphs, $\beta = -.29$, $t = -2.03$, $p = .049$, and textual-conflict graphs, $\beta = -.48$ $t = -3.7$, $p = .001$, with higher graph literacy associated with fewer incorrect responses. Furthermore, graph literacy scores significantly predicted the mean log-transformed times spent viewing the relevant conventional features for *y*-axis-conflict graphs, $\beta = .30$, $t = 2.1$, $p = .042$, and for textual-conflict graphs, $\beta = .31$, $t = 2.2$, $p = .033$. Participants with higher graph literacy spent more time fixating on the relevant conventional features. When the mean log-transformed time spent viewing the conventional features was included in the regression equation for *y*-axis-conflict graphs, viewing time predicted the number of incorrect mapping responses, $\beta = -.41$, $t = -2.97$, $p = .005$, whereas the direct effect of graph literacy on mapping responses was no longer significant, $\beta = -.17$, $t = -1.21$, $p = .230$ (see Figure 1a). In contrast, for textual-conflict graphs when the mean log-transformed time spent viewing the conventional features was included in the regression equation, viewing time did not predict the number of incorrect mapping responses, $\beta = -.14$, $t = -1.06$, $p = .294$, whereas the direct effect of graph literacy on mapping responses remained significant, $\beta = -.44$, $t = -3.2$, $p = .002$.[14]

[14] Results remained unchanged when either the time spent viewing the *y*-axis label or the title were included in regression equations, instead of the sum of the time spent viewing both areas.

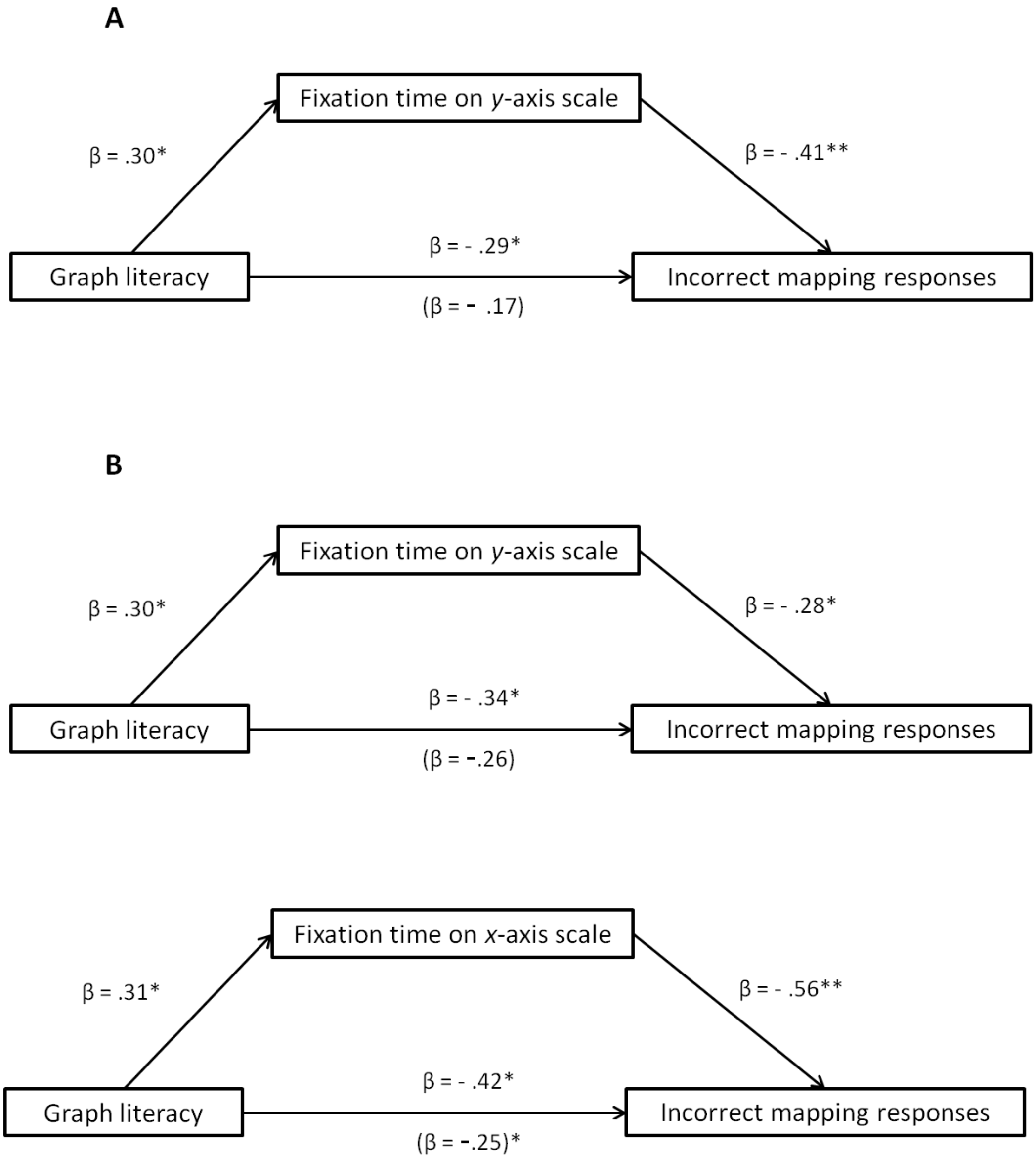


Figure 1. Mediational analyses of the effect of graph literacy on the percentage of incorrect mapping responses, and the mediational effect of viewing time of relevant conventional features. The value in parentheses shows the relationship between graph literacy and incorrect mapping responses after controlling for viewing time of conventional features. (a) Results for Experiment 1, for graphs with *y* axis-scale conflicts; (b) Results for Experiment 2, for graphs with *y*-axis-scale conflicts (top panel) and *x*-axis-scale conflicts (bottom panel). Note: Standardized coefficients are shown. $^*p < .05$, $^{**}p < .01$.

To test the mediation effect for *y*-axis-conflict graphs we performed a bootstrap analysis with 1,000 resamples using the PROCESS procedure developed by Hayes (2013). The analysis provided evidence of a significant indirect effect of time spent viewing the *y*-axis scale: The mean estimate was –1.66, within the 95% confidence interval [–3.76, –0.36]. This supports H3. Finally, to further explore more general patterns of eye fixations, we performed additional analyses examining the different types of transitions between different regions, following Carpenter and Shah (1998). Results of these further analyses can be found in Appendix B.

Discussion

The results of Experiment 1 show that people tend to make erroneous inferences indicating an overreliance on spatial-to-conceptual mappings. Crucially, the tendency to rely primarily on such mappings was larger for less graph literate individuals, in line with our predictions. This tendency held both when the essential information to override the mappings was conveyed in textual information in the title and *y*-axis label (i.e., textual conflicts) and when it was included in numerical values on the *y*-axis scale (i.e., *y*-axis-scale conflicts). Of note, Okan et al. (2012b) found a similar pattern of findings for graphs that were oriented horizontally. However, when graphs were oriented vertically (as in the present study), the larger tendency to misinterpret graphs among less graph literate individuals was reliable only for graphs with textual conflicts and not for graphs with *y*-axis-scale conflicts. However, the study reported by Okan et al. (2012b) was conducted online and the stimulus set employed in that study was related but not equivalent to that used in the present study. The inclusion of a larger set of stimulus materials for each type of conflict in Experiment 2 enabled us to determine the extent to which the pattern of findings observed in Experiment 1 generalizes to a more diverse set of graphs.

In line with our predictions, analyses of participants' eye movements in Experiment 1 revealed that less graph literate participants spent less time fixating on the relevant conventional features containing essential information in each case, as compared to highly graph literate participants. These findings are in line with the information reduction framework proposed by Haider and Frensch (1996, 1999), according to which skill acquisition leads to differences in attention allocation, with more skilled individuals focusing more on task-relevant information. Although in the task of graph comprehension all regions of a graph are arguably relevant to reach an accurate interpretation, those determined by arbitrary graph conventions (e.g., axis labels, values on scales) become particularly relevant when graphs contain the conflicts described above. Viewers with low graph literacy can be less likely to have stored graph schemas (Maichle, 1994; Pinker, 1990) that would direct their attention to conventional features. This might lead them to spend less time viewing these features. That is, the findings of Experiment 1 support the notion that lower levels of graph literacy are associated not just with a failure to understand and integrate information in conventional features at a conceptual level, but also with a tendency to spend less time encoding them.

Importantly, the total time spent viewing the graphs did not vary reliably as a function of graph literacy. This finding indicates that individuals with high graph literacy do not just engage in a more thorough exploration of all regions of the graphs but instead allocate more attention to those regions containing the most relevant information for the task at hand.

Finally, in line with our hypotheses we also found that the influence of graph literacy on the tendency to show incorrect responses could be accounted for by the time spent viewing numerical scales on the *y* axis. However, this was not the case for the time spent viewing textual elements (i.e., title and *y*-axis labels). For *y*-axis-conflict graphs, the full mediation observed suggests that the shorter times spent viewing scales among individuals with low graph literacy satisfactorily accounted for their performance. In contrast, for graphs with textual conflicts, the absence of a significant indirect effect of the time spent viewing textual elements suggests that the effect of graph literacy on performance is mediated by a factor other than viewing times (for a discussion concerning different types of mediation, see Zhao, Lynch, & Chen, 2010). However, the small number of graphs employed in Experiment 1 prompts us to suggest caution in our interpretations, as the observed pattern of findings may not generalize to a more diverse set of graphs.

Experiment 2

Experiment 2 was designed to address four new questions. First, as noted above, we sought to determine whether the findings observed in Experiment 1 would generalize to a more diverse set of graphs and types of conflict. To this end, we expanded our set of stimuli to include four additional graphs with textual conflicts and four additional graphs with *y*-axis-scale conflicts, as well as four graphs containing essential information in the *x*-axis scale (i.e., *x*-axis-scale conflicts; see graphs G7 to 10 in Appendix A). Second, as Experiment 1 included only graphs containing conflicts, it was not possible to determine to what extent such conflicts affected interpretations and viewing times, as compared to graphs without conflicts. Therefore, in Experiment 2 for each of the graphs with conflicts we constructed an equivalent graph without conflict. This enabled us to determine the extent to which the inclusion of conflicts affected interpretations and viewing times. In addition, these nonconflict graphs enabled us to determine individual baseline viewing times and to control for them when analyzing viewing times in conflict graphs, as will be described below.

Third, in Experiment 2, we sought to examine the effect of content knowledge in graph comprehension processes. Knowledge of the content being depicted can affect graph interpretations independently of graph literacy (Freedman & Shah, 2002; Shah & Freedman, 2011; Shah et al., 2005). For instance, viewers with some content knowledge can rely on expected relationships in the data to interpret graphs and to avoid potential errors that could occur by relying solely on spatial relations (Shah & Freedman, 2011). The graphs used in the present studies included relatively abstract information in most cases, even if embedded in the medical context (e.g., “Treatments A, B, C”), and therefore we did not expect that medical knowledge would be associated with strong expectations concerning the trends depicted. However, prior medical knowledge could potentially aid in the comprehension of the information conveyed in conventional features, thus helping to avoid wrong interpretations. Therefore, in Experiment 2 we also measured participants’ level of knowledge concerning relevant clinical conditions (e.g., myocardial infarction and stroke).

Finally, in Experiment 2 we also aimed to examine if the differences observed between participants with low and high graph literacy in Experiment 1 were driven by other confounding factors. To illustrate, participants with low graph literacy may merely be more careless in interpreting the information, thus making more errors both while completing the graph literacy scale and while processing the graphs presented in the experiment. Alternatively, less graph literate individuals may simply be less knowledgeable about the fact that data can be represented misleadingly in graphs, and this may affect the way they explore graphs. Additionally, we also measured numeracy using a recently developed scale (the Berlin Numeracy Test; Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) that has better psychometric properties than the numeracy scale used in Experiment 1.

We hypothesized that the effect of graph literacy on accuracy of understanding would be smaller for graphs without conflicts than for graphs with conflicts (H1a). The reason is that in the former type of graphs, conventional features and spatial features point to the same (correct) interpretation. In contrast, and in line with Experiment 1, for graphs with conflicts we expected that participants with low graph literacy would more often make erroneous interpretations corresponding to spatial-to-conceptual mappings (H1b). We also hypothesized that graph literacy would not affect times spent viewing conventional features in graphs without conflicts, as they did not contain essential information for accurate interpretations (H2a). In contrast, for graphs with conflicts we expected that participants with high graph literacy would spend a longer time than those with low graph literacy viewing the relevant conventional features in each case (H2b). Finally, we expected that the longer time spent viewing *y*- and *x*-axis scales would mediate the relationship between graph literacy and the proportion of incorrect responses (H3a). In line with Experiment 1 we expected that this would not be the case for the time spent viewing textual elements (H3b).

Method

Participants.

Ninety-one participants from the database of the Max Planck Institute for Human Development in Berlin were prescreened with the graph literacy scale used in Experiment 1 (13 items from Galesic & Garcia-Retamero, 2011b, plus 4 more difficult items). Their graph literacy scores ranged from 10 to 17, with a mean of 14.5 (*SD* = 1.6). We invited 38 participants in the top and bottom quartiles (i.e., with scores ranging from 10 to 13, and from 16 to 17). Additionally, given the limited availability of participants to take part in the laboratory experiment, we invited another 13 participants with scores from 14 to 15. Thus, the final sample included 51 participants (61% female), with a mean age of 25.3 years (*SD* = 4.7, range 18–38 years), 49% with up to high school education and 51% with at least some college. The group of participants with low graph literacy included those who obtained 14 or fewer correct responses (*n* = 24, mean score 12.7, *SD* = .9); the group of participants with high graph literacy included those who obtained 15 or more correct responses (*n* = 29, mean score 16.0, *SD* = .8). Participants were paid 10 euros for taking part in the study.

Materials.

Eye-tracking equipment. The eye-tracking equipment was identical to that used in Experiment 1, and fixations were determined using the same procedure.

Stimuli. In addition to the 4 graphs constructed in Experiment 1, we constructed 12 new graphs presenting medical information that contained conflicts. In 6 of the graphs essential information was included in the numerical scale for the *y*-axis (Graphs G1 to G6). In 4 graphs essential information was included in the numerical scale for the *x*-axis (Graphs G7 to G10). Finally, 6 graphs contained essential information in the title and in the textual label for the *y*-axis (Graphs G11 to G16). Description of all items is given in Table 2, and all graphs can be seen in Appendix A.

Coding of eye fixations. The same AOIs as in Experiment 1 were defined (i.e., titles, labels for the *y* axes, and scales on the *y* axes). Additionally, we defined an AOI corresponding to the scales on the *x* axes. All AOIs were defined for graphs with and without conflicts. The inclusion of graphs without conflicts enabled us to control for baseline individual variability in viewing times in analysis of graphs with conflicts, as will be described below. As in Experiment 1, the number of fixations and the total viewing times on each AOI were highly correlated (mean correlation = .94 across all variables computed for graphs with conflicts, and .93 for graphs without conflicts).

Measurement of graph literacy. Graph literacy was measured using the same items as in Experiment 1, for a total score of 17.

Measurement of numeracy. In addition to the numeracy scale used in Experiment 1, we administered the Berlin Numeracy Test—an adaptive measure of numeracy that was designed for more educated samples (Cokely et al., 2012).

Measurement of knowledge that graphs can be misleading and careless responses. To measure participants' knowledge that graphs can be misleading, we included six items developed by the current authors. Three items concerned graphs in general and three items focused on the medical domain (see Appendix C). Cronbach's alpha for the six items was .81. To identify careless responses, we administered the self-report participant engagement items developed by Meade and Craig (2012). Specifically, we used eight items selected from the Diligence subscale and the six items in the Interest subscale.[15] We included a response scale ranging from 1 (*Completely disagree*) to 4 (*Completely agree*). Additionally, we included the three single-item measures developed by Meade and Craig evaluating effort expended on the study (Effort), attention to the study (Attention), and whether the respondent felt his or her data should be used for analysis (Use Me), all adapted for the context of our study. All items were translated into German by a native German speaker with excellent knowledge of English and were reviewed by two of the authors.

[15] Item 7 from the Diligence subscale was excluded as it was not applicable to the context of the current study.

Table 2. Summary of the experimental materials

Item	Task
y-axis-scale conflict	
G1. Line graph with inverted scale (values increase from top to bottom), *x* axis below	Find the year in which the percentage of people with a disease was highest
G2. Bar graph with scale with negative values	Identify the therapy that resulted in the *lowest* change in the percentage of people with the disease
G3. Stacked bar graph with excised scale	Identify the ethnic group for which the proportion of people with one type of flu was larger than the proportion of people with another kind of flu
G4. Bar graph with scale with both positive and negative values and bars rising from lower *x* axis. Zero baseline not indicated.	Identify the treatment that resulted in the smallest change in patients' body weight
G5. Line graph with logarithmic scale	Find the year in which the difference between the number of men and women dying after suffering an infection was larger
G6. Line graph with inverted scale (values increase from top to bottom), *x* axis above	Find the age at which the recovery time from a disease was lowest
x-axis-scale conflict	
G7. Bar graph with inverted scale (values increasing from right to left)	Identify the pill that resulted in an increase in the values of a hormone over time
G8. Line graph with inverted scale (values increasing from right to left)	Identify the disease for which the number of affected people increased over time
G9. Bar graph with values not placed at proportional distances	Find the month after which patients' blood iron levels started to increase more slowly
G10. Line graph with values not placed at proportional distances	Find the week after which pain scores started to decrease more slowly
Textual conflict	
G11. Bar graph showing percentages of people *without* a disease	Identify the clinic in which the percentage of people *with* the disease was highest
G12. Bar graph showing the *change in the percentage* of people with different types of cancer during the previous year	Identify the type of cancer that affected the smallest *percentage* of people during the previous year
G13. Bar graph showing the number of *patients per doctor* in different countries	Identify the country that had the highest number of *doctors per patient*
G14. Line graph showing the percentage of people testing *negative* for a disease at different ages	Identify the age at which the percentage of people *diagnosed* with the disease was highest
G15. Line graph showing the number of *patients per nurse* in different years	Find the year in which the number of *nurses per patient* was lowest
G16. Line graph showing the percentage of people who *died* after different weeks of having been exposed to a virus	Find the week in which the percentage of people who *survived* after being expose to the virus was lowest

Note: Graphs G1, G2, G11, and G12 with conflicts were used in Experiment 1. All graphs were used in Experiment 2.

Procedure.

The experiment took on average 42 min (*SD* = 7.2) to complete and included three sections. In the first section participants signed a consent form and successfully completed a standardized calibration exercise. They were then presented with the 16 graphs without conflicts. In the second section, participants were presented with the 16 graphs with conflicts. In the third section, participants completed (1) the Berlin Numeracy Test, (2) the items assessing knowledge that graphs can be misleading, (3) the MMK questionnaire (Bachmann et al., 2007), (4) demographic questions, and (5) the items to identify careless responses (Meade & Craig, 2012). All remaining aspects of the procedure were identical to that of Experiment 1. The study was approved by the ethics committee of the Max Planck Institute for Human Development.

Results

Are Effects of Graph Literacy Confounded by Other Skills, Knowledge, and Motivational Factors? The correlation of graph literacy with numeracy measured with the Berlin Numeracy Test (Cokely et al., 2012) was .33, while it was .32 with the Lipkus et al. (2001) numeracy scale. This indicates that even though some of the same abilities might contribute to both graph literacy and numeracy, the amount of shared variance is relatively small. The correlation of graph literacy with MMK was –.09, indicating that no linear relationship existed between these variables. Finally, the correlations of graph literacy with knowledge that graphs can be misleading and with scales measuring careless responding developed by Meade and Craig (2012; Diligence, Interest, Effort, and Attention) ranged from –.10 to .09, suggesting that the effects of graph literacy are unlikely to be confounded with these factors. The item Use Me from this scale was not included in analyses, as all participants provided the same response (i.e., “yes”). Of note, the time that participants spent viewing the interpretation questions for the graphs did not differ reliably as a function of graph literacy (ps > .5), also suggesting that careless responding is unlikely to constitute a confounding factor for the effect of graph literacy. When numeracy, knowledge that graphs can be misleading, and careless responding were included as covariates in the analyses reported below, results remained unchanged in all cases.

Does Graph Literacy Affect Interpretations of Graphs With and Without Conflicts? Table 3 shows the percentage of respondents who gave correct responses to the graphs, as a function of graph literacy. First, we sought to examine differences in accuracy of interpretations of graphs with and without conflicts, for participants with high and low graph literacy. To this end, we conducted a 2 × 2 ANOVA with graph literacy as between-subjects factor and the presence of conflict as within-subject factor, on the average percentage of correct responses. This analysis revealed a main effect of graph literacy, $F(1,49) = 11.22$, $p = .002$, $\eta_p^2 = .19$, a main effect of the presence of conflict, $F(1,49) = 256.25$, $p = .001$, $\eta_p^2 = .84$, and an interaction between graph literacy and presence of conflict, $F(1,49) = 4.23$, $p = .045$, $\eta_p^2 = .08$. As can be seen in Table 3, graphs with conflicts had significantly lower rates of correct responses, as compared to their equivalent versions without conflicts. Overall, the percentage of correct responses was higher for participants with high graph literacy. Additionally, Bonferroni-corrected pairwise comparisons revealed that the difference in accuracy between participants with high and low graph literacy was larger for graphs with conflicts, $p = .008$, $d = .80$, than for graphs without conflicts, $p = .034$, $d = .59$, supporting H1a.

Table 3. Percentage of Respondents Who Gave Correct Responses to the Graphs in Experiment 2, as a Function of Whether They Contained Conflicts or Not, and Graph Literacy

Item	Low graph literacy		High graph literacy	
	Nonconflict	Conflict	Nonconflict	Conflict
y-axis-scale conflict				
G1. Inverted scale, *x* axis below	100%	27%	100%	45%
G2. Scale with negative values	86%	64%	97%	83%
G3. Excised scale	100%	5%	100%	21%
G4. Scale not indicating zero baseline	95%	14%	100%	45%
G5. Logarithmic scale	100%	0%	100%	28%
G6. Inverted scale, *x* axis above	100%	50%	100%	52%
x-axis-scale conflict				
G7. Inverted scale (bar)	91%	27%	97%	38%
G8. Inverted scale (line)	100%	18%	97%	38%
G9. Values not placed at proportional distances (bar)	95%	14%	97%	48%
G10. Values not placed at proportional distances (line)	100%	9%	100%	34%
Textual conflict				
G11. People without disease	100%	45%	100%	55%
G12. Change in the percentage of people with different types of cancer	82%	32%	100%	59%
G13. Number of patients per doctor	100%	50%	100%	59%
G14. People testing negative	95%	45%	100%	66%
G15. Number of patients per nurse	91%	36%	93%	48%
G16. People who died after virus exposure	73%	68%	90%	76%
Average overall	94%	32%	98%	50%

Next, we examined the different types of responses provided for graphs with conflicts. The average proportion of correct responses to the questions across graphs was 42% (*SE* = 4.5), while the average proportion of incorrect responses corresponding to direct spatial-to-conceptual mappings (*mapping responses*) was 55% (*SE* = 4.7). As in Experiment 1, the average proportion of incorrect responses that were not related to the mapping was low (4%; *SE* = 0.9), indicating that the majority of participants who misinterpreted the graphs did so on the basis of direct spatial-to-conceptual mappings. Therefore, in the following analyses we focus on the percentage of mapping responses.

The average percentage of mapping responses among participants with low graph literacy was 70% (*SE* = 4.7) for graphs with *y*-axis-scale conflict, 48% (*SE* = 7.0) for graphs with textual conflict, and 80% (*SE* = 5.2) for graphs with *x*-axis-scale conflict. In contrast, participants with high graph literacy showed on average 55% (*SE* = 5.4) mapping responses for graphs with *y*-axis-scale conflict, 35% (*SE* = 5.8) for graphs with textual conflict, and 57% (*SE* = 6.8) for graphs with *x*-axis-scale conflict. A 2 × 3 ANOVA with graph literacy as between-subjects factor and type of conflict as within-subject factor on the average percentage of mapping responses revealed a main effect of graph literacy, $F(1,49) = 6.35$, $p = .015$, $\eta_p^2 = .12$, supporting H1b. This analysis also revealed a main effect of type of conflict, $F(2,98) = 20.74$, $p = .001$, $\eta_p^2 = .30$, indicating that the percentage of mapping responses was significantly lower for textual-conflict graphs, as compared to *y*-axis-conflict ($p = .001$) and *x*-axis-conflict ($p = .001$) graphs. All other effects were not reliable ($F < 1$, $p > .4$). These results remained unchanged when MMK was also included as a factor in the ANOVA. Additionally, a main effect of MMK was observed, $F(1,48) = 4.00$, $p = .051$, $\eta_p^2 = .08$, indicating that mapping responses were more likely among participants with low content knowledge than for those with high content knowledge. Specifically, for participants with low MMK the percentage of mapping responses was 60% (*SE* = 4.5), while it was 49% (*SE* = 5.2) for participants with high MMK.

Does Graph Literacy Affect the Viewing Time of Relevant Conventional Features? As in Experiment 1 distributions were skewed right, and therefore viewing times were log-transformed to reduce skew. Table 4 shows the mean viewing times for the different areas of interest for graphs with and without conflicts, as a function of the type of conflict and graph literacy.

First, to examine differences in viewing times for graphs with and without conflicts, we conducted a 2 × 3 ANOVA with graph literacy as between-subjects factor and presence of conflict as within-subject factor on the mean log-transformed time spent viewing the conventional features. This analysis revealed a main effect of the presence of conflict, $F(1,49) = 28.20$, $p = .001$, $\eta_p^2 = .37$, indicating that conventional features were overall viewed longer when graphs contained conflicts. The analysis also revealed a marginally significant interaction between graph literacy and presence of conflict, $F(1,49) = 3.64$, $p = .06$, $\eta_p^2 = .07$. Bonferroni-corrected pairwise comparisons revealed no reliable differences in viewing times as a function of graph literacy in graphs without conflicts ($M = 1.4$, $SE = 0.1$ for participants with high graph literacy vs. $M = 1.6$, $SE = 0.2$ for those with low graph literacy), supporting H2a. Although this was also the case for graphs with conflicts, for this kind of graph the pattern of results at the descriptive level was in the expected direction. That is, we found longer viewing times for participants with high graph literacy ($M = 3.4$ s, $SE = 0.3$) than for those with low graph literacy ($M = 2.9$ s, $SE = 0.3$). The latter result provides some support for H2b. Viewing times for graphs with conflicts are explored in more detail in the next section. As in Experiment 1, the total time that participants spent viewing the graphs did not vary as a function of graph literacy ($F < 3$, ps > .1). However, as can be seen in Table 4, participants with high graph literacy spent significantly longer overall viewing graphs with conflicts ($M = 11.1$, $SE = 0.9$) than graphs without conflicts ($M = 8.6$, $SE = 0.5$; $p = .001$).

Next, we performed analyses only for graphs with conflicts, controlling for baseline individual variability in viewing times. To do so, for each conflict type we deducted the mean time spent viewing the relevant conventional features in graphs without conflicts from the mean time spent viewing the relevant conventional features in graphs with conflicts and divided the resulting value by the mean viewing time for graphs without conflicts. A 2 × 3 ANOVA with graph literacy as between-subjects factor and type of conflict as within-subject factor on the baseline-adjusted times spent viewing the relevant conventional features revealed a main effect of graph literacy, $F(1,49) = 8.19$, $p = .006$, $\eta_p^2 = .14$. Viewing times overall were higher for individuals with high graph literacy, providing additional support for H2b. This analysis also yielded a main effect of type of conflict, $F(2,98) = 7.24$, $p = .001$, $\eta_p^2 = .13$, reflecting that baseline-adjusted viewing times were higher for graphs with *y*-axis-scale conflict than for those with *x*-axis ($p = .010$) and textual ($p = .001$) conflict. All results remained unchanged when MMK was also included as a factor in the ANOVA. This analysis did not reveal a main effect of MMK or an interaction involving MMK, implying that this factor did not affect times spent viewing the relevant conventional features.

Table 4. Raw and Log-Transformed Mean Times Spent Viewing the Relevant Areas of the Graphs and Total Viewing Times in Experiment 2, as a Function of Whether They Contained Conflicts or Not, Type of Conflict, and Graph Literacy (SEM in Parentheses)

	Graphs with y-axis-scale conflict				Graphs with x-axis-scale conflict				Graphs with textual conflict			
	Low graph literacy		High graph literacy		Low graph literacy		High graph literacy		Low graph literacy		High graph literacy	
Area of graph	Nonconflict	Conflict	Nonconflict	Conflict	Nonconflict	Conflict	Nonconflict	Conflict	Nonconflict	Conflict	Nonconflict	Conflict
y-axis scale												
Time	**1.01**	**2.18**	**0.55**	**2.53**	1.34	1.02	1.06	1.39	1.01	1.05	0.64	1.36
	(0.20)	**(0.40)**	**(0.07)**	**(0.40)**	(0.22)	(0.20)	(0.18)	(0.29)	(0.14)	(0.15)	(0.07)	(0.20)
Log time	**-0.74**	**-0.34**	**-1.20**	**-0.17**	-0.48	-0.76	-0.61	-0.77	-0.55	-0.53	-0.84	-0.47
	(0.14)	**(0.23)**	**(0.17)**	**(0.24)**	(0.19)	(0.26)	(0.19)	(0.22)	(0.15)	(0.17)	(0.12)	(0.18)
x-axis scale												
Time	1.41	1.55	1.27	1.67	**0.67**	**1.14**	**0.79**	**2.26**	1.35	1.33	1.36	1.40
	(0.13)	(0.13)	(0.07)	(0.12)	**(0.07)**	**(0.39)**	**(0.10)**	**(0.40)**	(0.13)	(0.13)	(0.10)	(0.11)
Log time	0.10	0.12	0.01	0.23	**-0.96**	**-1.05**	**-0.80**	**-0.53**	0.03	0.00	-0.01	0.10
	(0.10)	(0.10)	(0.07)	(0.08)	**(0.13)**	**(0.28)**	**(0.13)**	**(0.27)**	(0.11)	(0.11)	(0.08)	(0.08)
Title & *y*-axis label												
Time	2.10	2.33	2.13	2.59	2.26	1.44	2.30	1.54	**2.94**	**4.92**	**2.71**	**5.00**
	(0.23)	(0.20)	(0.18)	(0.27)	(0.28)	(0.16)	(0.26)	(0.22)	**(0.35)**	**(0.65)**	**(0.23)**	**(0.53)**
Log time	0.28	0.40	0.38	0.55	0.10	-0.31	0.26	-0.09	**0.65**	**1.12**	**0.75**	**1.25**
	(0.13)	(0.14)	(0.11)	(0.13)	(0.15)	(0.20)	(0.15)	(0.16)	**(0.18)**	**(0.19)**	**(0.09)**	**(0.13)**
Graph total												
Time	9.46	10.14	8.16	11.12	12.66	9.51	11.18	11.10	8.65	10.72	7.38	11.15
	(0.90)	(0.85)	(0.48)	(1.01)	(0.94)	(1.09)	(0.83)	(1.37)	(0.78)	(1.08)	(0.50)	(1.03)
Log time	2.07	2.41	1.99	2.41	2.30	2.01	2.25	2.10	1.97	2.13	1.86	2.20
	(0.09)	(0.09)	(0.06)	(0.08)	(0.07)	(0.11)	(0.07)	(0.12)	(0.10)	(0.11)	(0.06)	(0.10)

Note: Relevant conventional features for each type of conflict are marked in bold.

Is the Effect of Graph Literacy on Accuracy Mediated by Differences in the Viewing Time of Conventional Features? Individual correlation coefficients for mapping responses in all items with graph literacy ranged from –.02 to –.38. As our question concerned whether viewing time mediated the effect of graph literacy on accuracy of comprehension, for the mediational analyses we selected items with a correlation of at least .2 between graph literacy and mapping responses. This resulted in a total of three graphs per type of conflict: Graphs G3, G4, and G5 for *y*-axis-conflict graphs; Graphs G8, G9 and G10 for *x*-axis-conflict graphs, and Graphs G11, G12, and G14 for textual-conflict graphs. The average correlation with graph literacy for the remaining items was –.076. Thus, for the analyses that will be presented below, mean times spent viewing the relevant conventional features were calculated for the three selected graphs for each type of conflict. We controlled for baseline individual variability in viewing times following the same method outlined above.

In a regression equation, as expected, graph literacy significantly predicted the percentage of mapping responses for *y*-axis-conflict graphs, $\beta = -.34$, $t = -2.51$, $p = .015$, and *x*-axis-conflict graphs, $\beta = -.42$, $t = -3.25$, $p = .002$, and marginally for textual-conflict graphs, $\beta = -.26$, $t = -1.9$, $p = .06$, with higher graph literacy associated with fewer incorrect responses. Furthermore, graph literacy significantly predicted the baseline-adjusted viewing times on the relevant conventional features for *y*-axis-conflict graphs, $\beta = .30$, $t = 2.2$, $p = .034$, and *x*-axis-conflict graphs, $\beta = .31$, $t = 2.3$, $p = .027$ (see Figure 1b), but not for textual-conflict graphs, $\beta = -.12$, $t = -.85$, $p = .39$.

When baseline-adjusted times spent viewing the conventional features were included in the regression equations, this factor predicted the percentage of mapping responses for *y*-axis-conflict graphs, $\beta = -.28$, $t = -2.03$, $p = .047$, and *x*-axis-conflict graphs, $\beta = -.56$, $t = -5.06$, $p = .001$, whereas the direct effect of graph literacy on mapping responses was reduced both for *y*-axis-scale conflict graphs, $\beta = -.26$, $t = -1.87$, $p = .068$, and *x*-axis-conflict graphs, $\beta = -.25$, $t = -2.22$, $p = .031$. In contrast, for textual-conflict graphs when the baseline-adjusted times spent viewing the conventional features were included in the regression equation, this factor did not predict the percentage of mapping responses, $\beta = -.06$, $t = -.44$, $p = .659$, while the direct effect of graph literacy on mapping responses remained marginally significant, $\beta = -.26$, $t = -1.9$, $p = .065$.

To test the mediation effect for graphs with *y*-axis and *x*-axis-scale conflicts, we performed two bootstrap analyses with 1,000 resamples using the PROCESS procedure developed by Hayes (2013). The analyses provided evidence of a significant indirect effect of baseline-adjusted times spent viewing *y*-axis and *x*-axis scales, as the mean estimate was −5.36, (within the 95% confidence interval [−15.83, −1.30]) for the former and −12.25 (within the 95% confidence interval [−20.80, −4.09]) for the latter. This supports H3a, while the lack of relationship between viewing times, incorrect responses, and graph literacy for graphs with textual conflicts supports H3b. As in Experiment 1, we also performed additional analyses examining the different types of transitions between different global regions. Results can be found in Appendix B.

Discussion

In Experiment 2 we replicated and extended the findings of Experiment 1. Using a more diverse set of graphs containing different types of conflicts, we found that people with lower graph literacy often relied on spatial-to-conceptual mappings in their interpretations. In contrast, people with higher graph literacy were more likely to use information from conventional features to override the mappings leading to erroneous conclusions.

Eye-tracking recordings showed that participants with higher graph literacy spent more time viewing conventional features in graphs involving scale conflicts. Longer times spent viewing scales, in turn, were associated with fewer errors in interpretations. In contrast, time spent viewing relevant conventional features in graphs involving textual conflicts did not predict accuracy of understanding (in line with Experiment 1) and was not related to graph literacy. A possible explanation for these differences between graphs involving scale and textual conflicts is that longer viewing times of textual elements might not necessarily lead to a correct interpretation, as this may require not only attending to these elements but also making appropriate inferences. For instance, in two graphs with textual conflicts, the questions asked about a certain ratio (e.g., doctors per patient) but the graphs showed the inverse of that ratio (e.g., patients per doctor). The correct interpretation of these graphs might require a stronger engagement in mental operations that are not necessarily reflected in longer viewing times, as well as the integration of information in conventional features with other types of conceptual understanding and literacy not measured here (such as the broader concepts of prose and document literacy, e.g., Kutner et al., 2006). In contrast, attending to numerical scales might be sufficient to avoid misinterpretations in graphs with scale conflicts, as such conflicts originate from relatively simple manipulations of numerical values (e.g., inverting their order of changing the intervals between them).

Experiment 2 also included graphs that did not contain any conflicts. These graphs were designed to be equivalent in every way to the graphs involving conflicts (same graph type, type of question, number of data points and variables) except that spatial-to-conceptual mappings corresponded to the information conveyed by conventional features. For those graphs, we found high rates of correct answers for participants with both high and low graph literacy. In line with our predictions, graph literacy did not affect times spent viewing conventional features in graphs without conflicts. This can be accounted for by the fact that for such graphs, attending to these features is not crucial to reach a correct interpretation. Independently of graph literacy, we found that content knowledge about medical domain improved understanding of the graphs. This finding supports the idea that knowledge of the content being depicted can also aid in graph interpretations (Freedman & Shah, 2002; Shah & Freedman, 2011; Shah et al., 2005; see also Novick, 2006). Because the graphs in this study involved information related to health and medicine, acquaintance with this area might have helped participants to more easily understand the concepts included in conventional features (e.g., prevalence of diseases, effects of cancer treatments, hormones, and vitamins), particularly for graphs with textual conflicts. Content knowledge did not affect viewing times, suggesting that it does not affect allocation of attention to specific graph elements.

Finally, in Experiment 2 we excluded a number of possible confounds of the relationship of graph literacy and graph processing. Participants with low graph literacy were not merely more careless, as suggested by the lack of differences linked to graph literacy in items measuring participant engagement (Meade & Craig, 2012) and by the fact that no reliable differences existed in the overall time spent exploring graphs or viewing the questions assessing interpretations. Additionally, participants with low graph literacy knew as well as those with high graph literacy that graphs can be plotted misleadingly.

General Discussion

In two experiments, we found that lower graph literacy was associated with a stronger tendency to rely on spatial features of graphs (such as slope of a line or height of bars) rather than on conventional features (such as axis labels and scales). When information conveyed by spatial features was incongruent with that included in conventional features, participants with lower graph literacy misinterpreted the graphs more frequently than those with higher graph literacy. Analyses of participants' eye movements revealed that lower graph literacy was associated with less time spent viewing the conventional features containing essential information for detecting conflicts. These differences in viewing times mediated the link between graph literacy and interpretations of graphs that involved a conflict between spatial features and information conveyed in numerical scales on *x* or *y* axes.

Theoretical Implications

The present findings expand previous research on perceptual and cognitive processes in graph comprehension (Carpenter & Shah, 1998; Kosslyn, 1989; Lohse, 1993; Pinker, 1990; Simkin & Hastie, 1987), documenting the existence of differences in these processes that are linked to individual differences in graph literacy. Eye-fixation patterns in our study suggest that higher graph literacy is associated with a larger tendency to direct attention to conventional features containing essential information. This finding is in accord with the information reduction framework proposed by Haider and Frensch (1996, 1999), which suggests that skilled individuals are more able to recognize and focus on task-relevant information.

Our finding that graph literacy did not affect time spent viewing conventional features in graphs without conflicts is also in line with the information reduction framework, because in such graphs, attending to conventional features is not crucial to reach a correct interpretation. That is, our findings indicate that graph literacy is associated with strategic differences in allocation of attention and encoding. In Experiment 2 we also found that people with high graph literacy spent more time overall viewing graphs with conflicts than graphs without conflicts.[16] This finding also supports the notion that highly skilled individuals adjust their encoding strategies more adaptively, as compared to less skilled individuals (Cokely & Kelley, 2009; Cokely et al., 2006). As noted by Cokely and Kelley (2009), highly skilled individuals might rely more on elaborative search when this is necessary for comprehension (here, when graphs contain conflicts) but less when this can be advantageous (when graphs do not contain conflicts, and therefore a thorough search is not necessary).

Our results are also in line with studies documenting a widespread tendency to interpret graphs on the basis of non-arbitrary spatial-to-conceptual mappings that emerge

[16] Unreported analyses on the mean duration of fixations on conventional features also revealed that for individuals with high graph literacy, the mean duration was longer when graphs contained conflicts, while the presence of conflicts did not affect the mean duration for individuals with low graph literacy. Additionally, no reliable differences were observed in the mean duration of fixations between individuals with low and high graph literacy for graphs with or without conflicts. This result contrasts with studies that have revealed that experts can have shorter fixation durations in processing visual displays, owing to extended capacities that allow for more rapid encoding of information (for a review, see Gegenfurtner, Lehtinen, & Säljö, 2011).

consistently in adults and children with no graphing experience (e.g., “higher equals more”; Gattis, 2002, 2004). When spatial features are incongruent with the relationships that the graph is supposed to show (e.g., when higher bars do not indicate larger quantity), relying on such mappings can lead to errors in interpretation. Crucially, our findings demonstrate that less graph literate individuals show a bias toward basing their interpretations of graphs primarily on such translations. Such mappings can be grounded on associations acquired through experience with the environment (Tversky, 2001, 2009), as well as on other types of general cognitive constraints or sources of similarity (Gattis, 2002, 2004). Future research could seek to determine how graph literacy affects overreliance on different kinds of mappings when they come into conflict (e.g., “steeper equals faster” vs. “higher equals more”; Gattis & Holyoak, 1996).

Taken together, our findings provide converging evidence that graph literacy can exert a top-down influence on the interpretations that viewers give to graphs, in conjunction with other types of prior knowledge, such as specific content knowledge (Freedman & Shah, 2002; Shah & Freedman, 2011; Shah & Hoeffner, 2002). The examination of the constituting features of graph schemas in viewers with low and high graph literacy was beyond the scope of the current investigation and methods. However, it is plausible that such mental representations may have contributed to direct allocation of attention to the relevant conventional features (Maichle, 1994; Pinker, 1990).

Implications for Graph Design and Visual Communication of Medical Information

Research indicates that graphs that are available to the public often include misleading characteristics similar to those manipulated in the present study, such as improperly scaled axes (Beattie & Jones, 2002; Cooper, Schriger, Wallace, Mikulich, & Wilkes, 2003) or longer bars representing lower values (Kosslyn, 2006), and that such manipulations can substantially affect viewers’ judgments and decisions (Arunachalam, Pei, & Steinbart, 2002; Pennington & Tuttle, 2009). In line with Okan et al.’s (2012b) findings, our results suggest that such graphs are likely to be misinterpreted by people lacking graph literacy skills and point to an important principle for designing graphs that are suitable even for people with low graph literacy: Spatial and conventional features should convey the same meaning. For some graphs, this could help less graph literate people reach the correct interpretation even without attending to the conventional features.

In addition, methods could be developed to direct attention to essential information in conventional features, to increase the likelihood that this information will be encoded and integrated. For instance, people could be presented with interactive displays that require using mouse clicks to uncover the different regions. Forcing people to uncover conventional features in a first step could help them identify referents of the concepts that will be depicted before they make direct spatial-to-conceptual mappings. The current results suggest that this method might be more effective for graphs containing essential information in scales. For graphs containing essential information in textual elements, specific training might be required at a conceptual level (e.g., understanding the difference between rates of change and event rates).

Open Questions for Future Research

The current work leaves a number of questions open for future research. First, both the materials and the graph literacy instrument focused on the medical domain. In some instances, embedding graphs in a particular context may hinder performance, as context information might activate different knowledge structures than graphs presented abstractly (Mevarech & Stern, 1977). Future research should include more diverse materials varying in complexity and involving different tasks (see, e.g., Ratwani, Trafton, & Boehm-Davis, 2008; Trickett & Trafton, 2006). Second, our participants were relatively well educated. It is possible that the differences in the graph comprehension processes outlined above will become more salient with a more diverse, less educated group of participants. Finally, future research should aim to achieve a more precise specification of how graph comprehension is affected by interactions between graph literacy and other individual differences, such as domain-general cognitive abilities (Stanovich & West, 2000, 2008), decision making-skills (Bruine de Bruin, Parker, & Fischhoff, 2007), working memory limitations (Huestegge & Philipp, 2011; Peebles & Cheng, 2001, 2003), and spatial abilities (Feeney, Adams, Webber, & Ewbank, 2004).

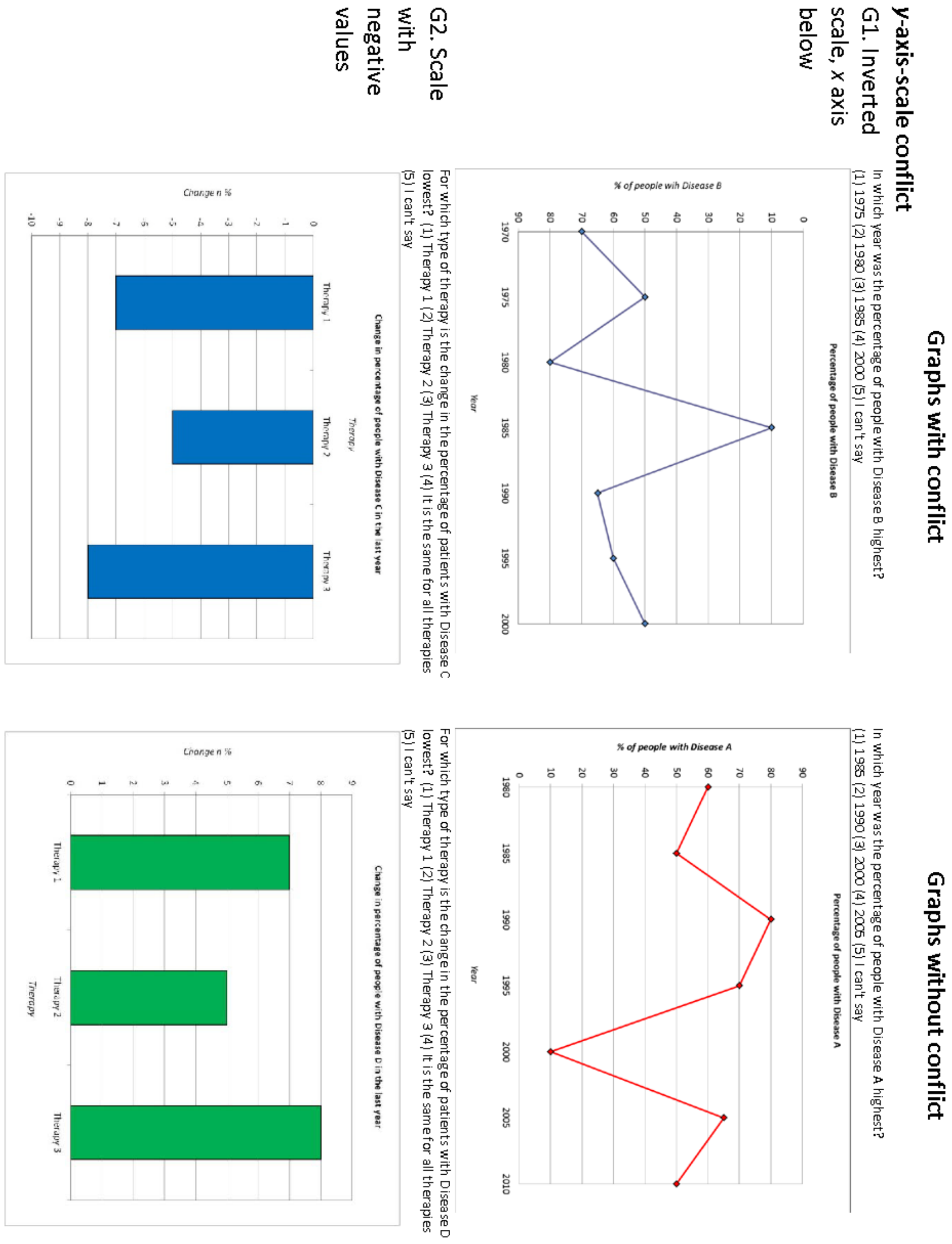
In sum, we have demonstrated that graph comprehension processes can be affected by individual differences in graph literacy. A lack of experience with arbitrary graphic conventions can limit the attention directed to some features that do not map onto viewers' experience with their physical environment. Less graph literate individuals can be less likely to engage in integration processes that involve incorporating essential information in such features and combining this information with that in other elements. In contrast, these individuals seem to rely to a larger extent on their real-world knowledge to interpret graphs and thus more often misinterpret data. A precise theoretical understanding of the nature and causes of our judgment biases allows the anticipation of potential errors and development of improved educational interventions. The current findings can play a central role in the development of custom-tailored decision support systems built to inoculate professionals, policy makers, and the general public against potentially distorted and misleading communication.

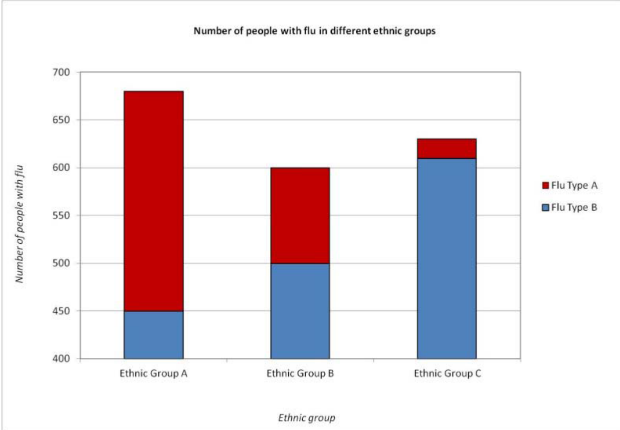
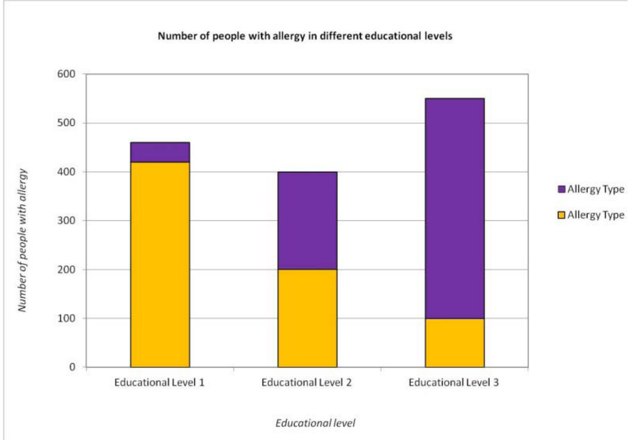
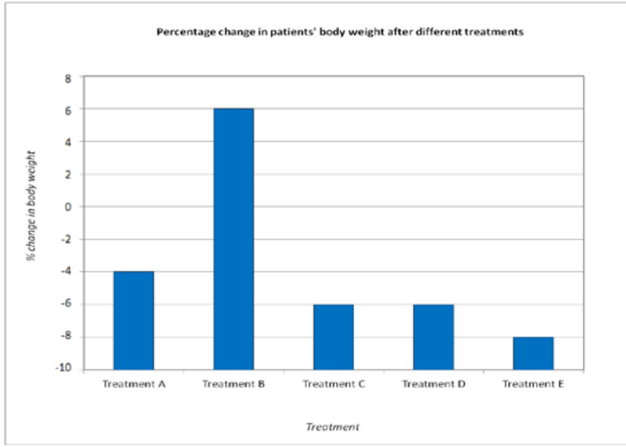
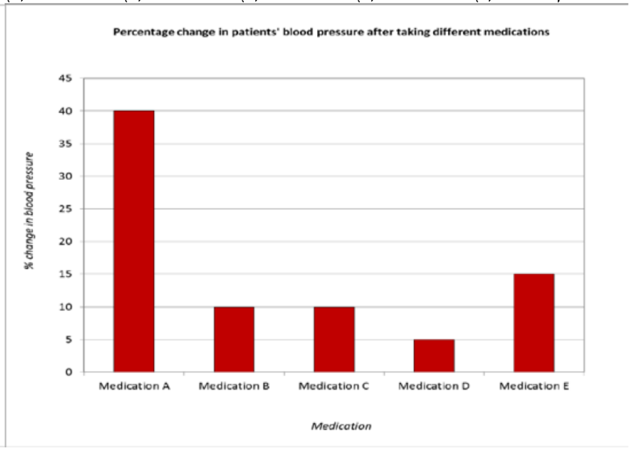
Acknowledgments

This study is part of grants awarded by the Informed Medical Decisions Foundation (United States) to Yasmina Okan (ref. # 0230-1), and to Rocio Garcia-Retamero by the Ministerio de Economía y Competitividad (Spain; ref. # PSI2011-22954), and was supported by the Max Planck Society. We would like to thank Edward T. Cokely and Andrés Catena for their valuable comments, as well as Priti Shah, Rebecca Rhodes, Merideth Gattis, and two other anonymous reviewers for their helpful suggestions. We are also grateful to Mona Merkert, Gregor Caregnato, and Christian Elsner for their help in conducting the study and to Anita Todd for editing the manuscript.

Appendix A

Graphs used in Experiments 1 and 2. *Note: Graphs G1, G2, G11, and G12 with conflicts were used in Experiment 1. All graphs were used in Experiment 2. In Experiment 1, response options for graphs G1 and G2 did not include "I can't say."*



	Graphs with conflict	Graphs without conflict
***y*-axis-scale conflict (continued)**		
G3. Excised scale	For which ethnic group was the proportion of people with flu Type A larger than the proportion of people with flu Type B? (1) Ethnic Group A (2) Ethnic Group B (3) Ethnic Group C (4) it was never larger (5) I can't say	For which educational level was the proportion of people with allergy Type A larger than the proportion of people with allergy Type B? (1) Educational Level 1 (2) Educational Level 2 (3) Educational Level 3 (4) it was never larger (5) I can't say
		
G4. Scale not indicating zero baseline	Which treatment resulted in the smallest change in patients' body weight? (1) Treatment A (2) Treatment B (3) Treatment C (4) Treatment E (5) I can't say	Which medication resulted in the smallest change in patients' blood pressure? (1) Medication A (2) Medication B (3) Medication C (4) Medication D (5) I can't say
		

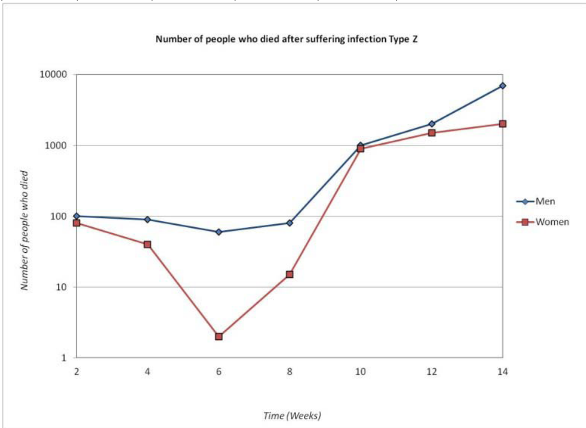
Graphs with conflict / Graphs without conflict

***y*-axis-scale conflict (continued)**

G5. Logarithmic scale

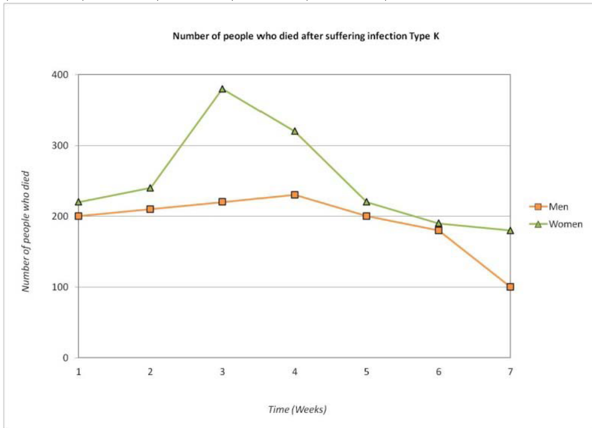
Graph with conflict:

When was there a larger difference between the number of men and women dying after suffering infection Type Z?
(1) Week 2 (2) Week 6 (3) Week 10 (4) Week 14 (5) I can't say



Graph without conflict:

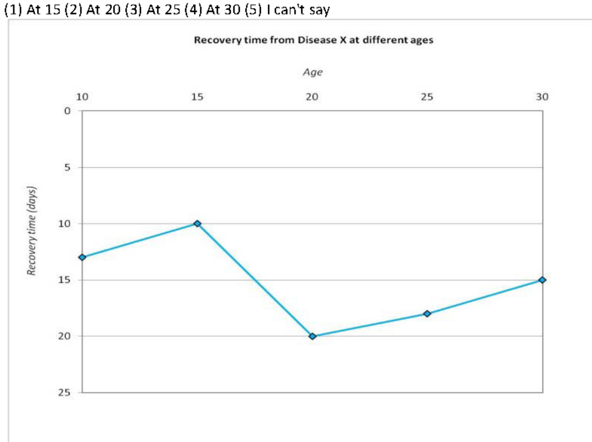
When was there a larger difference between the number of men and women dying after suffering infection Type K?
(1) Week 1 (2) Week 3 (3) Week 5 (4) Week 7 (5) I can't say



G6. Inverted scale, *x* axis above

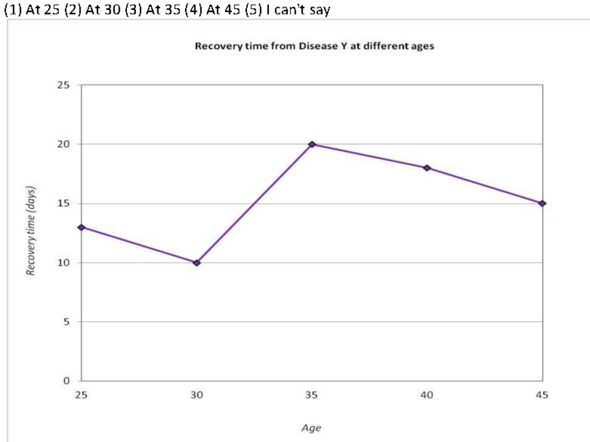
Graph with conflict:

At which age was the recovery time from Disease X lowest?
(1) At 15 (2) At 20 (3) At 25 (4) At 30 (5) I can't say



Graph without conflict:

At which age was the recovery time from Disease Y lowest?
(1) At 25 (2) At 30 (3) At 35 (4) At 45 (5) I can't say

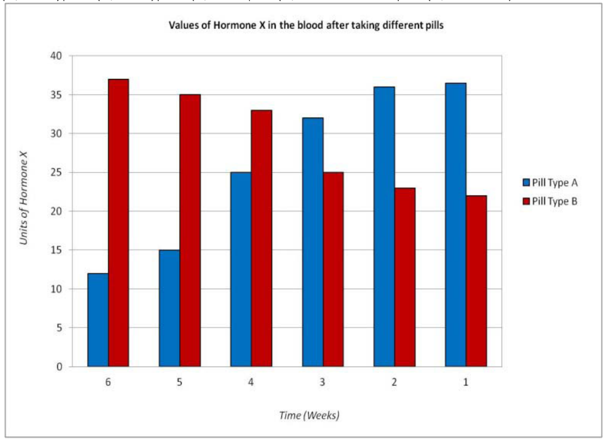


	Graphs with conflict	Graphs without conflict

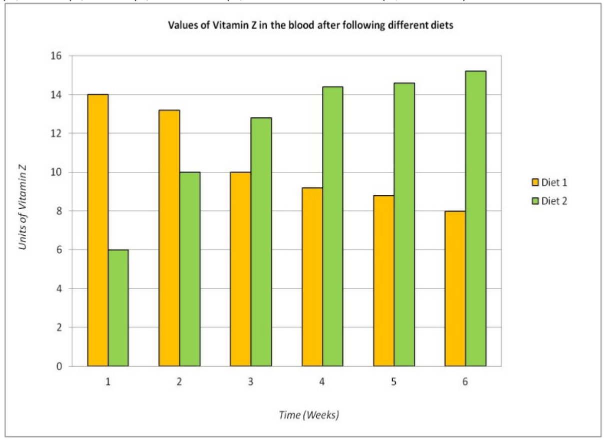
***x*-axis-scale conflict**

G7. Inverted scale (bar)

Which pill resulted in an increase in the values of Hormone X in the blood over time?
(1) Pill Type A (2) Pill Type B (3) Both pills (4) Neither of the pills (5) I can't say

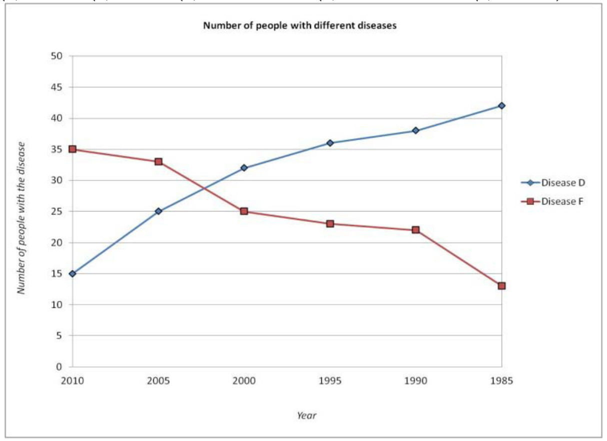


Which diet resulted in an increase in the values of Vitamin Z in the blood over time?
(1) Diet 1 (2) Diet 2 (3) Both diets (4) Neither of the diets (5) I can't say

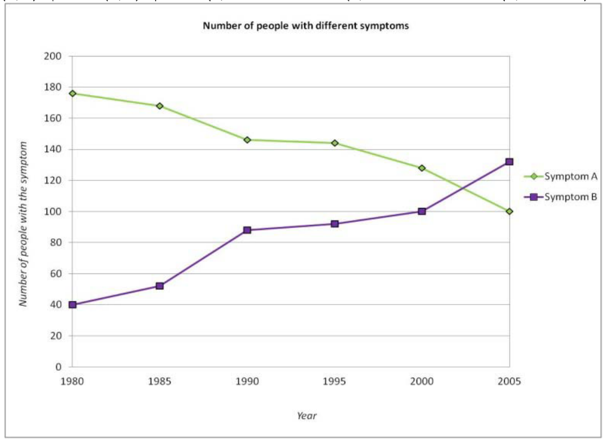


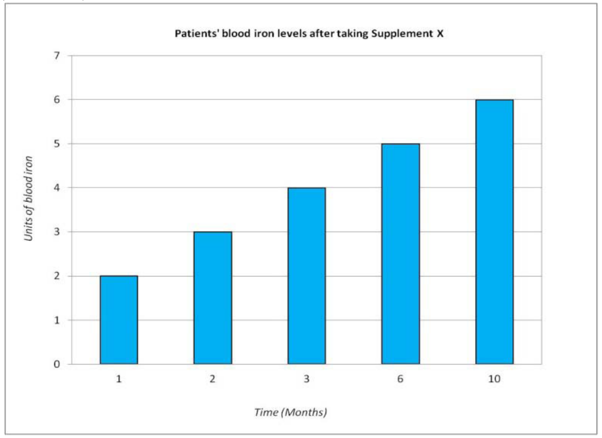
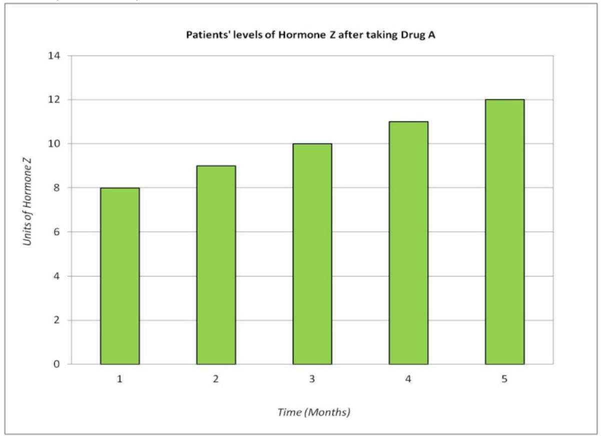
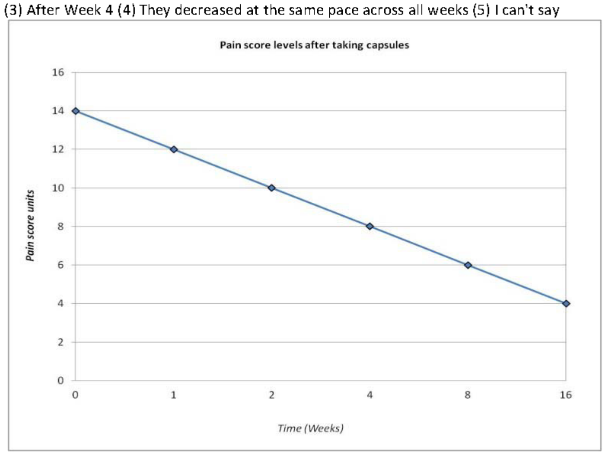
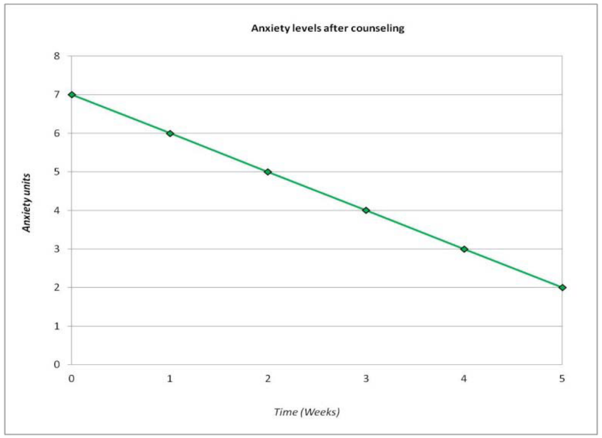
G8. Inverted scale (line)

For which disease did the number of affected people increase over time?
(1) Disease D (2) Disease F (3) For both of them (4) For neither of them (5) I can't say



For which symptom did the number of affected people increase over time?
(1) Symptom A (2) Symptom B (3) For both of them (4) For neither of them (5) I can't say



	Graphs with conflict	Graphs without conflict
***x*-axis-scale conflict (continued)**		
G9. Values not placed at proportional distances (bar)	When did patients' blood iron levels start to increase more slowly? (1) After Month 2 (2) After Month 3 (3) After Month 6 (4) They increased equally quickly across all months (5) I can't say 	When did patients' levels of hormone Z start to increase more slowly? (1) After Month 2 (2) After Month 3 (3) After Month 4 (4) They increased equally quickly across all months (5) I can't say 
G10. Values not placed at proportional distances (line)	When did pain scores start decreasing more slowly? (1) After Week 1 (2) After Week 2 (3) After Week 4 (4) They decreased at the same pace across all weeks (5) I can't say 	When did anxiety levels start decreasing more slowly? (1) After Week 1 (2) After Week 2 (3) After Week 4 (4) They decreased at the same pace across all weeks (5) I can't say 

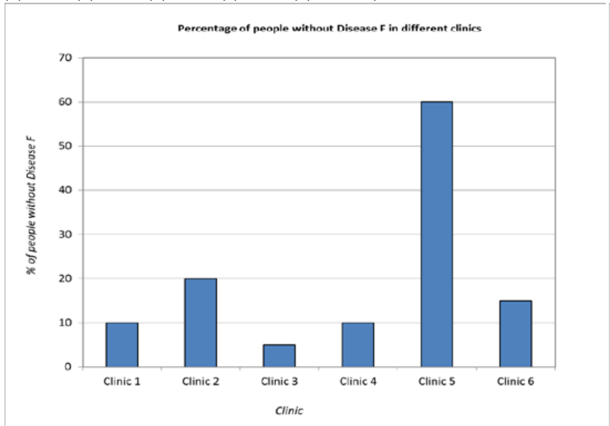
Graphs with conflict / Graphs without conflict

Textual conflict

G11. People without disease

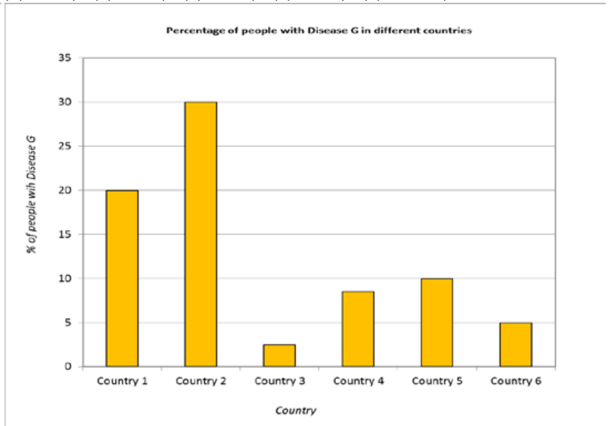
Graphs with conflict

In which clinic was the percentage of people with Disease F highest?
(1) Clinic 2 (2) Clinic 3 (3) Clinic 5 (4) Clinic 6 (5) I can't say



Graphs without conflict

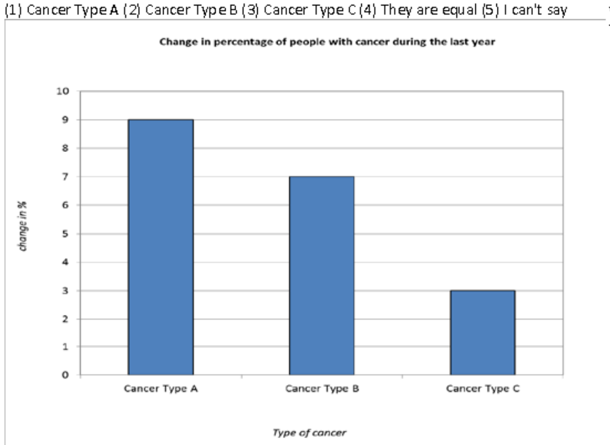
In which country was the percentage of people with Disease G highest?
(1) Country 1 (2) Country 2 (3) Country 3 (4) Country 5 (5) I can't say



G12. Change in the percentage of people with different types of cancer

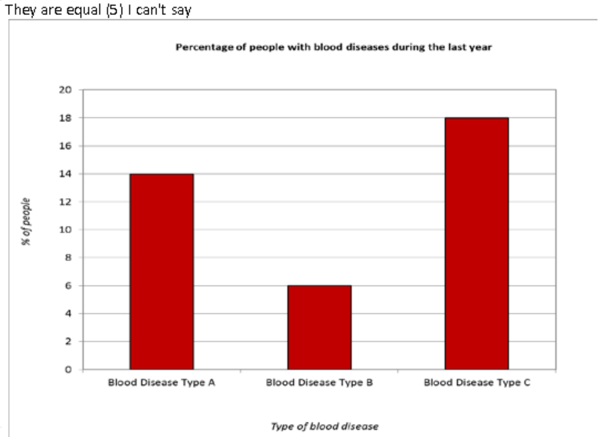
Graphs with conflict

Which type of cancer affected the smallest percentage of people during the last year?
(1) Cancer Type A (2) Cancer Type B (3) Cancer Type C (4) They are equal (5) I can't say



Graphs without conflict

Which type of blood disease affected the smallest percentage of people during the last year? (1) Blood disease Type A (2) Blood disease Type B (3) Blood disease Type C (4) They are equal (5) I can't say

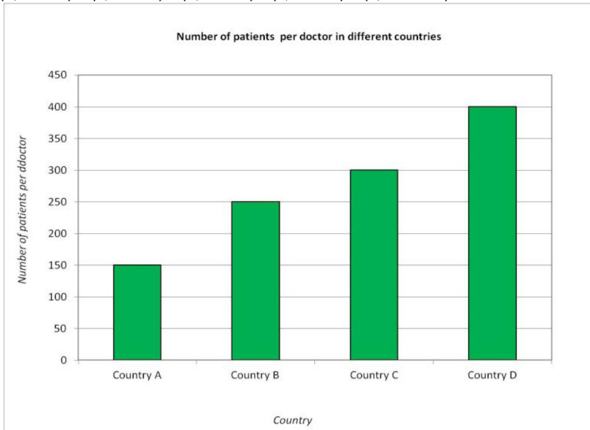


Graphs with conflict | **Graphs without conflict**

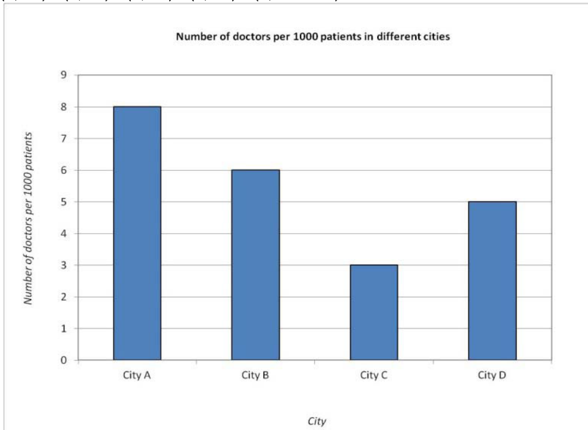
Textual conflict (continued)

G13. Number of patients per doctor

Which country has the highest number of doctors per patient?
(1) Country A (2) Country B (3) Country C (4) Country D (5) I can't say

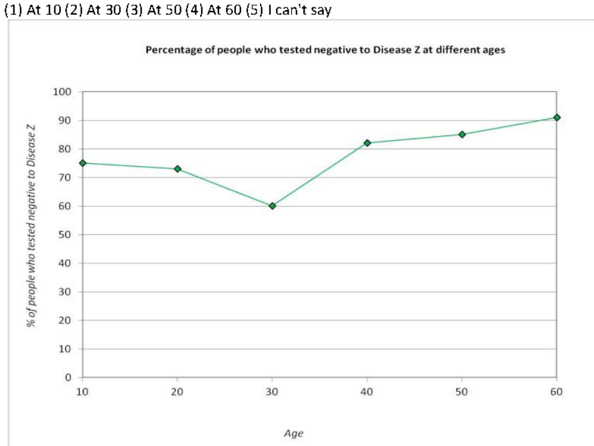


Which city has the highest number of doctors per patient?
(1) City A (2) City B (3) City C (4) City D (5) I can't say

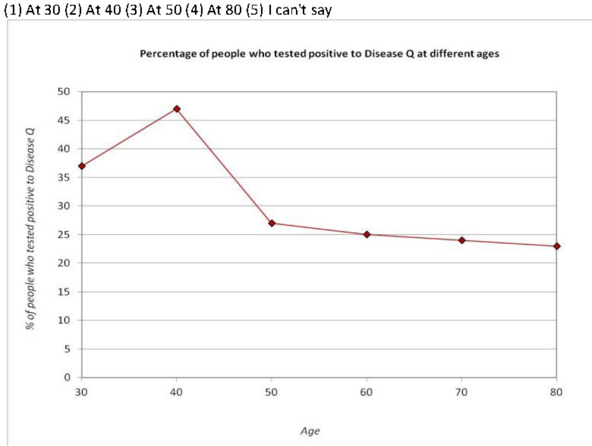


G14. People testing negative

At which age is the percentage of people diagnosed with Disease Z highest?
(1) At 10 (2) At 30 (3) At 50 (4) At 60 (5) I can't say



At which age is the percentage of people diagnosed with Disease Q highest?
(1) At 30 (2) At 40 (3) At 50 (4) At 80 (5) I can't say

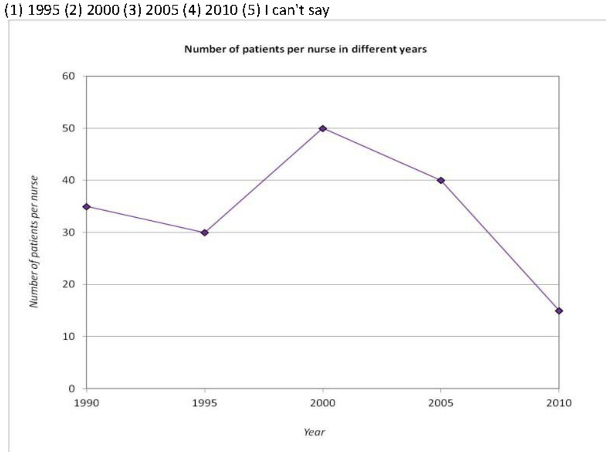


Graphs with conflict | **Graphs without conflict**

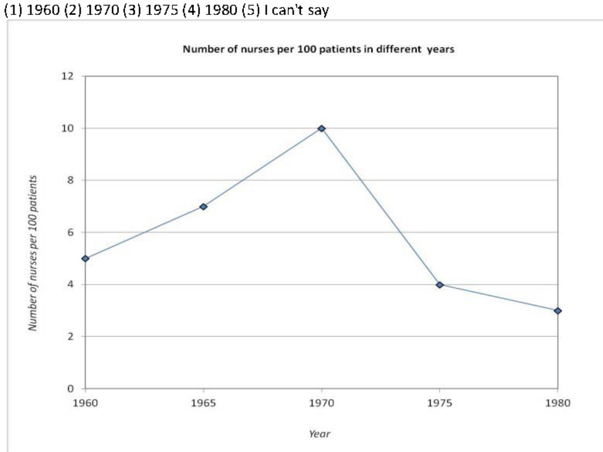
Textual conflict (continued)

G15. Number of patients per nurse

In which year was the number of nurses per patient lowest?
(1) 1995 (2) 2000 (3) 2005 (4) 2010 (5) I can't say

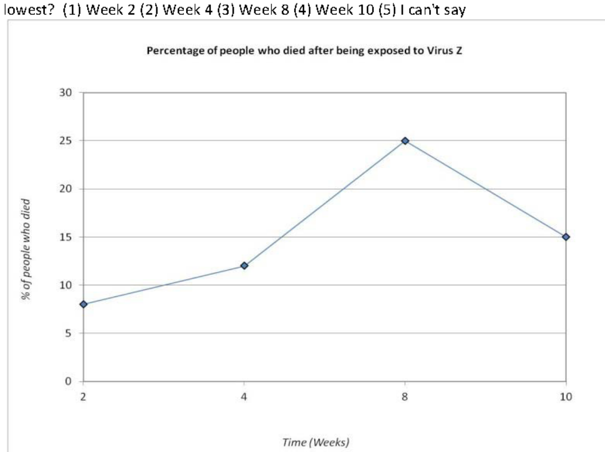


In which year was the number of nurses per patient lowest?
(1) 1960 (2) 1970 (3) 1975 (4) 1980 (5) I can't say

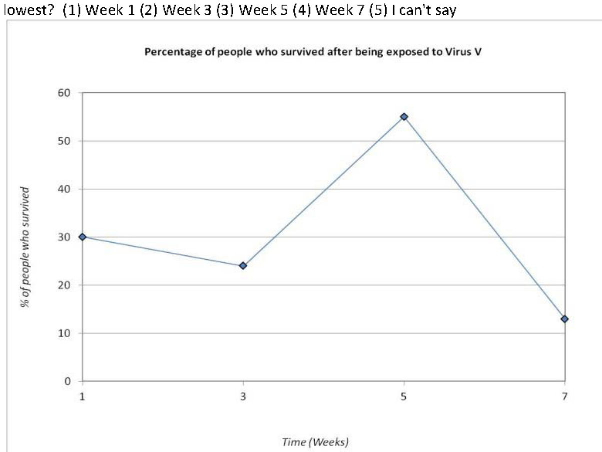


G16. People who died after virus exposure

When was the percentage of people who survived after being exposed to Virus Z lowest? (1) Week 2 (2) Week 4 (3) Week 8 (4) Week 10 (5) I can't say



When was the percentage of people who survived after being exposed to Virus V lowest? (1) Week 1 (2) Week 3 (3) Week 5 (4) Week 7 (5) I can't say



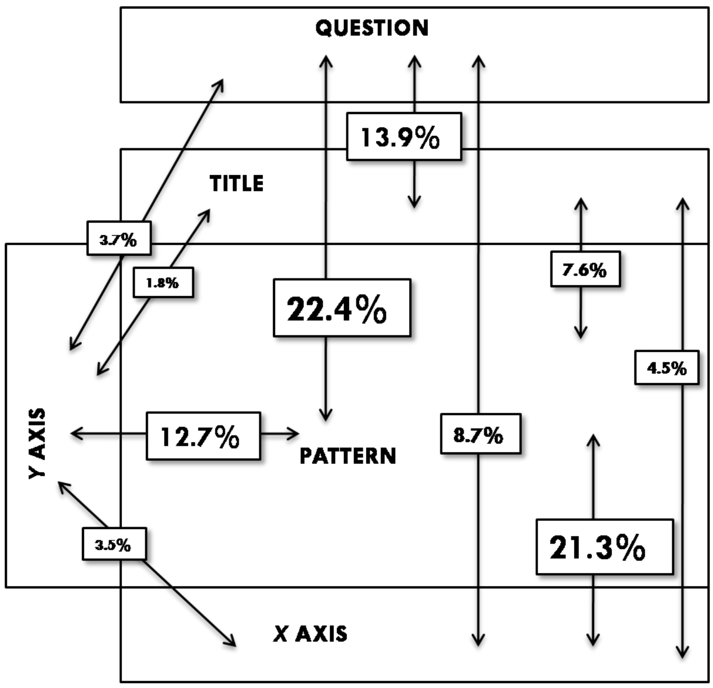
Appendix B

Further analyses: Transitions between global areas.

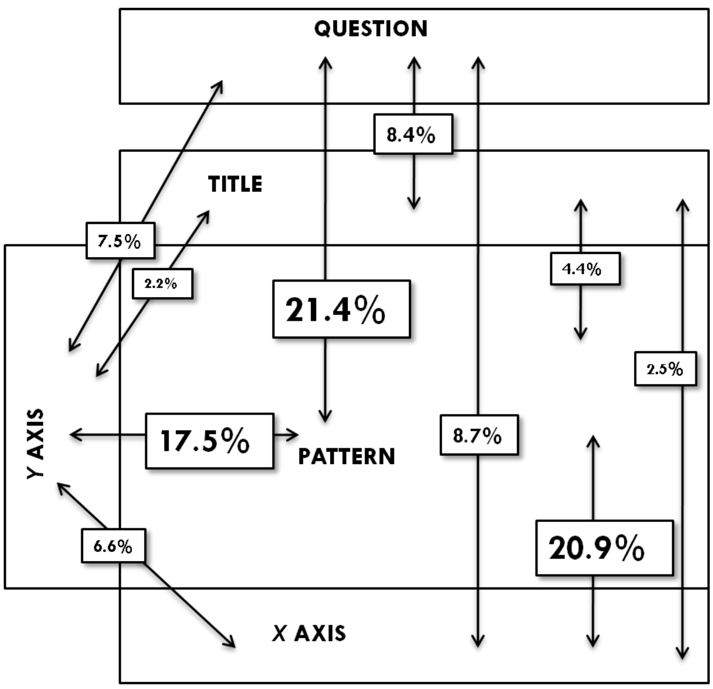
To broaden our exploration of the patterns of eye fixations in the current study, we further defined a set of areas of interest (AOIs) for all graphs that corresponded to the global elements of bar charts and line plots outlined in previous research (see, e.g., Carpenter & Shah, 1998; Kosslyn, 2006). Specifically, we divided the graphs into four global parts: the pattern, the *x* axis, the *y* axis, and the title. For this division, the *x* axis and *y* axis included the respective *x*-axis and *y*-axis values and labels. Following Carpenter and Shah (1998; see also Huestegge & Philipp, 2011), we then computed the number of transitions between these global areas. A transition was counted each time the participant broke a sequence of consecutive fixations on a given AOI to fixate on a different AOI. The question was also included as an AOI. Figure B1 shows the types of transitions made between the different global areas and how often each type occurred across graphs. In Experiment 1, the mean number of transitions across graphs was 18.3 (*SE* = 0.9), while in Experiment 2 it was 20.6 (*SE* = 1.1) for graphs with conflicts and 18.5 (*SE* = 0.8), for graphs without conflicts. In all cases, the most frequent types of transition across graphs were those between the pattern and the question, and between the pattern and the *x* axis (Figure B1). These results are in line with Carpenter and Shah's integrative model (1998), which predicts that a large proportion of transitions occur between the pattern and regions used to determine referents (e.g., *x* and *y* axes), as consequence of viewers' need to integrate information across these regions.

Figure B1. (a) The proportions of transitions made by participants between different global areas in Experiment 1; (b) The proportions of transitions made by participants between different global areas for graphs with conflicts in Experiment 2; (c) The proportions of transitions made by participants between different global areas for graphs without conflicts in Experiment 2.

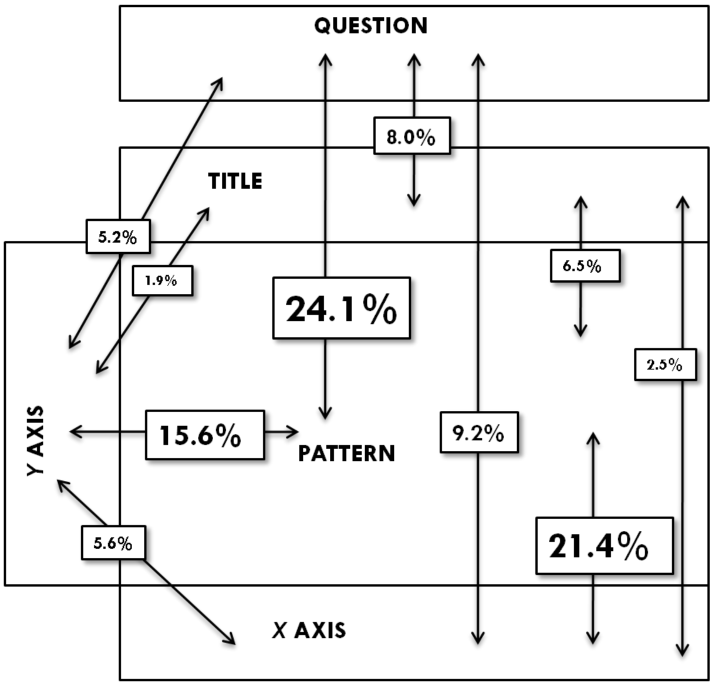
A



B



C


QUESTION
8.0%
TITLE
5.2%
1.9%
24.1%
6.5%
2.5%
Y AXIS
15.6%
PATTERN
9.2%
5.6%
21.4%
X AXIS

Chapter V

Appendix C

Items measuring knowledge that graphs can be misleading. *Note: The response options provided were "Yes" and "No." Yes responses were coded with 1 and No responses with 0, for a total possible score of 6.*

Thinking about graphs that you might have seen in different contexts, such as graphs presenting data for different financial, nutritional, or political options and trends …
Do you think they are sometimes designed in a way that…
Makes some options look better or worse than they really are (e.g., by making differences in the data presented look larger or smaller)?
Directs attention to a particular option or aspects of that option (e.g., by directing attention to specific values in the data)?
Makes trends look more positive or negative than they really are (e.g., by distorting or misrepresenting the trends in the data)?

Thinking about graphs that you might have seen presenting medical information, such as data for different treatments and screenings (e.g., results of medical trials and pharmaceutical advertisements)…
Do you think they are sometimes designed in a way that…
Makes some options look better or worse than they really are (e.g., by making differences in the data presented look larger or smaller)?
Directs attention to a particular option or aspects of that option (e.g., by directing attention to specific values in the data)?
Makes trends look more positive or negative than they really are (e.g., by distorting or misrepresenting the trends in the data)?

CHAPTER VI.

BIASING AND DEBIASING HEALTH DECISIONS WITH BAR GRAPHS: COSTS AND BENEFITS OF GRAPH LITERACY

Biasing and Debiasing Health Decisions with Bar Graphs:
Costs and Benefits of Graph Literacy

Abstract

Bar graphs are often recommended to improve risk communication in medicine and health. Unfortunately, when people view a bar graph depicting a mean they tend to believe that data points located within bars are more likely to be part of the underlying distribution than equidistant points outside bars. Here, we investigate potential consequences, key mechanisms, and generalizability of the within-the-bar bias in the medical domain. Results revealed a *within-the-bar bias* that led participants to prefer to modify their blood glucose levels, even when the information provided gave them no justifiable reason to do so. Interestingly, individuals with higher levels of graph literacy showed the largest biases. Graph literate individuals also benefited from the inclusion of bidirectional error bars in graphs, suggesting that debiasing efforts may be more beneficial for relatively high-skilled individuals. Theoretical mechanisms and prescriptive implications for graph design are discussed.

Submitted as:

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (submitted). Biasing and debiasing health decisions with bar graphs: Costs and benefits of graph literacy. *Quarterly Journal of Experimental Psychology*.

Introduction

Visual displays play an increasingly important role in modern societies, facilitating the communication of complicated information in medicine, economics, sport, weather, climate, and politics (Garcia-Retamero & Cokely, in press; Spiegelhalter, Pearson, & Short, 2011). Unfortunately, graphical communication can also cause judgment and decision making errors. For example, when people are shown a bar graph representing a mean and are asked to judge the likelihood that a data point is part of its underlying distribution, they often believe that the likelihood is larger for points located within the bars than for equidistant points located outside the bar. This tendency, called the "within-the-bar bias" (Newman & Scholl, 2012), is thought to occur because bars are unique visual objects defined by the closure of their boundaries, which originate from one particular axis. Consequently, people's attention is drawn to the region within the bar, such that it takes precedence over regions outside the bar.

Newman and Scholl (2012) demonstrated that the within-the-bar bias affects not only judgments concerning the likelihood of different data points, but also decisions made on the basis of bar graphs. They asked participants to imagine they were the CEO of a large car tire manufacturer, and presented them with information concerning the tensile strength of tires. Participants were told that the mean tensile strength of tested tires was zero, and that zero was the ideal value for safety. No objective reasons were provided to either increase or decrease the tensile strength of the tires. However, participants who viewed the value of zero represented in a graph where the bar originated from a lower x axis (i.e., situated below the mean) often preferred to increase the tensile strength. In contrast, those who viewed this value in a graph where the bar originated from an upper x axis (i.e., situated above the mean) often preferred to decrease the tensile strength.

Here, we report a study on the generalizability and mechanisms of the within-the-bar bias. Our aim in the present paper was threefold. First, we sought to investigate whether the within-the-bar bias extends to more common health and medical treatment decisions. Specifically, we examined the effect of this bias on people's preferences for treatments to alter blood glucose levels. If such preferences are affected by the within-the-bar bias, people who receive their blood test results in a graph with a bar originating from a lower x axis may seek to increase their blood glucose levels, even if the information provided gives them no reasons to do so. In contrast, those presented with a graph containing a bar descending from an upper x axis may prefer a treatment that decreases their blood glucose levels.

Second, we aimed to investigate the relations between the within-the-bar bias and graph literacy (Galesic & Garcia-Retamero, 2011b). Graph literacy refers to the ability to understand graphically presented information, and includes general knowledge about making inferences from different graphic formats (Freedman & Shah, 2002). Individuals with high graph literacy have been found to extract knowledge of a high level of complexity when viewing line graphs (Maichle, 1994) and to be more likely to make relevant inferences when viewing bar graphs depicting interactions (Shah & Freedman, 2011). Graph literacy is also associated with a larger likelihood to incorporate important information in titles of graphs, axes labels, and scales, and with a lower reliance on salient spatial features to interpret graphs (e.g., heights of bars; Okan, Garcia-Retamero, Galesic, & Cokely, 2012). Accordingly, graph literacy might moderate the within-the-bar bias.

Finally, we investigated the effectiveness of an intervention that could reduce the effect of the within-the-bar bias—i.e., error bars—that might help emphasize that values can come from both below and above the mean. Previous research suggests that the bias can persist even in graphs with error bars (Newman & Scholl, 2012). However, in previous experiments the error bars were not visible during the test phase. If graphs containing error bars are available at the time of judgment, as is common in medical decision making, graph literate viewers might be more prone to take into account the information conveyed by the error bars in their decisions.

Method

Participants

Participants were 458 undergraduate students from the University of Granada (307 female), aged 17–60 years (M = 21, SD = 4.7).

Materials & Procedure

The questionnaire was administered in the laboratory of the University of Granada, and all materials were implemented as an electronic survey in Unipark (www.unipark.de). Besides the experiment presented here, it included other unrelated tasks concerning medical risks, and took approximately 50 minutes to complete. The tasks relevant for the present study took between 15 and 20 minutes. All participants were presented with a hypothetical scenario in which they received their blood glucose levels from the previous week, with information structured in a manner similar to Newman and Scholl's vignettes (2012; see Appendix). Materials stated that a previous measurement of the participants' blood glucose (at the start of the week) had been ideal (120 mg/dL); however, since the start of the week, the last 30 blood tests performed showed that their blood glucose levels had varied between –20 and +20 in percentage change. Participants were then reminded that deviation from ideal levels could lead to a high risk of severe health consequences, and that blood glucose levels can vary throughout the day (e.g., dependent on one's last meal). Participants were then informed that their average percentage change throughout the week was zero.

Participants were randomly assigned into one of five experimental conditions. In the *numerical* (control) condition, participants were presented only with a text containing the numerical information (see Appendix). In the remaining conditions, participants were presented both with the numerical information in text and with a bar graph depicting this information, which appeared immediately below the text. Bar graphs were constructed following Newman and Scholl (2012). Specifically, in the *rising* condition, participants were presented with a bar graph showing the average percentage change where the bar was rising from a lower x axis (see Figure 1a). In the *falling* condition, participants saw the bar descending from an upper x axis (see Figure 1b). In the *rising with error bar* and *falling with error bar* conditions (Figures 1c and 1d, respectively), participants viewed the same graphs as in the first two conditions, with the exception that the graphs included bidirectional error bars. In all cases the y axis scale in the graph ranged from –20 to +20.

Participants were then instructed that, based on the information provided, they could choose to follow a treatment that would either slightly increase their blood glucose levels or take a treatment that would slightly decrease their blood glucose levels. They responded using a slider ranging from "slightly decrease my blood glucose levels" to "slightly increase my blood glucose levels", with a mid-point indicating "neither increase nor decrease my blood glucose levels." The numeric slider values ranged from –50 to 50. Following Newman and Scholl (2012), the participants did not see the numerical values. Time to read the scenario and to answer the decision question was unlimited.

Next, graph literacy was measured using the scale developed by Galesic and Garcia-Retamero (2011b). We also measured participants' numeracy (i.e., the ability to understand and manipulate different numerical expressions of probability; Lipkus, Samsa, & Rimer, 2001) using the Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012) and the 11 items included in the general and expanded numeracy scales developed by Lipkus et al. (2001). The experiment ended following basic demographic questions and debriefing.

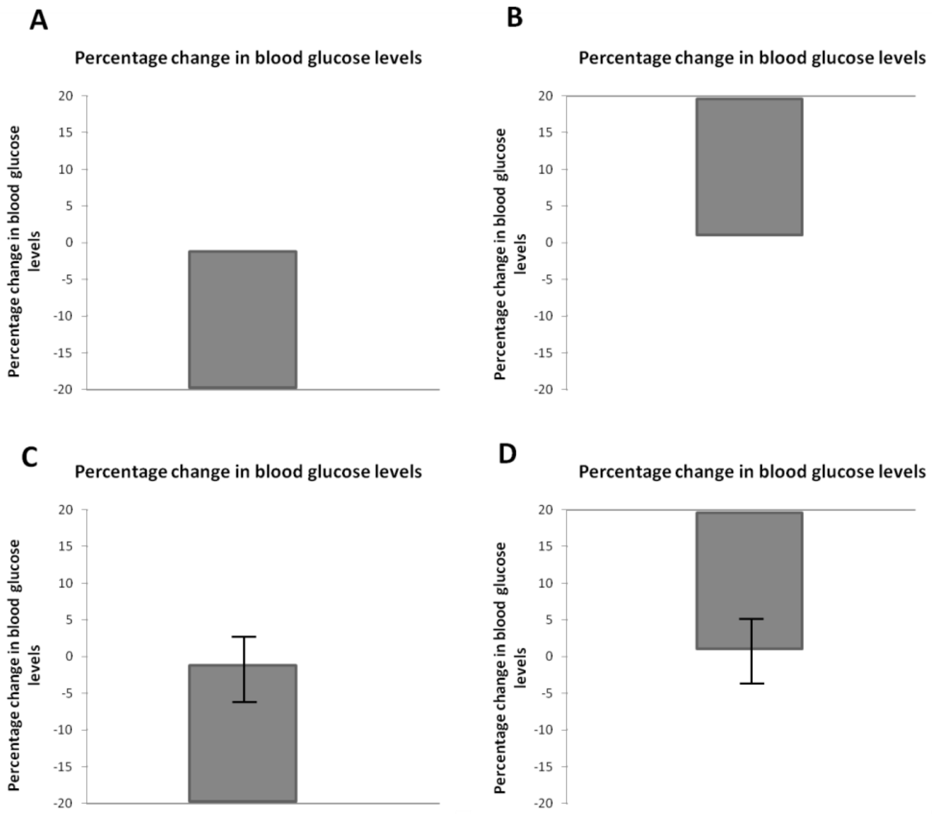


Figure 1. Graphs viewed by participants in the A. *rising* condition, B. *falling* condition, C. *rising with error bar* condition, and D. *falling with error bar* condition.

Results

First, we sought to examine whether the within-the-bar bias affected medical treatment decisions. Mean preference ratings were not normally distributed and showed homogeneous variances even after log-transformation. Therefore, we conducted a non-parametric Kruskall-Wallis test to compare ratings in all conditions. Mann-Whitney tests (one-tailed) were used for planned comparisons between the conditions. Results revealed that preference ratings varied significantly across conditions, $H(4)$, = 34.13, p = .001. These variations were in the anticipated directions, suggesting the presence of a within-the-bars bias. As can be seen in Figure 2, the presence of a rising bar led participants overall to show a preference to increase their glucose levels relative to the numerical condition, both for graphs without error bars (U = 3062.00, r = −.23, p = .001) and with error bars (U = 3235.00, r = −.23, p = .002). In contrast, the presence of a falling bar led participants overall to show a preference to decrease their levels relative to the numerical condition, for graphs without error bars (U = 3442.00, r = −.13, p = .045), and with error bars (U = 3906.50, r = −.14, p = .027).[17]

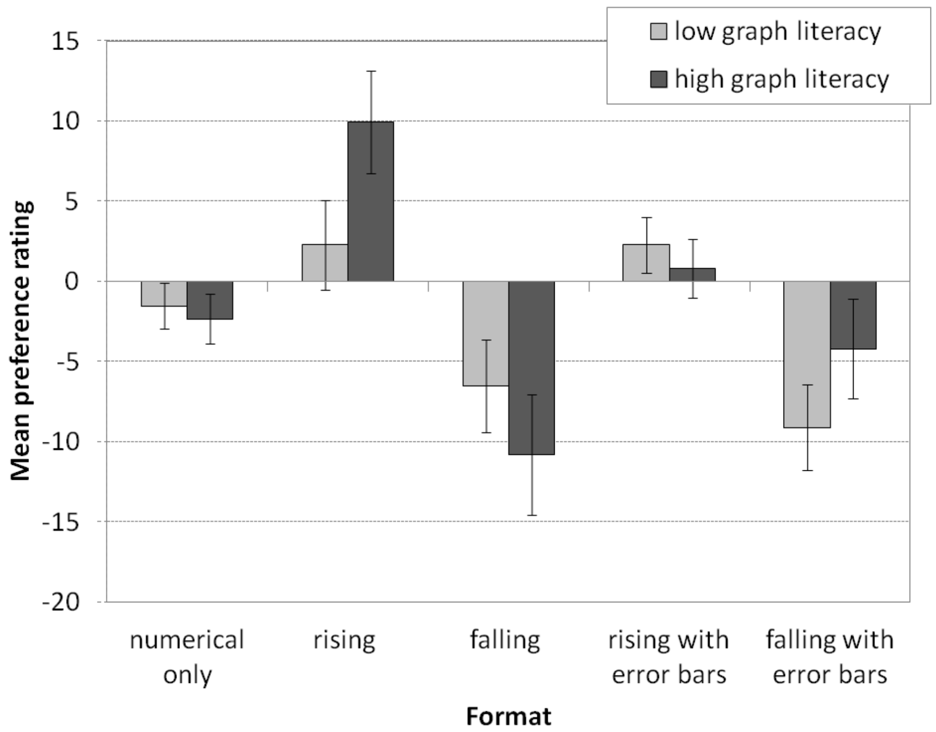


Figure 2. Mean preference rating, by format and graph literacy. Note: A mean rating of 0 indicates a preference for maintaining current glucose levels, while ratings over and below 0 indicate preference to increase and decrease levels, respectively. Error bars represent one standard error of the mean.

Second, we sought to examine relations between the within-the-bar bias and graph literacy, and the effectiveness of error bars to reduce the bias. To this end, we split participants into two groups according to the median graph literacy score for the total sample (i.e., 10 of a possible maximum of 13). The group of participants with low graph literacy included those who obtained 10 or fewer correct responses (n = 285, mean score 8.5, SD = 1.5), while the group of participants with high graph literacy included those who obtained 11

[17] In line with this analysis, a one-way ANOVA on the mean preference ratings for all conditions revealed a main effect of condition in the anticipated directions, $F(4, 453) = 8.80$, $p = .001$, $\eta_p^2 = .07$. When graph literacy was also included as a between-subjects factor, a main effect of condition was also observed, $F(4, 448) = 9.57$, $p = .001$, $\eta_p^2 = .08$. The interaction between condition and graph literacy did not reach statistical significance, $F(4, 448) = 1.61$, $p = .17$, $\eta_p^2 = .01$. However, simple effects contrasts to answer the specific questions of interest are justified when a priori expectations exist (Tybout et al., 2001). All results remained unchanged after controlling for numeracy and for the presence of a chronic disease (e.g., diabetes, thyroid disease).

or more (n = 173, mean score 11.5, SD = .6). Kruskall-Wallis tests revealed that preference ratings varied significantly across conditions both for individuals with low graph literacy, $H(4)$, = 19.06, p = .001, and for those with high graph literacy, $H(4)$, = 19.15, p = .001. To examine differences in the tendency to be affected by the bias in both groups of individuals, we compared ratings for the rising and falling conditions without error bars with ratings for the numerical condition. For participants with low graph literacy, differences were in the anticipated directions but only reached significance for the rising condition (U = 1455.00, r = −.17, p = .030 for rising vs. numerical, and U = 1522.50, r = −.08, p = .18 for falling vs. numerical). In contrast, highly graph literate participants in the rising condition showed a marked preference to increase their glucose levels, relative to the numerical condition, (U = 279.00, r = −.39, p = .001), while those in the falling condition showed a preference to decrease their levels, (U = 388.00, r = −.20, p = .05).

Finally, to test the effect of the intervention for participants with low and high graph literacy, we compared ratings in conditions without error bars to ratings in conditions with error bars. For participants with low graph literacy, no reliable differences were observed between rising and rising with error bars conditions (U = 1537.50, r = −.003, p = .49) or between falling and falling with error bars conditions (U = 1527.50, r = −.08, p = .19). This suggests that the presence of error bars in graphs did not affect preferences among less graph literate individuals. In contrast, error bars significantly reduced the bias among participants with high graph literacy, for graphs with rising bars (U = 487.00, r = −.19, p = .05). The difference between falling and falling with error bars also showed a trend in the anticipated direction (see Figure 2), but did not reach conventional levels of significance (U = 485.00, r = −.13, p = .14). This indicates that error bars may be an effective means to reduce the within-the-bar bias among individuals with sufficient levels of graph literacy, in some conditions.

Discussion

The current study extends prior research showing that participants tend to believe that data points located within bars are more likely to be part of the underlying distribution than equidistant points outside bars (Newman & Scholl, 2012). Here, the *within-the-bar bias* led participants to show a preference to modify their blood glucose levels, even when the information provided gave them no justifiable reason to do so. The bias affected participants who were more graph literate to a larger extent than those who were less skilled with graphs. However, the more graph literate participants were able to benefit from the inclusion of error bars in graphs, which attenuated the within-the-bar bias.

In previous research, graph literacy has been found to protect participants against biases and errors. For example, individuals with high graph literacy tend to rely less on spatial features such as the heights of bars, when such features are non-diagnostic (e.g., when information conveyed by spatial features is in conflict with that conveyed by axes labels or scales; Okan et al., 2012b). Eye-tracking and memory assessments show that superior performance among more graph literate individuals is often driven by strategic allocation of attention and elaborative encoding of task-relevant elements (e.g., values on scales; Okan, Galesic, & Garcia-Retamero, 2013; Woller-Carter, Okan, Cokely, & Garcia-Retamero, 2012; see also Cokely & Kelley, 2009). Interestingly, here we found that higher graph literacy was associated with a marked bias, while less skilled individuals more often avoided the bias. Theoretically, this difference partially reflects individual differences in the extent to which people encoded and attended to the graphs. In the present work, relevant information could be extracted both from the text and from graphs. Less graph literate individuals can be less comfortable with graphs, and thus it seems likely that they spent more time focusing instead on the numerical and text-based information, largely avoiding the graph and the associated bias (for converging evidence, see also Gaissmaier et al., 2012). In contrast, individuals with high graph literacy might have focused more on the graphs, resulting in a detailed representation of the bar graph which would likely be highly accessible and available to bias decision making.

Our finding that error bars reduced the within-the-bars bias differs from results found by Newman and Scholl (2012). The difference in results likely reflects at least two key features of our study. First, our graphs were available during decision making while Newman and Scholl's participants had to rely on memory of their graphs. Second, we measured individual differences in graph literacy and found that the benefits of the error bars were primarily concentrated in highly graph literate individuals. For participants who had some familiarity with error bars and who were more likely to either infer or recognize their meaning (i.e., those with high graph literacy), the inclusion of error bars reduced the within-the-bar bias. Of note, although error bars reduced the bias among highly graph literate individuals, the reduction was only reliable when bars were rising from a lower *x* axis, with a marginal trend when bars were falling from an upper *x* axis. We suspect this may be linked to the specific context of our task, as hyperglycemia might be perceived as having more severe health consequences than hypoglycemia. This may be associated with a general preference to decrease glucose levels, which could magnify the effect of the within-the-bar bias for graphs with falling bars. More research is needed to investigate this issue, as well as the generalizability of the phenomenon, with emphasis on studies that are representative of actual medical or health risk communication practices.

Concerning the perceptual mechanisms that give rise to the within-the-bar bias, Newman and Scholl (2012) argued that the bias occurs because bars are unique visual objects defined by the closure of their boundaries, which originate from one particular axis. Relatedly, Peebles (2008) demonstrated that people presented with bar graphs underestimated the distance of target values to the average (represented by a horizontal line parallel to the *x* axis). Theoretically, visual attention is drawn to the length of bars, which are identified as objects attached to the *x* axis. These accounts converge to indicate that the within-the-bar bias is likely triggered by basic principles of object perception. Bottom-up factors such as the format of graphs can influence the visual chunks that are created, often driven by Gestalt principles including proximity, similarity, and connectedness (Ali & Peebles, 2013; Pinker, 1990). While the visual chunks formed by bars can facilitate tasks such as making discrete comparisons between individual data points (Pinker, 1990) or interpreting interaction data (Ali & Peebles, 2013), they can also lead to systematic misinterpretations of bar graphs, at least among those who encode and represent these features (e.g., graph literate participants). Cognitive process tracing methodologies such as eye-tracking and verbal protocol analysis could shed further light on the role of perceptual and attentional processes in the within-the-bar bias, as well as on their relation with conceptual processes that can also affect the interpretation of bar graphs (e.g., different mappings that viewers establish between spatial features and conceptual relations; Okan et al., 2012b).

Conclusions. Graphical displays are increasingly being used and recommended for the communication of medical risks to the public, and advantages of bar graphs over other kinds of displays (e.g., pie charts) have been documented (e.g., Feldman-Stewart, Brundage, & Zotov, 2007). Unfortunately, the present work accords with previous research showing that bar graphs can lead to judgment and decision making errors (Okan et al., 2012b). This is often the case when graphs are poorly designed, or contain distortions that can mislead graph viewers (e.g., improperly scaled or split axes; Cooper, Schriger, Wallace, Mikulich & Wilkes, 2003; Woller-Carter et al., 2012). Here we demonstrated that bar graphs can also be associated with systematic biases likely caused by basic principles of object perception. We also found that graph literacy does not necessarily protect individuals from this bias, although more graph literate individuals can in some conditions show debiasing when error bars are included in graphs.

Taken together, our findings emphasize that recommendations to use bar graphs to communicate health-related statistics may not always be warranted. Designers and policy makers should proceed with caution, and consider the inclusion of error bars, or the use of alternative graphical formats (e.g., points or depictions of the distributions, as suggested by Newman & Scholl, 2012), where applicable. Ensuring that bar graphs comply with principles of good graph design is most likely necessary, but not sufficient, to promote accurate comprehension and informed decision making.

Acknowledgments

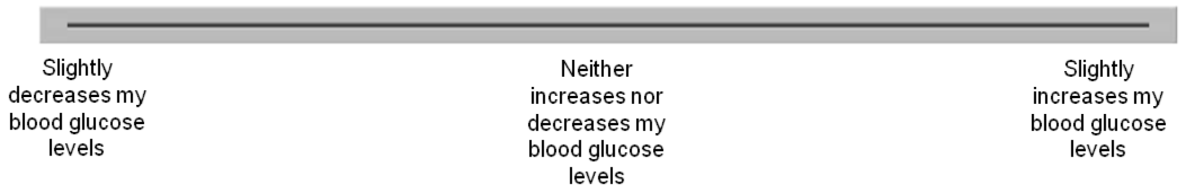
This study has been funded by a grant awarded by the Informed Medical Decisions Foundation (United States) to Yasmina Okan (ref. # 0230-1), and by grants awarded by the Ministerio de Economía y Competitividad (Spain) to Rocio Garcia-Retamero (ref. # PSI2011-22954) and to Antonio Maldonado (ref. # PSI2009-12217). We are grateful to Mirta Galesic and Andrés Catena for their valuable comments and suggestions.

Appendix

Scenario presented. *Note: The text viewed by participants was in Spanish.*

Imagine that you receive the results of your blood glucose levels from the past week. Results are presented using a measurement taken at the start of the week as a reference point. That day, your measurement was ideal (120 mg/dL). Since the start of the week, your blood glucose levels have varied between –20 and +20 in percentage change. If your blood glucose levels are too high (above the measurement taken at the start of the week), or too low (below the measurement taken at the start of the week), you could have a high risk of severe health consequences. Your blood glucose levels were measured several times last week, as levels can vary throughout the day, and can depend on the time passed since the last meal. Your average percentage change throughout the week was zero. However, in some measurements the percentage change was above zero, while in others it was below zero.
Based on this information, you can choose to follow a treatment that will slightly increase your blood glucose levels, or a treatment that slightly decrease your blood glucose levels.

In this case, I would prefer to follow a treatment that…



CHAPTER VII.

GENERAL DISCUSSION AND CONCLUSIONS

Introduction

Well designed graphical displays are often considered to be an effective, transparent, and ethically desirable option to present medical information and improve the communication of risks (Garcia-Retamero & Cokely, in press). However, graphs can readily be employed to obscure information or create favorable impressions of different medical treatments or products (Gigerenzer, Gaissmaier, Kurz-Milcke, Schwartz, & Woloshin, 2007; Kurz-Milcke, Gigerenzer, & Martignon, 2008; Tufte, 2001; Wainer, 1984; Woller-Carter, Okan, Cokely, & Garcia-Retamero, 2012), and individuals with low *graph literacy* can be at a higher risk of misinterpreting information. The experiments reported in this dissertation sought to investigate different design features that can either aid or mislead viewers with varying levels of graph literacy from a theoretically grounded perspective, integrating notions from theories of graph comprehension (Carpenter & Shah, 1998), embodied cognition (Tversky, 2009; Wilson, 2002), skill acquisition and expert performance (Cokely & Kelley, 2009; Haider & Frensch, 1996, 1999). In what follows, I will outline theoretical and practical implications of the reported findings, as well as limitations of the work and promising avenues for future research.

Processes involved in the comprehension of icon arrays

The literature concerned with the graphical communication of health-related statistics has established some links with classic work on perceptual and cognitive judgments involved in graph comprehension (e.g., Ancker, Weber, & Kukafka, 2011a; Price, Cameron, & Butow, 2007; Waters, Weinstein, Colditz, & Emmons, 2006, 2007). However, experiments evaluating the effectiveness of different graphical formats have seldom been grounded on theories of graph comprehension. As noted by Cheng (2001), this may be at least partially due to the fact that research in the cognitive science of graphs has often focused on highly specific aspects of cognition that are hard to scale up to real-world contexts. Accordingly, it is not easy to anticipate how the processes necessary to achieve accurate and meaningful interpretations will be influenced by different design features in graphs. This holds particularly true for icon arrays, as this type of display has received little attention in prominent theories of graph comprehension (e.g., Carpenter & Shah, 1998; Kosslyn, 2006; Pinker, 1990).

Chapters II and **III** attempted to achieve a better understanding of key mechanisms underlying the efficacy of icon arrays in individuals with varying levels of graph literacy. In particular, the experiment reported in **Chapter II** showed that static icon arrays increased accuracy of risk understanding to a larger extent among individuals with high graph literacy. This finding highlighted that the mechanisms thought to underlie the effect of icon arrays (i.e., disentangling overlapping classes; Brainerd & Reyna, 1990; Reyna, 2009; and bringing attention to background information; Stone et al., 2003) might not be sufficient to enhance performance to the same extent for everyone, pointing to differences in processing of icon arrays linked to graph literacy.

The findings reported in **Chapter III** contributed to shed some light on the processes underlying superior performance, demonstrating that risk understanding can be improved when icon arrays are designed to elicit a more active, elaborative, processing of information (c.f., Cokely & Kelley, 2009; Natter & Berry, 2005). Displays that promoted such kind of processing contributed to increase accuracy of risk understanding even among less graph literate individuals. In contrast, encouraging the allocation of attention and encoding of all regions of icon arrays (i.e., first process involved in graph comprehension; Carpenter & Shah, 1998) or facilitating the identification and integration of referents of icon arrays via the inclusion of explanatory labels (i.e., third process involved in graph comprehension; Carpenter & Shah, 1998) was not sufficient to enhance performance. These findings did not enable to determine whether elaborative processing of icon arrays is necessary or sufficient for improved comprehension (see Woller-Carter et al., 2012, for related arguments). However, they stressed that achieving a comprehensive understanding of the mechanisms underlying the efficacy of icon arrays for viewers with varying skills requires considering global graph comprehension processes, as well as the extent to which the displays promote the generalization of risk information. Such considerations will likely help to explain and predict effects of the manipulations of design features of icon arrays such as grouping vs. random arrangement of icons (Ancker et al., 2011a; Feldman-Stewart, Brundage, & Zotov, 2007; Feldman-Stewart, Kocovski, McConnell, Brundage, & Mackillop, 2000; Han et al., 2012; Schapira, Nattinger, & McHorney, 2001; Wright, Whitwell, Takeichi, Hankins, & Marteau, 2009).

Processes involved in the comprehension of bar graphs and line graphs

Processes underlying the comprehension of bar graphs and line graphs have been documented more extensively than those underlying the comprehension of icon arrays. However, the effect of graph literacy on such processes is not well understood. **Chapters IV** and **V** addressed this issue, and showed that graph literacy was associated with a larger likelihood to identify and integrate referents of the variables depicted in bar graphs and line graphs (i.e., third process involved in graph comprehension). This tendency can help viewers to avoid misinterpretations, when the message conveyed by spatial features is in conflict with that conveyed by features linked to arbitrary conventions. These findings expand previous research on perceptual and cognitive processes in graph comprehension (Carpenter & Shah, 1998; Kosslyn, 1989; Lohse, 1993; Pinker, 1990; Shah & Carpenter, 1995; Simkin & Hastie, 1987), documenting the existence of differences in these processes that are linked to individual differences in graph literacy. Additionally, they emphasize that the process of translating visual features into conceptual relations (i.e., second process involved in graph comprehension) can be better understood by taking into account the influence of the associations that viewers may have acquired through their experience with the environment (see also Tversky, 2001, 2009). Further analyses examining transitions between fixations on the different regions of the display revealed that a large proportion of transitions occurred between the areas used to determine referents (e.g., x and y axes) and the pattern, also in line with Carpenter and Shah's (1998) model.

The finding that the link between graph literacy and performance was mediated by the viewing time of numerical scales (but not of titles and axes labels; **Chapter V**) also stresses the theoretical relevance of establishing a conceptual distinction between different types of sources of graph-related errors. For graphs including essential information in scales, differences in allocation of attention can, at least partially, account for differences in performance. However, overcoming errors related to textual elements likely hinges to a larger extent on a certain level of conceptual understanding that supports appropriate inferences. A promising approach that could help to further delineate different theoretically relevant sources of error could involve distinguishing features in graphs that are task-relevant from those that are task-redundant (Haider & Frensch, 1996, 1999). Future research could also examine whether comprehension can be enhanced by manipulations that elicit a more active, elaborative processing of information in textual elements, akin to the manipulations employed in **Chapter III** for icon arrays.

Finally, the study reported in **Chapter VI** contributed to outline some conditions under which higher graph literacy can be associated with larger biases. Results suggest that this can be the case for biases which are triggered by basic principles of object perception, driven by the creation of particular visual chunks (e.g., bars defined by the closure of their boundaries, originating from a particular axis; Newman & Scholl, 2012; Peebles, 2008). This finding was interpreted in terms of a larger tendency among less graph literate individuals to avoid graphs altogether (and therefore the associated biases), when relevant information is also available in text. It is unclear whether individuals with higher levels of graph expertise than those of our more graph literate subjects will show a qualitatively different pattern of results. We suspect that the relationship between graph literacy and the bias might be a curvilinear one, such that highest graph literacy will be associated with lower bias.

Theoretical implications

Differences in people's tendency to rely on information conveyed by spatial features (e.g., bars) vs. on numerical elements in scales to interpret graphs have also been discussed within the framework of a prominent theory in judgment and decision making and probabilistic reasoning, namely fuzzy trace theory (e.g., Brainerd & Reyna, 1990; Reyna & Brainerd, 1995; Reyna et al., 2009). This theory posits that people encode both gist and verbatim representations in parallel, and tend to rely on the former. According to Reyna (2008), in bar graphs the salient gist would be the heights of bars, while verbatim representations would be concrete numbers in scales. Fuzzy trace theory predicts that people will tend to neglect specific numbers, showing an overreliance on the information conveyed by heights of bars. This prediction is in accord with the results reported in **Chapters IV** and **V**.

However, while fuzzy trace theory predicts that increasing expertise will be associated with an increasing preference to operate on crude gist, the present results can be interpreted as evidence that higher levels of graph literacy can in some cases be associated with a larger tendency to operate on verbatim information in graphs. This idea is also supported by evidence from recent studies showing that more graph literate individuals had better memory for verbatim details (e.g., specific numerical information) of graphs used by advertisers and news organizations to communicate risks (Woller-Carter et al., 2012). In contrast, a large body of evidence has showed that expertise in many domains is associated with an increased reliance on semantic gist (Reyna & Brainerd, 2011; Reyna, Chick, Corbi, & Hsia, in press). Thus, the conclusion that this relation is reversed in the domain of graph comprehension in particular (i.e., that experts rely more on verbatim representations) seems hard to substantiate.

Instead, the current findings may be interpreted as evidence that, for specific types of tasks and graphs (i.e., those containing incongruencies between the message conveyed by spatial features and numerical scales), operating on verbatim information becomes essential for accurate interpretations, and graph literacy can be associated with a larger likelihood to incorporate such information. It is plausible that the existence of conflicts in graphs prompts an adaptive adjustment of encoding strategies among more graph literate individuals (c.f., Cokely, Kelley, & Gilchrist, 2006; Cokely & Kelley, 2009), who might strategically shift to a larger reliance on a verbatim-based processing.

Summing up, the experiments reported provide new evidence on the theoretical and practical relevance of the construct of graph literacy, which plays an important role in achieving a unified understanding of graphical communication of medical risks. Future research should seek to determine the exact nature of the representations that individuals with varying levels of graph literacy store and operate on (i.e., Marr's second level of analysis of information-processing systems; 1982), including the extent to which they incorporate propositional elements (Pinker, 1990) vs. analogical or spatial ones (Trafton & Trickett, 2001; Trafton, Marshall, Mintz, & Trickett, 2002). There is some evidence suggesting that higher spatial abilities can be linked to the use of analogical representations for graphical reasoning (Feeney, Adams, Webber, & Ewbank, 2004), suggesting that different types of representations may also be used by individuals with different levels of graph literacy. Addressing these questions would enable to achieve a better *cognitive fit* (Shaft & Vessey, 2006; Vessey, 1991) between people's problem representations and the tasks required to extract information from graphs.

Practical implications

The reported experiments have a number of practical implications for the graphical communication of health-related statistics. First, the results reported in **Chapters II** and **III** suggest that, while icon arrays can constitute a highly efficient means to reduce biases such as denominator neglect, it cannot be assumed that such displays will be beneficial for everyone. The advantages of icon arrays documented in past research (e.g., their capacity to display part-to-whole relations and to represent risks as frequencies; Brase, 2009; Kurz-Milcke et al., 2008) do not guarantee an accurate understanding of risk information for all individuals (see Garcia-Retamero & Galesic, 2010, for converging evidence) and in some cases may even hinder performance (e.g., Ruiz et al., 2013). Taking this into account, strong recommendations concerning the use of icon arrays to communicate medical risks (e.g., Burkell, 2004; Fagerlin & Peters, 2011; Fagerlin, Zikmund-Fisher, & Ubel, 2011) might need to be qualified.

Promoting an active engagement in the information depicted in icon arrays seems to be a promising means to enhance comprehension (**Chapter III**). However, often less can be more (Zikmund-Fisher, 2013; Zikmund-Fisher et al., 2012; Zikmund-Fisher, Dickson, & Witteman, 2011), and therefore icon arrays including dynamic or interactive components should be designed carefully to avoid an increase in cognitive burden. Caution should be taken to avoid overwhelming people's working memory, causing interference, or directing people's attention *away* from relevant information.

Second, the results reported in **Chapters IV** and **V** highlight the negative impact of designing graphs in which information in conventional features is incongruent with that conveyed by spatial features (see Sun, Li, & Bonini, 2011, for related arguments on the effect of manipulations in physical sizes of numerical scales). Moreover, they indicate that visual aids incorporating logarithmic scales (e.g., Paling perspective scales) may in some cases be misunderstood by less graph literate viewers. Accordingly, it may be advisable to bring attention to the scales included in such visual aids, either through features implemented in the displays themselves, or through instructions provided by those administering the visual aids. Additionally, graph designers should seek to preserve the compatibility between the message conveyed by spatial features, textual elements, and the question to be answered. That is, our results converge to suggest that designers of graphs should decide which are the questions that viewers may need to address, and aim to organize data in a way that facilitates those questions to be answered (Kosslyn, 2006).

The results reported in **Chapter IV** also show that horizontally oriented bar graphs can increase the likelihood that highly graph literate viewers incorporate essential elements in scales in their interpretations. However, given that this interaction between graph literacy, orientation, and type of conflict was found unexpectedly, replications of this effect would be desirable before clear prescriptive implications can be derived. The need to interpret the current findings with caution is further emphasized by previous results showing that horizontal bars graphs can in some cases lead to *lower* accuracy (e.g., for gist judgments concerning treatment decisions; Feldman-Stewart et al., 2007). The effect of orientation could not be examined further in the eye tracking studies reported in **Chapter V** due to limitations in the sample size and to the increased complexity of the experimental design employed. However, future work examining how the allocation of attention to scales is affected by their orientation would yield insights with important implications for guidelines of effective graph design. Analyses of reaction times could also contribute to this end, as graph orientation may affect response latencies to a larger extent than accuracy.

Finally, the results reported in **Chapter VI** demonstrate that preferences for medical treatments represented in bar graphs can be biased in systematic ways due to principles of object perception. This expands previous research showing that specific features of bar graphs can be associated with misinterpretations (e.g., for stacked bars, people may often read off the top of bars instead of calculating the difference with the bar below; Mt-Isa et al., 2013). A promising means of reducing the within-the-bar bias was also identified, namely the inclusion of error bars. However, given that less educated people may perceive graphs including representations of uncertainty to be vague and less trustworthy (Schapira et al., 2001), future work should examine alternative debiasing means.

Limitations

One limitation of the reported studies is that, in most cases, participants were relatively well educated. In contrast, the instrument used to evaluate graph literacy (Galesic & Garcia-Retamero, 2011b) was designed to be used with the general public, and validated in nationally representative samples. Accordingly, the distribution of graph literacy scores was often negatively skewed. Thus, future studies should examine whether the observed patterns of results do indeed hold with more diverse samples. Moreover, the quasi-experimental design employed does not allow complete control of extraneous variables. Although some threats to internal validity were reduced by measuring and controlling for the effects of potential confounding factors (e.g., numeracy, motivational factors or carelessness), more research is needed to establish clear causal links between graph literacy, processing differences, and performance.

Another limitation is linked to the use of eye tracking as a means to track cognitive processes, as the interpretation of the specific metrics used is debatable. According to the "eye-mind" hypothesis (Just & Carpenter, 1976) the eye fixates on referents of symbols being processed mentally, implying that eye movements can indicate where a person's attention is being directed in a visual display. However, eye fixations can only be used as indirect indication of cognitive processes, as mental computations do not necessarily produce eye movements, and people do not always see what they fixate on (Wickens & Hollands, 2000). Moreover, interpretations for different eye tracking metrics including fixation frequency, duration of fixations, as well as saccade-derived metrics can vary from task to task and from study to study (Ehmke & Wilson, 2007; Jacob & Karn, 2003; Poole & Ball, 2006; Renshaw, Finlay, Tyfa, & Ward, 2004). Eye movements show researchers *where* participants looked, but not *why* (Eger, Ball, Stevens, & Dodd, 2007).

The adequacy of the metrics analyzed in **Chapter V** and of the ways in which metrics were interpreted is supported by a recent meta-analysis of eye tracking research on differences in the comprehension of visualization liked to expertise (Gegenfurtner, Lehtinen, & Säljö, 2011). However, a more comprehensive understanding of the steps involved in graph comprehension could be achieved by complimenting oculomotor measures with other process tracing techniques such as concurrent think aloud reports (Ericsson & Simon, 1980) or potentially less reactive methods such as retrospective reports cued by eye-movement replay (e.g., Post-Experience Eye-Tracked Protocols; Ball, Eger, Stevens, & Dodd, 2006; Eger et al., 2007) and sequential qualitative interviews (e.g., Nicolson, Knapp, Gardner, & Raynor, 2011). Recent examples of the use of such techniques in graph comprehension research (both in isolation and in conjunction with eye tracking) can be found in Ali and Peebles (2011), Peebles and Ali (2009), Ratwani, Trafton, and Boehm-Davis (2008) and Trafton et al. (2000).

Finally, the dependent variables measured in the current studies and the question formats selected to assess such variables were in some cases limited. For instance, to evaluate understanding of treatment risk reduction, **Chapters II** and **III** employed the procedure reported by Schwartz et al. (1997), which had also been employed in numerous other studies (e.g., Galesic, Garcia-Retamero, & Gigerenzer, 2009). However, different question forms can prompt different kinds of computations and reasoning mechanisms, enhancing the salience of different elements of problems, and thus significantly affecting performance (Girotto & Gonzalez, 2001; Perales & Shanks, 2008). Similarly, **Chapters IV** and **V** examined understanding of information in bar graphs and line graphs using specific information extraction questions. As noted by Ratwani et al. (2008), more complex integration questions might prompt different kinds of graph comprehension processes. Therefore, it would be desirable to determine whether varying the specific format and complexity of the questions used affects results in significant ways. Additionally, other relevant variables, including risk perceptions and perceived threat (see e.g., Keller & Siegrist, 2009; Schapira, Nattinger, & McAuliffe, 2006; Schapira et al., 2001; Siegrist, Orlow, & Keller, 2008; Stone et al., 2003; Stone, Yates, & Parker, 1997; Timmermans, Molewijk, Stiggelbout, & Kievit, 2004), could be examined. The links between accuracy of understanding, treatment decisions (e.g., willingness to take a drug), and behavioral change are not straightforward (Ahmed, Naik, Willoughby, & Edwards, 2012; Waters, Weinstein, Colditz, & Emmons, 2007). Thus, it is necessary to determine the extent to which different design features in graphs can affect real-life decisions and actual behavior.

Open questions for future research

Taken together, the current work leaves a number of questions open for future research. First, it is necessary to achieve a more fine-grained specification of the processes underlying graph comprehension in different viewers (Carpenter & Shah, 1998). For instance, authors investigating the comprehension of choropleth graphs (Ratwani et al., 2008) have suggested that cognitive integration (i.e., the comparison and contrast of aggregate visual clusters) should be distinguished from visual integration (i.e., the integration of individual data components into clusters of information and relation to the referents), and that both types of integration are part of the process of identifying referents contemplated in Carpenter and Shah's (1998) work. Future research could examine whether such distinction can also contribute to outline with more precision the processes underlying the comprehension of graphs depicting health-related data.

Concerning the process of translating visual features into conceptual relations, **Chapters IV** and **V** focused on situations in which mappings could be readily established through information acquired in the environment. However, when visual chunks cannot readily be associated with conceptual relations or viewers lack knowledge to establish the associations, the process of translating visual features into conceptual relations will involve complex inferential processes (Shah, Mayer, & Hegarty, 1999). Thus, future research could also examine how graph literacy affects processing of graphs in such cases. Open questions also remain concerning the interpretation of more complex visual displays. Studies have indicated that the interpretation of some types of complex visualizations (i.e., weather maps) involves the allocation of attention to information in a goal-driven manner and the construction of qualitative mental models from which quantitative information is abstracted (Trafton et al., 2000), as well as spatial transformations and visual imagery (Trafton et al., 2002). The need to engage in such processes and subprocesses for more complex graphs communicating health-relevant information could also be examined.

Future research should also seek to determine the specific factors that give rise to graph literacy, as well as how this skill relates to other types of prior knowledge. Training in statistics acquired through formal education is likely to be a crucial factor affecting graph literacy, as evidenced by the moderate correlations existing between graph literacy scores and educational level (Galesic & Garcia-Retamerob, 2011; Ruiz et al., 2013). However, longitudinal studies are needed to establish the effect of change in graph-related skills as a result of interventions (i.e., explicit instructions; Glazer, 2011). Additionally, links between graph literacy and specific content knowledge should be examined further, as some studies have suggested that novice viewers may in some cases be more likely than experts to rely on knowledge about typical relations to interpret graphs (e.g., accidents and drunk driving; Freedman & Shah, 2002). Another fruitful avenue of research would be the investigation of links between graph literacy and spatial abilities (Kellen, Chan, & Fang, 2013) or object vs. spatial visualization preferences (Blazhenkova & Kozhevnikov, 2008; Garcia-Rodriguez, Summers, & Duxbury, 2011). Further empirical studies are also needed to shed further light on the nature and constituting features of graph schemas, as this widely accepted construct has seldom been examined empirically (see Ratwani & Trafton, 2008, for an exception).

Importantly, future research efforts should include carefully designed graphs, for which interdisciplinary collaborations including researchers in computer science and human computer interaction would be highly beneficial. This could help to overcome flaws of graphs tested in past research, including the failure to preserve proportionality between areas of the different regions and the quantities displayed (see Micallef, Dragicevic, & Fekete, 2012; Ottley, Metevier, Han, & Chang, 2012, for a discussion). This could help to avoid confounds in stimuli used in experiments and maximize the potential of the tested graphs to improve the comprehension of medical risks.

Finally, it would also be interesting to determine the actual use of graphs in clinical settings. The few studies that have examined this issue have revealed that the use of graphs to communicate risks is limited (Neuner-Jehle, Senn, Wegwarth, Rosemann, & Steurer, 2011), pointing to an important gap between recommendations for communicating risk and the reality. It is possible that physicians are aware of the potency of graphs to mislead viewers, thus resorting to employ other formats that may be less suited to improve understanding, but that may seem less potentially troublesome. Moreover, the informational needs of some patients will not always include quantitative information, implying that qualitative risk communications might be sufficient in some cases (Zikmund-Fisher, 2013). Future research examining the questions raised above holds the promise of exploiting the potential of graphs in the contexts in which they may be more needed.

Conclusions

In 1999, Lipkus and Hollands made a call for theoretically driven research examining which graphs are better suited to particular risk communication tasks. Fourteen years later, we are still far from a unified framework that enables to explain and, more importantly, predict, how the use of different types of graphs and the manipulation of different design features in graphs will affect the comprehension of health and medical statistics. The work reported in this dissertation aimed to take some steps toward this goal. Several graph design features that can hinder the comprehension of information were outlined, and mechanisms underlying graph comprehension in individuals of varying skills levels were uncovered.

Graph literacy was identified as a key moderator of the effectiveness of visual aids, as well as a skill that can protect people from biases linked to an overreliance on spatial-to-conceptual mappings, but not necessarily from more basic perceptual biases. Some debiasing methods were identified (e.g., the inclusion of error bars in bar graphs), and key principles for designing graphs that are suitable even for people with low graph literacy were proposed, namely ensuring that spatial and conventional features in graphs convey the same meaning. I hope that this work has contributed to the essential goal of increasing our understanding of the defining features, moderating factors, and mechanisms that give rise to *transparent* and *nontransparent* graphical displays in health.

REFERENCES

References

Ahmed, H., Naik, G., Willoughby, H., & Edwards, A. G. (2012). Communicating risk. *British Medical Journal*, *344*, 40–44.

Ali, N., & Peebles, D. (2011). The different effects of thinking aloud and writing on graph comprehension. In L. Carlson, C. Holscher, & T. Shipley (Eds.), *Proceedings of the twentieth annual conference of the Cognitive Science Society* (pp. 3143–3148). Mahwah, NJ: Lawrence Erlbaum.

Ali, N., & Peebles, D. (2013). The effect of gestalt laws of perceptual organization on the comprehension of three-variable bar and line graphs. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *55*, 183–203.

Ancker, J. S., Senathirajah, Y., Kukafka, R., & Starren, J. B. (2006). Design features of graphs in health risk communication: A systematic review. *Journal of the American Medical Informatics Association*, *13*, 608–618.

Ancker, J. S., Weber, E. U., & Kukafka, R. (2011a). Effect of arrangement of stick figures on estimates of proportion in risk graphics. *Medical Decision Making*, *31*, 143–150.

Ancker, J. S., Weber, E. U., & Kukafka, R. (2011b). Effects of game-like interactive graphics on risk perceptions and decisions. *Medical Decision Making*, *31*, 130–142.

Armstrong, K., FitzGerald, G., Schwartz, J. S., & Ubel, P. A. (2001). Using survival curve comparisons to inform patient decision making. *Journal of General Internal Medicine*, *16*, 482–485.

Armstrong, K., Schwartz, J. S., FitzGerald, G., Putt, M., & Ubel, P. A. (2002). Effect of framing as gain versus loss on understanding and hypothetical treatment choices: Survival and mortality curves. *Medical Decision Making*, *22*, 76–83.

Arunachalam, V., Pei, B. K., & Steinbart, P. J. (2002). Impression management with graphs: Effects on choices. *Journal of Information System, 16,* 183–202.

Bachmann, L. M., Gutzwiller, F. S., Puhan, M. A., Steurer, J., Steurer-Stey, C., & Gigerenzer, G. (2007). Do citizens have minimum medical knowledge? A survey. *BMC Medicine*, *31*, 5–14.

Ball, L. J., Eger, N., Stevens, R., & Dodd, J. (2006). Applying the PEEP method in usability testing. *Interfaces*, *67*, 15–19.

Banks, S. M., Salovey, P., Greener, S., Rothman, A. J., Moyer, A., Beauvais, J., & Epel, E. (1995). The effects of message framing on mammography utilization. *Health Psychology*, *14*, 178–184.

Baron, J. (2000). *Thinking and deciding* (3rd ed.). Cambridge, England: Cambridge University Press.

Baumeister, R. F., Bratslavsky, E., Finkenauer, C., & Vohs, K. D. (2001). Bad is stronger than good. *Review of General Psychology*, *5*, 323–370.

Beattie, V., & Jones, M. J. (1992). The use and abuse of graphs in annual reports: A theoretical framework and an empirical study. *Accounting and Business Research*, *22*, 291–303.

Beattie, V., & Jones, M. J. (2002). Measurement distortion of graphs in corporate reports: An experimental study. *Accounting, Auditing & Accountability Journal*, *15*, 546–564.

Bertin, J. (1983). *Semiology of Graphics: Diagrams, Networks, Maps*. Madison: University of Wisconsin Press.

Blazhenkova, O., & Kozhevnikov, M. (2009). The new object-spatial-verbal cognitive style model: Theory and measurement. *Applied Cognitive Psychology*, *23*, 638–663.

Bodemer, N., & Gaissmaier, W. (2012). Risk communication in health. In S. Roeser, R. Hillerbrand, P. Sandin, & M. Peterson (Eds.), *Handbook of risk theory: Epistemology, decision theory, ethics, and social implications of risk* (Vol. 2, pp. 621–660). Dordrecht: Springer Netherlands.

Bonato, M., Fabbri, S., Umiltà, C., & Zorzi, M. (2007). The mental representation of numerical fractions: Real or integer? *Journal of Experimental Psychology: Human Perception and Performance*, *33*, 1410–1419.

Bonner, C., & Newell, B. R. (2008). How to make a risk seem riskier: The ratio bias versus construal level theory. *Judgment and Decision Making*, *3*, 411–416.

Brainerd, C. J., & Reyna, V. F. (1990). Inclusion illusions: Fuzzy-trace theory and perceptual salience effects in cognitive development. *Developmental Review*, *10*, 365–403.

Brase, G. L. (2009). Pictorial representations in statistical reasoning. *Applied Cognitive Psychology*, *23*, 369–381.

Bruine de Bruin, W., & Bostrom, A. (2013). Assessing what to address in science communication. *Proceedings of the National Academy of Sciences*, *110*, 14062–14068.

Bruine de Bruin, W., Parker, A. M., & Fischhoff, B. (2007). Individual differences in Adult Decision-Making Competence. *Journal of Personality and Social Psychology*, *92*, 938–956.

Burkell, J. (2004). What are the chances? Evaluating risk and benefit information in consumer health materials. *Journal of the Medical Library Association*, *92*, 200–208.

Burns, R., Elzer, S., & Carberry, S. (2008). Estimating effort for trend messages in grouped bar charts. In G. Stapleton, J. Howse, & J. Lee (Eds.), *Diagrammatic representation and inference* (pp. 353–356). Berlin, Germany: Springer.

Campitelli, G., & Labollita, M. (2010). Correlations of cognitive reflection with judgments and choices. *Judgment and Decision Making, 5,* 182–192.

Canham, M., & Hegarty, M. (2010). Effects of knowledge and display design on comprehension of complex graphics. *Learning and Instruction*, *20,* 155–166.

Carpenter, P. A., & Shah, P. (1998). A model of the perceptual and conceptual processes in graph comprehension. *Journal of Experimental Psychology: Applied, 4,* 75–100.

Carswell, C. M. (1992). Choosing specifiers: An evaluation of the basic tasks model of graphical perception. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *34*, 535–554.

Cheng, P. C. H., Cupit, J., & Shadbolt, N. R. (2001). Supporting diagrammatic knowledge acquisition: An ontological analysis of Cartesian graphs. *International Journal of Human-Computer Studies*, *54*, 457–494.

Cleary, P. D., & Angel, R. (1984). The analysis of relationships involving dichotomous dependent-variables. *Journal of Health and Social Behavior*, *25*, 334—348.

Cleveland, W. S. (1994). *The elements of graphing data*. Summit, NJ: Hobart Press.

Cleveland, W. S., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, *79*, 531–554.

Cleveland, W. S., & McGill, R. (1986). An experiment in graphical perception. *International Journal of Man-Machine Studies*, *25*, 491–500.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making, 7,* 25–47.

Cokely, E. T., Ghazal, S., Galesic, M., Garcia-Retamero, R., & Schulz, E. (2013). How to measure risk comprehension in educated samples. In R. Garcia-Retamero & M. Galesic (Eds.), *Transparent communication of health risks: Overcoming cultural differences (pp. 29–52)*. New York, NY: Springer.

Cokely, E. T., & Kelley, C. M. (2009). Cognitive abilities and superior decision making under risk: A protocol analysis and process model evaluation. *Judgment and Decision Making, 4,* 20–33.

Cokely, E. T., Kelley, C. M., & Gilchrist, A. L. (2006). Sources of individual differences in working memory: Contributions of strategy to capacity. *Psychonomic Bulletin & Review*, *13*, 991–997.

Cooper, R. J., Schriger, D. L., Wallace, R. C., Mikulich, V. J., & Wilkes, M. S. (2003). The quantity and quality of scientific graphs in pharmaceutical advertisements. *Journal of General Internal Medicine*, *18*, 294–297.

Cosmides, L., & Tooby, J. (1996). Are humans good intuitive statisticians after all? Rethinking some conclusions from the literature on judgment under uncertainty. *Cognition*, *58*, 1–73.

Cox, R. (1999). Representation construction, externalised cognition and individual differences. *Learning and Instruction*, *9*, 343–363.

Dehaene, S., Bossini, S., & Giraux, P. (1993). The mental representation of parity and number magnitude. *Journal of Experimental Psychology: General*, *122*, 371–396.

Del Missier, F., Mäntylä, T., & Bruine de Bruin, W. (2010). Executive functions in decision making: An individual differences approach. *Thinking and Reasoning*, *16*, 69–97.

Del Missier, F., Mäntylä, T., & Bruine de Bruin, W. (2012). Decision-making competence, executive functioning, and general cognitive abilities. *Journal of Behavioral Decision Making*, *25*, 331–351.

Denes-Raj, V., Epstein, S., & Cole, J. (1995). The generality of the ratio-bias phenomenon. *Personality and Social Psychology Bulletin*, *21*, 1083–1092.

Di Stasi, L. L., Contreras, D., Cándido, A., Cañas, J. J., & Catena, A. (2011). Behavioral and eye-movement measures to track improvements in driving skills of vulnerable road users: First-time motorcycle riders. *Transportation Research Part F: Traffic Psychology and Behaviour*, *14*, 26–35.

Dolan, J. G., Qian, F., & Veazie, P. J. (2012). How well do commonly used data presentation formats support comparative effectiveness evaluations? *Medical Decision Making*, *32*, 840–850.

Edwards, A., Elwyn, G., Covey, J., Matthews, E., & Pill, R. (2001). Presenting risk information a review of the effects of framing and other manipulations on patient outcomes. *Journal of Health Communication*, *6*, 61–82.

Edwards, A., Elwyn, G., & Gwyn, R. (1999). General practice registrar responses to the use of different risk communication tools in simulated consultations: A focus group study. *British Medical Journal*, *319*, 749–752.

Edwards, A., Elwyn, G., & Mulley, A. (2002). Explaining risks: Turning numerical data into meaningful pictures. *British Medical Journal*, *324*, 827–830.

Eger, N., Ball, L. J., Stevens, R., & Dodd, J. (2007). Cueing retrospective verbal reports in usability testing through eye-movement replay. In L. J. Ball, M. A. Sasse, C. Sas, T. C. Ormerod, A. Dix, P. Bagnall, & T. McEwan (Eds.), *People and computers XXI – HCI. . . but not as we know it: Proceedings of HCI 2007* (pp. 129–137). Swindon: The British Computer Society.

Ehmke, C., & Wilson, S. (2007). Identifying web usability problems from eye-tracking data. In L. J. Ball, M. A. Sasse, C. Sas, T. C. Ormerod, A. Dix, P. Bagnall, & T. McEwan (Eds.), *People and computers XXI – HCI. . . but not as we know it: Proceedings of HCI 2007* (pp. 119–128). Swindon: The British Computer Society.

Elting, L. S., Martin, C. G., Cantor, S. B., & Rubenstein, E. B. (1999). Influence of data display formats on physician investigators' decisions to stop clinical trials: Prospective trial with repeated measures. *British Medical Journal*, *318*, 1527–1531.

Elzer, S., Green, N., Carberry, S., & Hoffman, J. (2006). A model of perceptual task effort for bar charts and its role in recognizing intention. *User Modeling and User-Adapted Interaction*, *16*, 1–30.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review*, *102*, 211–245.

Ericsson, K. A., Prietula, M. J., & Cokely, E. T. (2007). The making of an expert. *Harvard Business Review*, *85*, 114–121.

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*, 215–251.

Fagerlin, A., & Peters, E. (2011). Quantitative information. In B. Fischhoff, N. T. Brewer, & J. S. Downs (Eds.), *Communicating risks and benefits: An evidence-based user´s guide* (pp. 53–64). Silver Spring, MD: Food and Drug Administration.

Fagerlin, A., Ubel, P. A., Smith, D. M., & Zikmund-Fisher, B. J. (2007). Making numbers matter: Present and future research in risk communication. *American Journal of Health Behavior*, *31*, S47–S56.

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? *Medical Decision Making*, *25*, 398–405.

Fagerlin, A., Zikmund-Fisher, B. J., & Ubel, P. A. (2011). Helping patients decide: Ten steps to better risk communication. *Journal of the National Cancer Institute*, *103*, 1436–1443.

Feeney, A., Adams, J., Webber, L., & Ewbank, M. (2004). Individual differences in graphical reasoning. *Lecture Notes in Computer Science*, *2980*, 271–285.

Feldman-Stewart, D., Brundage, M. D., & Zotov, V. (2007). Further insight into the perception of quantitative information: Judgments of gist in treatment decisions. *Medical Decision Making*, *27*, 34–43.

Feldman-Stewart, D., Kocovski, N., McConnell, B. A., Brundage, M. D., & Mackillop, W. J. (2000). Perception of quantitative information for treatment decisions. *Medical Decision Making*, *20*, 228–238.

Feltz, A., & Cokely, E.T. (2011). Individual differences in theory-of-mind judgments: Order effects and side effects. *Philosophical Psychology*, *24*, 343–355.

Finucane, M. L., Mertz, C. K., Slovic, P., Schmidt, E. S. (2005). Task complexity and older adults' decision- making competence. *Psychology and Aging, 20*, 71–84.

Fischer, M. H. (2000). Do irrelevant depth cues affect the comprehension of bar graphs? *Applied Cognitive Psychology*, *14*, 151–162.

Fischer, M. H., Dewulf, N., & Hill, R. L. (2005). Designing bar graphs: Orientation matters. *Applied Cognitive Psychology*, *19*, 953–962.

Fischer, M. H. & Zwaan, R.A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *Quarterly Journal of Experimental Psychology*, *61*, 825–850.

Fortin, J. M., Hirota, L. K., Bond, B. E., O'Connor, A. M., & Col, N. F. (2001). Identifying patient preferences for communicating risk estimates: A descriptive pilot study. *BMC Medical Informatics and Decision Making, 1,* 2.

Frederick, S. (2005). Cognitive reflection and decision making. *Journal of Economic Perspectives*, *19,* 24–42.

Freedman, E. G., & Shah, P. (2002). Toward a model of knowledge-based graph comprehension. In M. Hegarty, B. Meyer, & N. H. Narayanan (Eds.), *Diagrammatic representation and inference* (pp. 59–141). Berlin, Germany: Springer.

Friel, S. N., Curcio, F. R., & Bright, G. W. (2001). Making sense of graphs: Critical factors influencing comprehension and instructional implications. *Journal for Research in Mathematics Education*, *32*, 124–158.

Friendly, M. (2008). A brief history of data visualization. In C. Chen, W. Härdle, & A.

Unwin (Eds.), *Handbook of data visualization* (pp. 15–56). Berlin, Germany: Springer.

Friendly, M., & Denis, D. (2005). The early origins and development of the scatterplot. *Journal of the History of the Behavioral Sciences*, *41*, 103–130.

Fuller, R., Dudley, N., & Blacktop, J. (2002). How informed is consent? Understanding of pictorial and verbal probability information by medical inpatients. *Postgraduate Medical Journal, 78*, 543–544.

Gaissmaier, W., Wegwarth, O., Skopec, D., Müller, 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. *Health Psychology, 31*, 286–296.

Galesic, M., & Garcia-Retamero, R. (2010). Statistical numeracy for health: A cross-cultural comparison with probabilistic national samples. *Archives of Internal Medicine*, *170,* 462–468.

Galesic, M., & Garcia-Retamero, R. (2011a). Do low-numeracy people avoid shared decision making? *Health Psychology*, *30*, 336–341.

Galesic, M., & Garcia-Retamero, R. (2011b). Graph literacy: A cross-cultural comparison. *Medical Decision Making, 31*, 444–457.

Galesic, M., Garcia-Retamero, R., & Gigerenzer, G. (2009). Using icon arrays to communicate medical risks: Overcoming low numeracy. *Health Psychology*, *28,* 210–216.

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 of Experimental Psychology: Applied*, *17*, 270–287.

Garcia-Retamero, R., & Cokely, E. T. (in press). Communicating health risks with visual aids. *Current Directions in Psychological Science.* doi: 10.1177/0963721413491570

Garcia-Retamero, R., & Dhami, M. K. (2009). Take-the-best in expert-novice decision strategies for residential burglary. *Psychonomic Bulletin and Review*, *16,* 163–169.

Garcia-Retamero, R., & Dhami, M. K. (2011). Pictures speak louder than numbers: on communicating medical risks to immigrants with limited non-native language proficiency. *Health Expectations*, *14*, 46–57.

Garcia-Retamero, R., & Galesic, M. (2009). Communicating treatment risk reduction to people with low numeracy skills: A cross-cultural comparison. *American Journal of*

Public Health, *99*, 2196–2202.

Garcia-Retamero, R., & Galesic, M. (2010a). How to reduce the effect of framing on messages about health. *Journal of General Internal Medicine, 25*, 1323–1329.

Garcia-Retamero, R., & Galesic, M. (2010b). Who profits from visual aids: Overcoming challenges in people's understanding of risks. *Social Science & Medicine*, *70*, 1019–1025.

Garcia-Retamero, R., & Galesic, M. (2011). Using plausible group sizes to communicate information about medical risks. *Patient Education and Counseling*, *84*, 245–250.

Garcia-Retamero, R., & Galesic, M. (2013). *Transparent communication of health risks: Overcoming cultural differences*. New York, NY: Springer.

Garcia-Retamero, R., Galesic, M., & Gigerenzer, G. (2010). Do icon arrays help reduce denominator neglect? *Medical Decision Making*, *30*, 672−684.

Garcia-Rodríguez, S., Summers, B., & Duxbury, D. (2011). *Visualization, cognitive style, numeracy and decision making*. Poster presented at the 23rd Subjective Probability, Utility, and Decision Making Conference, Kingston, United Kingdom.

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. *Patient Education and Counseling*, *83*, 398–403.

Gattis, M. (2002). Structure mapping in spatial reasoning. *Cognitive Development*, *17*, 1157–1183.

Gattis, M. (2004). Mapping relational structure in spatial reasoning. *Cognitive Science*, *28*, 589–610.

Gattis, M., & Holyoak, K. J. (1996). Mapping conceptual to spatial relations in visual reasoning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 231–239.

Gegenfurtner, A., Lehtinen, E., & Säljö, R. (2011). Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educational Psychology Review*, *23*, 523–552.

Gerend, M. A., & Shepherd, J. E. (2007). Using message framing to promote acceptance of the human papillomavirus vaccine. *Health Psychology*, *26*, 745–752.

Gevers, W., Reynvoet, B., & Fias, W. (2003). The mental representation of ordinal sequences is spatially organized. *Cognition*, *3*, B87–B95.

Gigerenzer, G., & Brighton, H. (2009). Homo heuristicus: Why biased minds make better inferences. *Topics in Cognitive Science*, *1*, 107–143.

Gigerenzer, G., Gaissmaier, W., Kurz-Milcke, E., Schwartz, L. M., & Woloshin, S. (2007). Helping doctors and patients make sense of health statistics. *Psychological Science in the Public Interest*, *8*, 53–96.

Gigerenzer, G., Todd, P. M., & the ABC Research Group (1999). *Simple heuristics that make us smart.* New York: Oxford University Press.

Gillan, D. J., Wickens, C. D., Hollands, J. G., & Carswell, C. M. (1998). Guidelines for presenting quantitative data in HFES publications. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *40*, 28–41.

Girotto, V., & Gonzalez, M. (2001). Solving probabilistic and statistical problems: A matter of information structure and question form. *Cognition*, *78*, 247–276.

Glazer, N. (2011). Challenges with graph interpretation: A review of the literature. *Studies in Science Education*, *47*, 183–210.

Grossarth-Maticek, R., & Ziegler, R. (2008). Randomized and non-randomized prospective controlled cohort studies in matched pair design for the long-term therapy of corpus uteri cancer patients with a mistletoe preparation. *European Journal of Medical Research*, *13*, 107–120.

Haider, H., & Frensch, P. A. (1996). The role of information reduction in skill acquisition. *Cognitive Psychology*, *30*, 304–337.

Haider, H., & Frensch, P. A. (1999). Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*, 172–190.

Han, P. K., Klein, W. M., Killam, B., Lehman, T., Massett, H., & Freedman, A. N. (2012). Representing randomness in the communication of individualized cancer risk estimates: Effects on cancer risk perceptions, worry, and subjective uncertainty about risk. *Patient Education and Counseling*, *86*, 106–113.

Hansen, C., & Hansen, R. (1988). Finding the face in the crowd: An anger superiority effect. *Journal of Personality and Social Psychology*, *54*, 917–924.

Haward, M. F., Murphy, R. O., & Lorenz, J. M. (2008). Message framing and perinatal decisions. *Pediatrics*, *122*, 109–118.

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. *Patient Education and Counseling*, *73*, 448–455.

Hayes, A. F. (2013). *An introduction to mediation, moderation, and conditional process analysis: A regression-based approach*. New York, NY: Guilford Press.

Hegarty, M., & Waller, D. (2005). Individual differences in spatial abilities. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 121–169). New York, NY: Cambridge University Press.

Hess, R., Visschers, V. H., & Siegrist, M. (2011). Risk communication with pictographs: The role of numeracy and graph processing. *Judgment and Decision Making*, *6*, 263–274.

Huestegge, L., & Philipp, A. M. (2011). Effects of spatial compatibility on integration processes in graph comprehension. *Attention, Perception, & Psychophysics*, *73*, 1903–1915.

Jacob, R. J., & Karn, K. S. (2003). Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In J. Hyona, R. Radach, & H. Deubel (Eds.) *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 573–605). Amsterdam, The Netherlands: Elsevier Science.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior*, *17*, 649–667.

Jarodzka, H., Scheiter, K., Gerjets, P., & van Gog, T. (2010). In the eyes of the beholder: How experts and novices interpret dynamic stimuli. *Learning and Instruction, 20,* 146–154.

Johnson, J. R., Rice, R., & Roemmich, R. A. (1980). Pictures that lie: The abuse of graphs in annual reports. *Management Accounting*, *62*, 50–56.

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*, 441–480.

Kellen, V., Chan, S., & Fang, X. (2013). Improving user performance in conditional probability problems with computer-generated diagrams. In M. Kurosu (Ed.), *Human-Computer Interaction. Users and contexts of use* (pp. 183–192). Berlin, Germany: Springer.

Keller, C., & Siegrist, M. (2009). Effect of risk communication formats on risk perception depending on numeracy. *Medical Decision Making*, *29*, 483–490.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A Construction-Integration model. *Psychological Review*, *95*, 163–182.

Kintsch, W. (2004). The construction-integration model of text comprehension and its implications for instruction. *Theoretical Models and Processes of Reading*, *5*, 1270–1328.

Kosslyn, S. M. (1989). Understanding charts and graphs. *Applied Cognitive Psychology*, *3*, 185–226.

Kosslyn, S. M. (2006). *Graph design for the eye and mind*. New York, NY: Oxford University Press.

Kramarski, B., & Mevarech, Z. R. (2003). Enhancing mathematical reasoning in the classroom: The effects of cooperative learning and metacognitive training. *American Educational Research Journal*, *40*, 281–310.

Kurz-Milcke, E., Gigerenzer, G., & Martignon, L. (2008). Transparency in risk communication. *Annals of the New York Academy of Sciences*, *1128*, 18–28.

Kutner, M. A., Greenberg, E., Jin, Y., & Paulsen, C. (2006). *The health literacy of America's adults: Results from the 2003 National Assessment of Adult Literacy*. Washington, DC: U.S. Department of Education, National Center for Education Statistics.

Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. Chicago, IL: University of Chicago Press.

Lee, A. Y., & Hutchison, L. (1998). Improving learning from examples through reflection. *Journal of Experimental Psychology: Applied*, *4*, 187–210.

Levin, I. P., Schneider, S. L., & Gaeth, G. J. (1998). All frames are not created equal: A typology and critical analysis of framing effects. *Organizational Behavior and Human Decision Processes*, *76*, 149–188.

Lewandowsky, S., Ecker, U. K., Seifert, C. M., Schwarz, N., & Cook, J. (2012). Misinformation and its correction continued influence and successful debiasing. *Psychological Science in the Public Interest*, *13*, 106–131.

Lichtenberg, P., Levinson, D., Sharshevsky, Y., Feldman, D., & Lachman, M. (2008). Clinical case management of revolving door patients–a semi-randomized study. *Acta Psychiatrica Scandinavica*, *117*, 449–454.

Lipkus, I. M. (2007). Numeric, verbal, and visual formats of conveying health risks: Suggested best practices and future recommendations. *Medical Decision Making*, *27*, 696–713.

Lipkus, I. M., & Hollands, J. G. (1999). The visual communication of risk. *Journal of the National Cancer Institute Monographs*, 25, 149–163.

Lipkus, I. M., & Peters, E. (2009). Understanding the role of numeracy in health: Proposed theoretical framework and practical insights. *Health Education & Behavior*, *36*, 1065–1081.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making*, *21*, 37–44.

Lloyd, F. J., & Reyna, V. F. (2001). A Web exercise in evidence-based medicine using cognitive theory. *Journal of General Internal Medicine*, *16*, 94–99.

Lohse, G. L. (1993). A cognitive model for understanding graphical perception. *Human-Computer Interaction*, *8*, 353–388.

Lunney, G. H. (1970). Using analysis of variance with a dichotomous dependent variable—An empirical study. *Journal of Educational Measurement*, *7*, 263—269.

Madden, C. J., & Zwaan, R.A. (2006). Perceptual representation as a mechanism of lexical ambiguity resolution: An investigation of span and processing time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *32*, 1291–1303.

Maichle, U. (1994). Cognitive processes in understanding line graphs. In W. Schnotz & R. W. Kulhavy (Eds.), *Comprehension of graphics* (pp. 207–227). Amsterdam, The Netherlands: North-Holland Elsevier Science.

Marr, D. (1982). *Vision: A computational investigation into the human representation and processing of visual information*. New York, NY: Freeman.

Marteau, T. M. (1989). Framing of information: Its influence upon decisions of doctors and patients. *British Journal of Social Psychology*, *28*, 89–94.

Mazur, D. J., & Hickam, D. H. (1990). Interpretation of graphic data by patients in a general medicine clinic. *Journal of General Internal Medicine*, *5*, 402–405.

Mazur, D. J., & Hickam, D. H. (1993). Patients' and physicians' interpretations of graphic data displays. *Medical Decision Making*, *13*, 59–63.

Mazur, D. J., & Merz, J. F. (1993). How the manner of presentation of data influences older patients in determining their treatment preferences. *Journal of the American Geriatrics Society*, *41*, 223–228.

McCaffery, K. J., Dixon, A., Hayen, A., Jansen, J., Smith, S., & Simpson, J. M. (2012). The influence of graphic display format on the interpretations of quantitative risk information among adults with lower education and literacy: A randomized experimental study. *Medical Decision Making*, *32*, 532–544.

McKenzie, D. L., & Padilla, M. J. (1986). The construction and validation of the test of graphing in science (TOGS). *Journal of Research in Science Teaching*, *23*, 571–579.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition and Instruction*, *14*, 1–43.

McNeil, B. J., Pauker, S. G., Sox, H. C., & Tversky, A. (1982). On the elicitation of preferences for alternative therapies. *New England Journal of Medicine*, *306*, 1259–1262.

Meade, A. W., & Craig, S. B. (2012). Identifying careless responses in survey data. *Psychological Methods*, *17*, 437–455.

Mevarech, Z. R., & Stern, E. (1997). Interaction between knowledge and contexts on understanding abstract mathematical concepts. *Journal of Experimental Child Psychology*, *65*, 68–95.

Micallef, L., Dragicevic, P., & Fekete, J.-D. (2012). Assessing the effect of visualizations on Bayesian reasoning through crowdsourcing. *IEEE Transactions on Visualization and Computer Graphics*, *18*, 2536–2545.

Mitchum, A. L., & Kelley, C. M. (2010). Solve the problem first: Constructive solution strategies can influence the accuracy of retrospective confidence judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 699–710.

Morgan, M. G., Fischhoff, B., Bostrom, A., & Atman, C. J. (2002). *Risk communication: A mental models approach*. Cambridge: Cambridge University Press.

Mt-Isa, S., Hallgreen, C. E., Asiimwe, A., Downey, G., Genov, G., Hermann, R., … Tzoulaki (2013). *Review of visualisation methods for the representation of benefit-risk assessment of medication: Stage 2 of 2*. Retrieved from http://www.imi-protect.eu/documents/ShahruletalReviewofvisualisationmethodsfortherepresentationofBRassessmentofmedicationStage2A.pdf

Munzner, T., Johnson, C., Moorhead, R., Pfister, H., Rheingans, P., & Yoo, T. S. (2006). NIH-NSF visualization research challenges report summary. *IEEE Computer Graphics and Applications*, *26*, 20–24.

Natter, H. M., & Berry, D. C. (2005). Effects of active information processing on the understanding of risk information. *Applied Cognitive Psychology*, *19*, 123–135.

Neuner-Jehle, S., Senn, O., Wegwarth, O., Rosemann, T., & Steurer, J. (2011). How do family physicians communicate about cardiovascular risk? Frequencies and determinants of different communication formats. *BMC Family Practice*, *12*, 15.

Newman, G. E., & Scholl, B. J. (2012). Bar graphs depicting averages are perceptually misinterpreted: The within-the-bar bias. *Psychonomic Bulletin & Review*, *19*, 601–607.

Ni, Y., & Zhou, Y. D. (2005). Teaching and learning fraction and rational numbers: The origins and implications of whole number bias. *Educational Psychologist*, *40*, 27–52.

Nicolson, D. J., Knapp, P., Gardner, P., & Raynor, D. K. (2011). Combining concurrent and sequential methods to examine the usability and readability of websites with information about medicines. *Journal of Mixed Methods Research*, *5*, 25–51.

Novick, L. R. (2006). The importance of both diagrammatic conventions and domain-specific knowledge for diagram literacy in science: The hierarchy as an illustrative case. In D. Barker-Plummer, R. Cox, & N. Swoboda (Eds.), *Diagrams 2006* (pp. 1–11). Berlin. Germany: Springer.

Öhman, A., Lundqvist, D., & Esteves, F. (2001). The face in the crowd revisited: A threat advantage with schematic stimuli. *Journal of Personality and Social Psychology*, *80*, 381–396.

Okan, Y., Galesic, M., & Garcia-Retamero, R. (2010). *Graph comprehension in medical contexts: An eye-tracking study*. Poster presented at the 13th Biennial European Conference of the Society for Medical Decision Making, Hall in Tyrol.

Okan, Y., Galesic, M., & Garcia-Retamero, R. (2013a). *How people with low and high graph literacy process health graphs: Evidence from eye-tracking*. Manuscript submitted for publication.

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2012a). Individual differences in graph literacy: Overcoming denominator neglect in risk comprehension. *Journal of Behavioral Decision Making*, *25*, 390–401.

Okan, Y., Garcia-Retamero, R., Galesic, M., & Cokely, E. T. (2012b). When higher bars are not larger quantities: On individual differences in the use of spatial information in graph comprehension. *Spatial Cognition & Computation*, *12*, 1–25.

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2013b). *Biasing and debiasing health decisions with bar graphs: Costs and benefits of graph literacy*. Manuscript submitted for publication

Okan, Y., Garcia-Retamero, R., Cokely, E. T., & Maldonado, A. (2013c). *Enhancing accuracy of risk comprehension using dynamic icon arrays*. Manuscript submitted for publication.

Ottley, A., Metevier, B., Han, P. K., & Chang, R. (2012). *Visually communicating Bayesian* *statistics to laypersons*. Retrieved from http://www.cs.tufts.edu/tech_reports/reports/2012-02/report.pdf

Paling, J. (2003). Strategies to help patients understand risks. *British Medical Journal*, *327*, 745–748.

Paolacci, G., Chandler, J., & Ipeirotis, P. G. (2010). Running experiments on Amazon Mechanical Turk. *Judgment and Decision Making*, *5*, 411–419.

Parker, A. M., Bruine de Bruin, W. B., Yoong, J., & Willis, R. (2012). Inappropriate confidence and retirement planning: Four studies with a national sample. *Journal of Behavioral Decision Making*, *25*, 382–389.

Parker, A. M., & Fischhoff, B. (2005). Decision-making competence: External validation through an individual-differences approach. *Journal of Behavioral Decision Making*, *18*, 1–27.

Peebles, D. (2008). The effect of emergent features on judgments of quantity in configural and separable displays. *Journal of Experimental Psychology: Applied*, *14*, 85–100.

Peebles, D., & Ali, N. (2009). Differences in comprehensibility between three-variable bar and line graphs. In N. Taatgen, H. van Rijn, J. Nerbonne & L. Schomaker (Eds.). *Proceedings of the thirty-first annual conference of the Cognitive Science Society* (pp. 2938–2943). Mahwah, NJ: Lawrence Erlbaum.

Peebles, D., & Cheng, P. C. H. (2001). Graph-based reasoning: From task analysis to cognitive explanation. In J. D. Moore & K. Stenning (Eds.), *Proceedings of the Twenty Third Annual Conference of the Cognitive Science Society* (pp. 762–767). Mahwah, NJ: Erlbaum.

Peebles, D., & Cheng, P. C. H. (2003). Modeling the effect of task and graphical representation on response latency in a graph reading task. *Human Factors*, *45*, 28–45.

Penna, M. P., Agus, M., Peró-Cebollero, M., Guàrdia-Olmos, J., & Pessa, E. (2012). The use of imagery in statistical reasoning by university undergraduate students: A preliminary study. *Quality & Quantity*, 1–15.

Pennington, R., & Tuttle, B. (2009). Managing impressions using distorted graphs of income and earnings per share: The role of memory. *International Journal of Accounting Information Systems*, *10*, 25–45.

Penrose, J. M. (2008). Annual report graphic use: A review of the literature. *Journal of Business Communication*, *45*, 158–180.

Perales, J. C., & Shanks, D. R. (2008). Driven by power? Probe question and presentation format effects on causal judgment. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 1482–1494.

Peters, E. (2012). Beyond comprehension: The role of numeracy in judgments and decisions. *Current Directions in Psychological Science*, *21*, 31–35.

Peters, E., Hibbard, J., Slovic, P., & Dieckmann, N. (2007). Numeracy skill and the communication, comprehension, and use of risk-benefit information. *Health Affairs, 26*, 741–748.

Peters, E. & Levin, I. P. (2008). Dissecting the risky-choice framing effect: Numeracy as an individual-difference factor in weighting risky and riskless options. *Judgment and Decision Making, 3,* 435–448.

Peters, E., Västfjäll, D., Slovic, P., Mertz, C. K., Mazzocco, K., & Dickert, S. (2006). Numeracy and decision making. *Psychological Science*, *17*, 407–413.

Pinker, S. (1990). A theory of graph comprehension. In R. Freedle (Ed.), *Artificial intelligence and the future of testing* (pp. 73–126). Hillsdale, NJ: Erlbaum.

Poole, A., & Ball, L. J. (2006). Eye tracking in HCI and usability research: Current status and future prospects. In Ghaoui, C. (Ed.), *Encyclopedia of Human-Computer Interaction* (pp. 211–219). Hershey, PA: Idea Group.

Pratto, F., & John, O. P. (1991). Automatic vigilance: The attention-grabbing power of negative social information. *Journal of Personality and Social Psychology*, *61*, 380–391.

Price, M., Cameron, R., & Butow, P. (2007). Communicating risk information: the influence of graphical display format on quantitative information perception—Accuracy, comprehension and preferences. *Patient Education and Counseling*, *69*, 121–128.

Ratwani, R. M., & Trafton, J. G. (2008). Shedding light on the graph schema: Perceptual features versus invariant structure. *Psychonomic Bulletin & Review*, *15*, 757–762.

Ratwani, R. M., Trafton, J. G., & Boehm-Davis, D. A. (2008). Thinking graphically: Connecting vision and cognition during graph comprehension. *Journal of Experimental Psychology: Applied*, *14,* 36–49.

Renshaw, J. A., Finlay, J. E., Tyfa, D., & Ward, R. D. (2003). Designing for visual influence: An eye tracking study of the usability of graphical management information. *Human-Computer Interaction*, *1*, 144–151.

Renshaw, J. A., Finlay, J. E., Tyfa, D., & Ward, R. D. (2004). Understanding visual influence in graph design through temporal and spatial eye movement characteristics. *Interacting with Computers*, *16*, 557–578.

Reyna, V. F. (1991). Class inclusion, the conjunction fallacy, and other cognitive illusions. *Developmental Review, 11*, 317–336.

Reyna, V. F. (2004). How people make decisions that involve risk. *Current Directions in Psychological Science*, *13*, 60–66.

Reyna, V. F. (2008). A theory of medical decision making and health: Fuzzy trace theory. *Medical Decision Making*, *28*, 850–865.

Reyna, V. F., & Brainerd, C. J. (1994). The origins of probability judgment: A review of data and theories. In G. Wright & P. Ayton (Eds.), *Subjective Probability* (pp. 239–272). New York: Wiley.

Reyna, V. F., & Brainerd, C. J. (1995). Fuzzy-trace theory: An interim synthesis. *Learning and Individual Differences*, *7*, 1–75.

Reyna, V. F., & Brainerd, C. J. (2008). Numeracy, ratio bias, and denominator neglect in judgments of risk and probability. *Learning and Individual Differences*, *18*, 89–107.

Reyna, V. F., & Brainerd, C. J. (2011). Dual processes in decision making and developmental neuroscience: A fuzzy-trace model. *Developmental Review*, *31*, 180–206.

Reyna, V. F., Chick, C. F., Corbi, J. C., & Hsia, A. N. (in press). Developmental reversals in risky decision-making: Intelligence agents show larger decision biases than college students. *Psychological Science*.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How numeracy influences risk comprehension and medical decision making. *Psychological Bulletin*, *135*, 943–973.

Rivers, S. E., Salovey, P., Pizarro, D. A., Pizarro, J., & Schneider, T. R. (2005). Message framing and pap test utilization among women attending a community health clinic. *Journal of Health Psychology*, *10*, 65–77.

Roth, W. M., & Bowen, G. M. (2003). When are graphs worth ten thousand words? An expert-expert study. *Cognition and Instruction*, *21*, 429–473.

Rothman, A. J., Martino, S. C., Bedell, B. T., Detweiler, J. B., & Salovey, P. (1999). The systematic influence of gain-and loss-framed messages on interest in and use of different types of health behavior. *Personality and Social Psychology Bulletin*, *25*, 1355–1369.

Rothman, A. J., & Salovey, P. (1997). Shaping perceptions to motivate healthy behavior: The role of message framing. *Psychological Bulletin*, *121*, 3–19.

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? *Patient Education and Counseling*. Advance online publication. doi: http://dx.doi.org/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 of Health Communication: International Perspectives*, *11*, 569–582.

Schapira, M. M., Nattinger, A. B., & McHorney, C. A. (2001). Frequency or probability? A qualitative study of risk communication formats used in health care. *Medical Decision Making*, *21*, 459–467.

Schapira, M. M., Walker, C. M., Cappaert, K. J., Ganschow, P. S., Fletcher, K. E., McGinley, E. L., … Jacobs, E. A. (2012). The Numeracy Understanding in Medicine Instrument: A Measure of health numeracy developed using Item Response Theory. *Medical Decision Making*, *32*, 851–865.

Schonlau, M., & Peters, E. (2008). Graph comprehension: An experiment in displaying data as bar charts, pie charts and tables with and without the gratuitous 3rd dimension. *RAND Working Paper Series No. WR- 618*. Retrieved from http://ssrn.com/abstract=1272627.

Schonlau, M., & Peters, E. (2012). Comprehension of graphs and tables depend on the task: Empirical evidence from two web-based studies. *Statistics, Politics, and Policy*, *3*; *Article 5*.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The role of numeracy in understanding the benefit of screening mammography. *Annals of Internal Medicine*, *127*, 966–972.

Shaft, T. M., & Vessey, I. (2006). The role of cognitive fit in the relationship between software comprehension and modification. *MIS Quarterly*, *30*, 29–55.

Shah, P. (2001). Graph comprehension: The role of format, content, and individual differences. In M. Anderson, B. Meyer, & P. Olivier (Eds.), *Diagrammatic representation and reasoning* (pp. 173–185). Berlin: Springer-Verlag.

Shah, P., & Carpenter, P. A. (1995). Conceptual limitations in comprehending line graphs. *Journal of Experimental Psychology: General*, *124*, 43–61.

Shah, P., & Freedman, E. G. (2011). Bar and line graph comprehension: An interaction of top-down and bottom-up processes. *Topics in Cognitive Science*, *3,* 560–578.

Shah, P., & Hoeffner, J. (2002). Review of graph comprehension research: Implications for instruction. *Educational Psychology Review*, *14*, 47–69.

Shah, P., Freedman, E. G., & Vekiri, I. (2005). The comprehension of quantitative information in graphical displays. In P. Shah & A. Miyake (Eds.), *The Cambridge handbook of visuospatial thinking* (pp. 426–476). New York, NY: Cambridge University Press.

Shah, P., Mayer, R. E., & Hegarty, M. (1999). Graphs as aids to knowledge construction: Signaling techniques for guiding the process of graph comprehension. *Journal of Educational Psychology*, *91*, 690–702.

Shaki, S., Fischer, M. H., & Petrusic, W. M. (2009). Reading habits for both words and numbers contribute to the SNARC effect. *Psychonomic Bulletin & Review*, *16*, 328–331.

Shanteau, J. (1992). How much information does an expert use? Is it relevant? *Acta Psychologica*, *81*, 75–86.

Shiloh, S., Salton, E., & Sharabi, D. (2002). Individual differences in rational and intuitive thinking styles as predictors of heuristic responses and framing effects. *Personality and Individual Differences*, *32,* 415–429.

Siegrist, M., Orlow, P., & Keller, C. (2008). The effect of graphical and numerical presentation of hypothetical prenatal diagnosis results on risk perception. *Medical Decision Making*, *28,* 567–574.

Simkin, D., & Hastie, R. (1987). An information-processing analysis of graph perception. *Journal of the American Statistical Association*, *82,* 454–465.

Simon, H. A. (1996). *The sciences of the artificial* (3rd ed.). Cambridge, MA: MIT Press.

Smith, N. K., Cacioppo, J. T., Larsen, J. T., & Chartrand, T. L. (2003). May I have your attention, please: Electrocortical responses to positive and negative stimuli. *Neuropsycologia*, *41*, 171–183.

Spence, I. (2005). No humble pie: The origins and usage of a statistical chart. *Journal of Educational and Behavioral Statistics*, *30,* 353–368.

Spence, I., & Lewandowsky, S. (1991). Displaying proportions and percentages. *Applied Cognitive Psychology*, *5*, 61–77.

Spiegelhalter, D., Pearson, M., & Short, I. (2011). Visualizing uncertainty about the future.

Science, *333*, 1393–1400.

Sprouse, J. (2011). A validation of Amazon Mechanical Turk for the collection of acceptability judgments in linguistic theory. *Behavior Research Methods*, *43*, 1–13.

Stanovich, K. E., & West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate. *Behavioral and Brain Sciences*, *23*, 645–726.

Stanovich, K. E., & West, R. F. (2008). On the relative independence of thinking biases and cognitive ability. *Journal of Personality and Social Psychology*, *94*, 672–695.

Stern, E., Aprea, C., & Ebner, H. G. (2003). Improving cross-content transfer in text processing by means of active graphical representation. *Learning and Instruction*, *13*, 191–203.

Sternberg, R. J. (1977). Component processes in analogical reasoning. *Psychological Review*, *84*, 353–378.

Stone, E. R., Sieck, W. R., Bull, B. E., Yates, J. F., Parks, S. C., & Rush, C. J. (2003). Foreground: background salience: Explaining the effects of graphical displays on risk avoidance. *Organizational Behavior and Human Decision Processes*, *90*, 19–36.

Stone, E. R., Yates, J. F., & Parker, A. M. (1997). Effects of numerical and graphical displays on professed risk-taking behavior. *Journal of Experimental Psychology: Applied*, *3*, 243–256.

Sun, Y., Li, S., & Bonini, N. (2010). Attribute salience in graphical representations affects evaluation. *Judgment and Decision Making*, *5*, 151–158.

Tabachneck-Schijf, H. J., Leonardo, A. M., & Simon, H. A. (1997). Camera: A computational model of multiple representations. *Cognitive Science*, *21*, 305–350.

Tait, A. R., Voepel-Lewis, T., Brennan-Martinez, C., McGonegal, M., & Levine, R. (2012). Using animated computer-generated text and graphics to depict the risks and benefits of medical treatment. *The American Journal of Medicine*, *125*, 1103–1110.

Tait, A. R., Voepel-Lewis, T., Zikmund-Fisher, B. J., & Fagerlin, A. (2010). Presenting research risks and benefits to parents: Does format matter? *Anesthesia & Analgesia*, *111*, 718–723.

Timmermans, D. R., Ockhuysen-Vermey, C. F., & Henneman, L. (2008). Presenting health risk information in different formats: The effect on participants' cognitive and emotional evaluation and decisions. *Patient Education and Counseling*, *73*, 443–447.

Timmermans, D., Molewijk, B., Stiggelbout, A., & Kievit, J. (2004). Different formats for communicating surgical risks to patients and the effect on choice of treatment. *Patient*

Education and Counseling, *54*, 255–263.

Toll, B. A., Martino, S., Latimer, A., Salovey, P., O'Malley, S., Calin-Menter, S., Hopkins J., Wu, R., Celestino, P., & Cummings, K. M. (2010). Randomized trial: Quitline specialist training in gain-framed vs standard-care messages for smoking cessation. *Journal of National Cancer Institute*, *102*, 96–16.

Trafton, J. G., Kirschenbaum, S. S., Tsui, T. L., Miyamoto, R. T., Ballas, J. A., & Raymond, P. D. (2000). Turning pictures into numbers: extracting and generating information from complex visualizations. *International Journal of Human-Computer Studies*, *53*, 827–850.

Trafton, J., Marshall, S., Mintz, F., & Trickett, S. (2002). Extracting explicit and implict information from complex visualizations. In M. Hegarty, B. Meyer, & N. H. Narayanan (Eds.), *Diagrammatic representation and inference* (pp. 206–220). Berlin, Germany: Springer.

Trafton, J. G., & Trickett, S. B. (2001). A new model of graph and visualization usage. In J. D. Moore & K. Stening (Eds.), *Proceedings of the twenty third annual conference of the Cognitive Science Society* (pp. 1048–1053). Mahwah, NJ: Erlbaum.

Trickett, S., & Trafton, J. (2006). Toward a comprehensive model of graph comprehension: Making the case for spatial cognition. In D. Barker-Plummer, R. Cox, & N. Swoboda (Eds.), *Diagrammatic representation and inference* (pp. 286–300). Berlin, Germany: Springer.

Tufte, E. R. (1997). *Visual explanations: images and quantities, evidence and narrative*. Cheshire, CT: Graphics Press.

Tufte, E. R. (2001). *The visual display of quantitative information* (2nd ed.). Cheshire, CT: Graphics Press.

Tuijnman, A. (2000). *Benchmarking adult literacy in America: An international comparative study (International Adult Literacy Survey)*. Washington, DC: U.S. Department of Education, Office of Vocational and Adult Education.

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, MA: Addison-Wesley.

Tversky, B. (2001). Spatial schemas in depictions. In M. Gattis (Ed.), *Spatial schemas and abstract thought* (pp. 79–112). Cambridge, MA: MIT Press.

Tversky, B. (2009). Spatial cognition: Embodied and situated. In P. Robbins & M. Aydede (Eds.), *The Cambridge handbook of situated cognition* (pp. 201–217). Cambridge, England: Cambridge University Press.

Tversky, B., Kugelmass, S., & Winter, A. (1991). Cross-cultural and developmental trends in graphic productions. *Cognitive Psychology*, *23*, 515–557.

Tversky, B., Morrison, J. B., & Betrancourt, M. (2002). Animation: Can it facilitate? *International Journal of Human-Computer Studies*, *57*, 247–262.

Tybout, A., Sternthal, B., Keppel, G., Verducci, J., Meyers-Levy, J., Barnes, J., … Maxwell, S. (2001). *Journal of Consumer Psychology*, *10*, 5−35.

Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, *22*, 219–240.

Vigneau, F., Caissie, A. F., & Bors, D. A. (2005). Eye-movement analysis demonstrates strategic influences on intelligence. *Intelligence*, *34*, 261–272.

Wainer, H. (1984). How to display data badly. *The American Statistician*, *38*, 137–147.

Wainer, H., & Velleman, P. F. (2001). Statistical graphics: Mapping the pathways of science. *Annual Review of Psychology*, *52*, 305–335.

Waters, E. A., Weinstein, N. D., Colditz, G. A., & Emmons, K. (2006). Formats for improving risk communication in medical tradeoff decisions. *Journal of Health Communication*, *11*, 167–182.

Waters, E. A., Weinstein, N. D., Colditz, G. A., & Emmons, K. M. (2007). Reducing aversion to side effects in preventive medical treatment decisions. *Journal of Experimental Psychology: Applied*, *13*, 11–21.

Weller, J. A., Dieckmann, N. F., Tusler, M., Mertz, C. K., Burns, W. J., & Peters, E. (2013). Development and testing of an abbreviated numeracy scale: A Rasch Analysis approach. *Journal of Behavioral Decision Making*, *26*, 198–212.

Wickens, C.D., & Hollands, J. G. (2000). *Engineering psychology and human performance*. (3rd ed.). Upper Saddle River, NJ: Prentice-Hall.

Wilson, D. K., Kaplan, R. M., & Schneiderman, L. J. (1987). Framing of decisions and selections of alternatives in health care. *Social Behaviour*, *2*, 51–59.

Wilson, M. (2002). Six views of embodied cognition. *Psychonomic Bulletin & Review*, *9*, 625–636.

Woller-Carter, M., Okan, Y., Cokely, E. T., & Garcia-Retamero, R. (2011). *On cognitive abilities and superior judgment: An eye tracking study of errors in medical, political, and consumer inferences*. Poster presented at the 32nd Annual Conference of the Society for Judgment and Decision Making, Seattle, WA.

Woller-Carter, M., Okan, Y., Cokely, E. T., & Garcia-Retamero, R. (2012). Communicating and distorting risks with graphs: An eye-tracking study. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *56*, 1723–1727.

Wright, A. J., Whitwell, S. C., Takeichi, C., Hankins, M., & Marteau, T. M. (2009). The impact of numeracy on reactions to different graphic risk presentation formats: An experimental analogue study. *British Journal of Health Psychology*, *14*, 107–125.

Yamagishi, K. (1997). When a 12.86% mortality is more dangerous than 24.14%: implications for risk communication. *Applied Cognitive Psychology, 11*, 495–506.

Zacks, J., Levy, E., Tversky, B., & Schiano, D. J. (1998). Reading bar graphs: Effects of extraneous depth cues and graphical context. *Journal of Experimental Psychology: Applied*, *4*, 119–138.

Zhao, X., Lynch, J. G., Jr., & Chen, Q. (2010). Reconsidering Baron and Kenny: Myths and truths about mediation analysis. *Journal of Consumer Research*, *37*, 197–206.

Zikmund-Fisher, B. J. (2013). The right tool is what they need, not what we have: A taxonomy of appropriate levels of precision in patient risk communication. *Medical Care Research and Review*, *70*, 37S–49S.

Zikmund-Fisher, B. J., Dickson, M., & Witteman, H. O. (2011). Cool but counterproductive: Interactive, web-based risk communications can backfire. *Journal of Medical Internet Research*, *13*, e60.

Zikmund-Fisher, B. J., Ubel, P. A., Smith, D. M., Derry, H. A., McClure, J. B., Stark, A., … Fagerlin, A. (2008). Communicating side effect risks in a tamoxifen prophylaxis decision aid: The debiasing influence of pictographs. *Patient Education and Counseling*, *73*, 209–214.

Zikmund-Fisher, B. J., Witteman, H. O., Fuhrel-Forbis, A., Exe, N. L., Kahn, V. C., & Dickson, M. (2012). Animated graphics for comparing two risks: A cautionary tale. *Journal of Medical Internet Research*, *14*, e106.

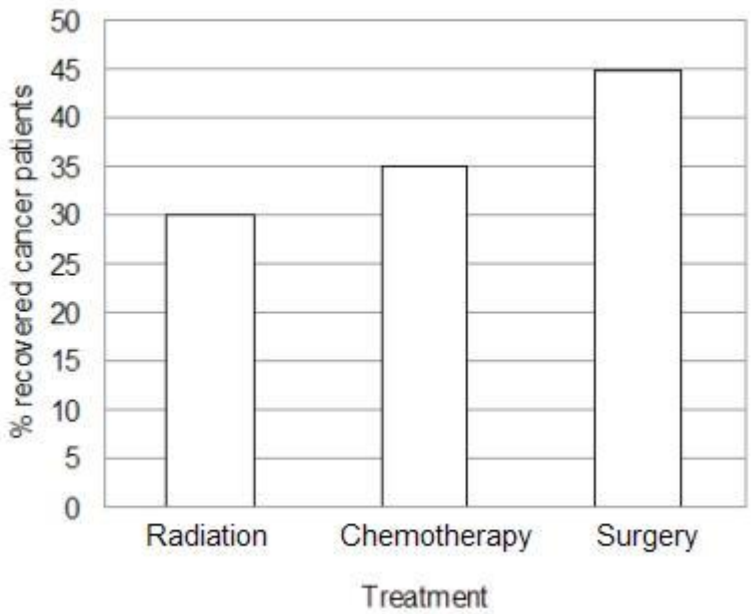
Zwaan, R. A., & Taylor, L. J. (2006). Seeing, acting, understanding: Motor resonance in language comprehension. *Journal of Experimental Psychology: General*, *135*, 1–11.

APPENDIX

Graph literacy scale

(Galesic & Garcia-Retamero, 2011b)

Here is some information about cancer therapies.



Q1. What percentage of patients recovered after chemotherapy?

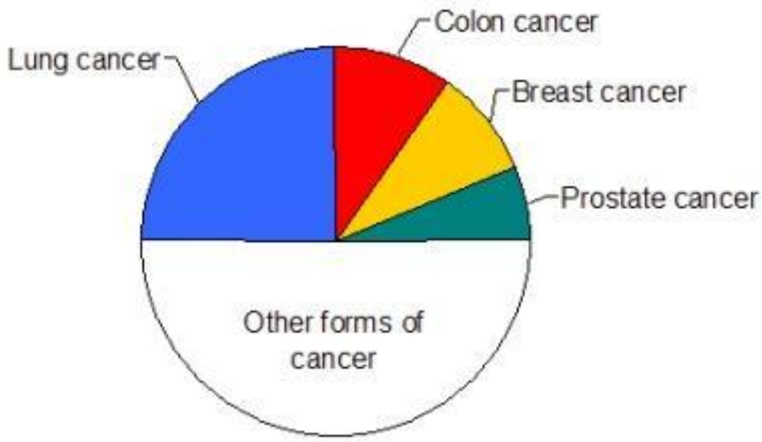
%

Q2. What is the difference between the percentage of patients who recovered after a surgery and the percentage of patients who recovered after radiation therapy?

%

Here is some information about different forms of cancer.

Percentage of people that die from different forms of cancer



Q3. Of all the people who die from cancer, approximately what percentage dies from lung cancer?

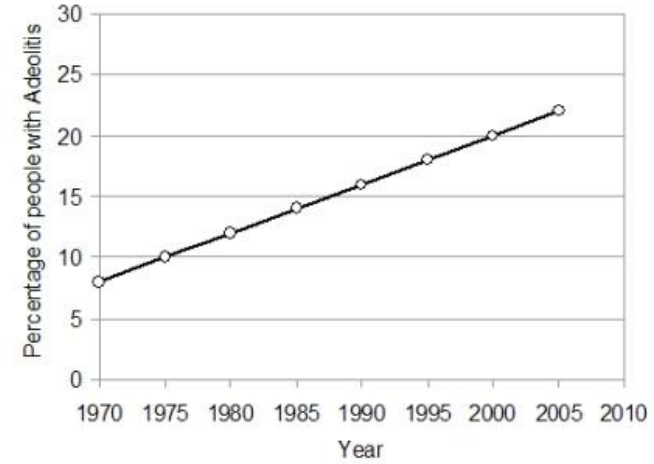
%

Q4. Approximately what percentage of people who die from cancer die from colon cancer, breast cancer, and prostate cancer taken together?

%

Here is some information about an imaginary disease called Adeolitis.

Percentage of people with Adeolitis



Q5. Approximately what percentage of people had Adeolitis in the year 2000?

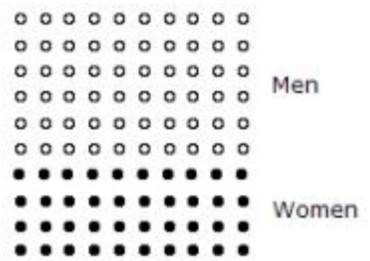
%

Q6. When was the increase in the percentage of people with Adeolitis higher?
From 1975 to 1980………………………………………..1
From 2000 to 2005………………………………………..2
Increase was the same in both intervals…………………...3
Don't know………………………………………………..4

Q7. According to your best guess, what will the percentage of people with Adeolitis be in the year 2010?

%

The following figure shows the number of men and women among patients with disease X. The total number of circles is 100.

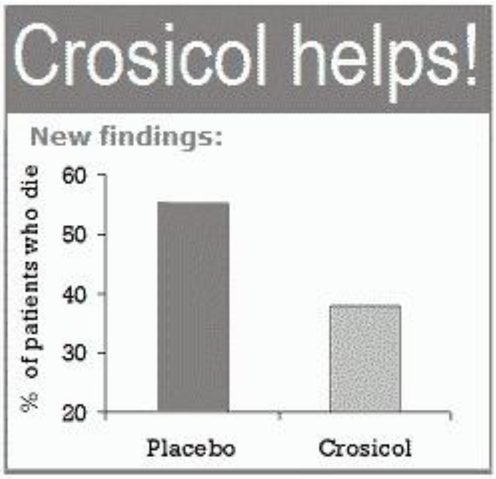


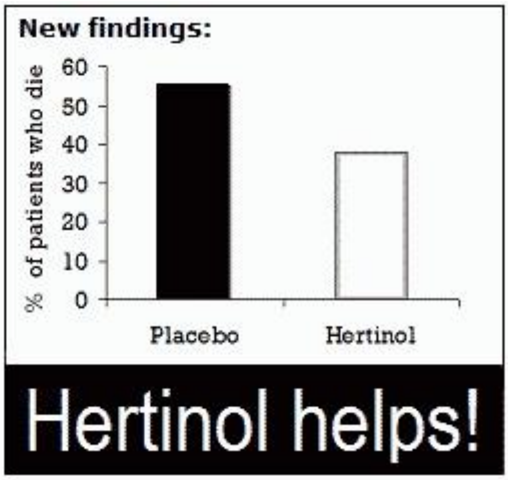
Q8. Of 100 patients with disease X, how many are women?

Q9. How many more men than women are there among 100 patients with disease X?

men

Q10. In a magazine you see two advertisements, one on page 5 and another on page 12. Each is for a different drug for treating heart disease, and each includes a graph showing the effectiveness of the drug compared to a placebo (sugar pill).

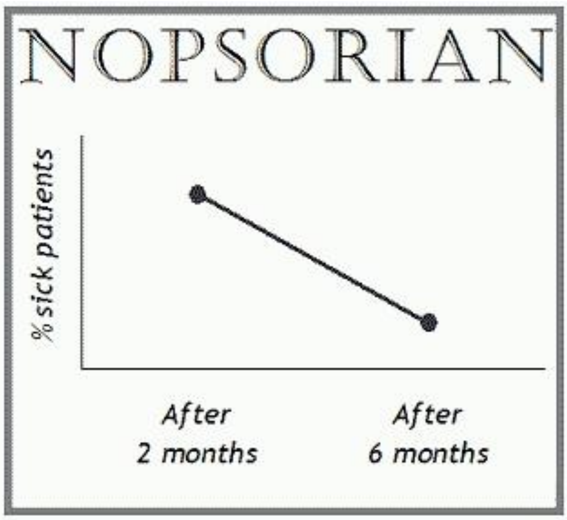




Compared to the placebo, which treatment leads to a larger decrease in the percentage of patients who die?

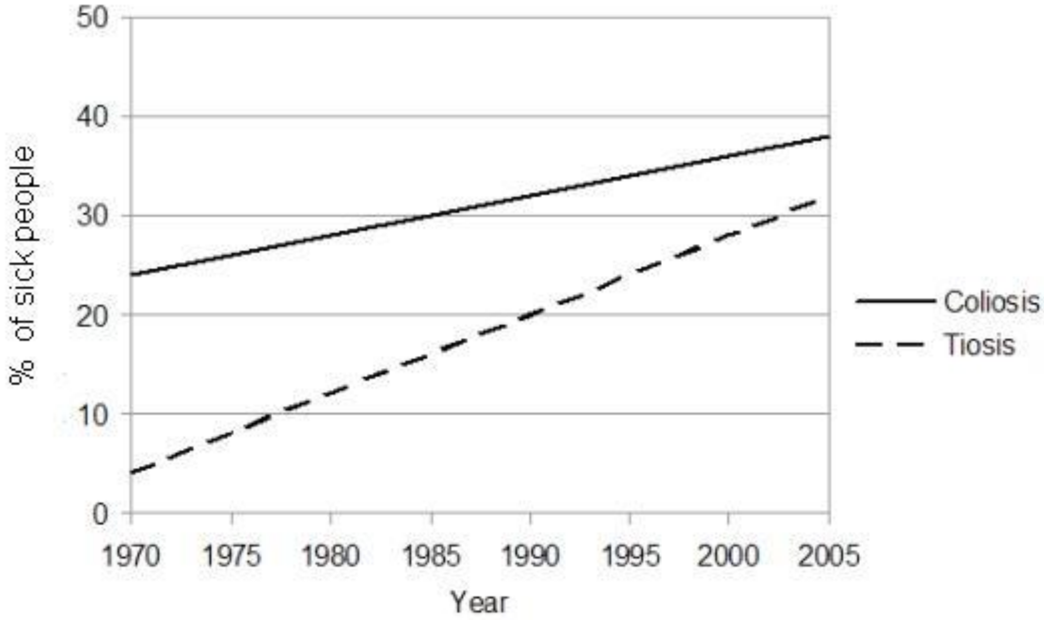
Crosicol………………………..1
Hertinol………………………..2
They are equal…………………3
Can't say………………………4

Q11. In the newspaper you see two advertisements, one on page 15 and another on page 17. Each is for a different treatment of psoriasis, and each includes a graph showing the effectiveness of the treatment over time.



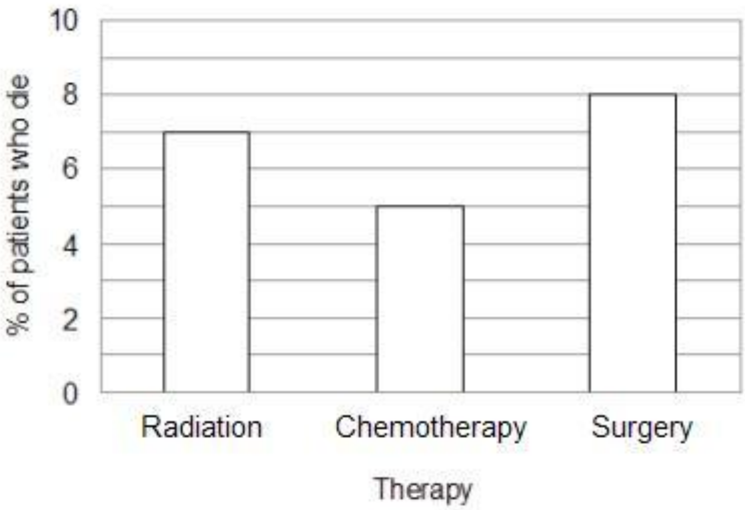
Which of the treatments contributes to a larger decrease in the percentage of sick patients?
Apsoriatin………………….1
Nopsorian………………….2
They are equal……………..3
Can't say…………………...4

Q12. Here is some information about the imaginary diseases Coliosis and Tiosis.



Between 1980 and 1990, which disease had a higher increase in the percentage of people affected?
Coliosis……………………………1
Tiosis……………………………...2
The increase was equal……………3
Can't say…………………………..4

Q13. Here is some information about cancer therapies.

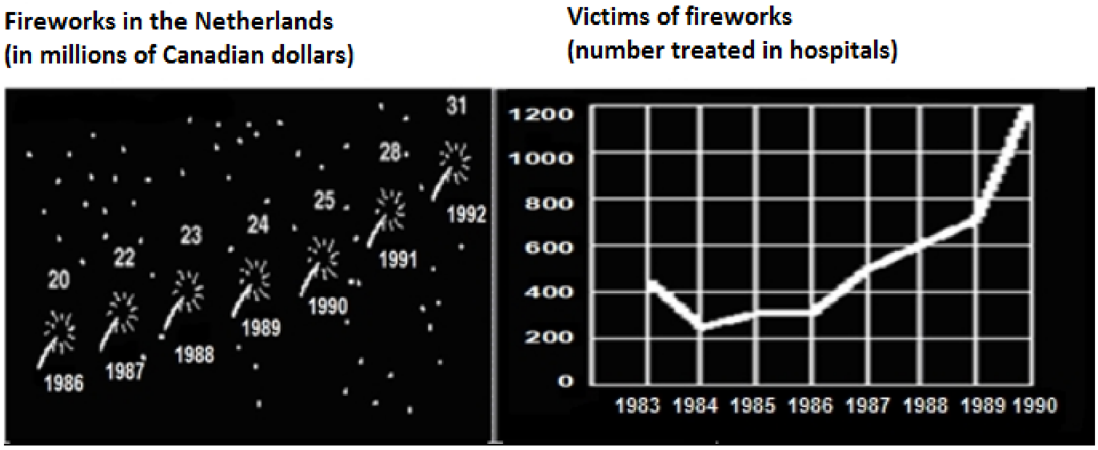


What is the percentage of cancer patients who die after chemotherapy?
%

Additional items to measure graph literacy

1. Below you can see information about Fireworks in the Netherlands.

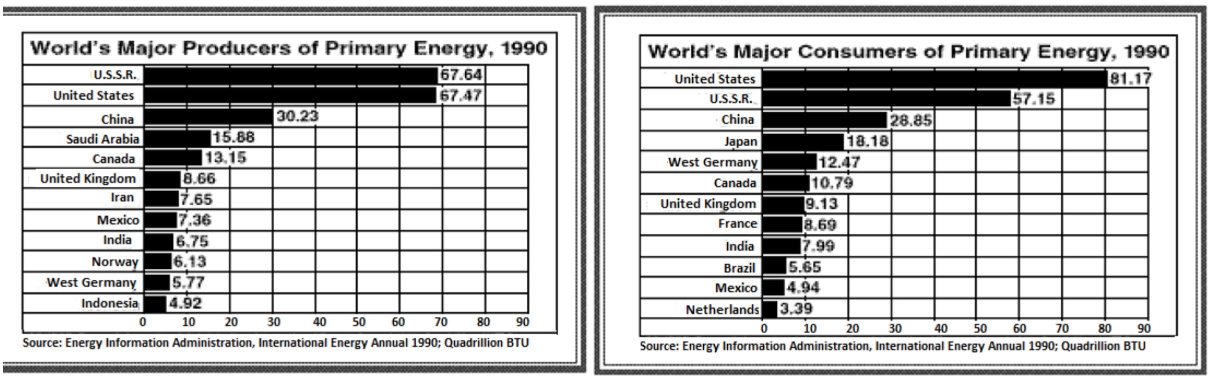
How many more people were injured in 1989 than in 1988? ☐ People



(Source: International Adult Literacy Survey, Document literacy)

2. Below you can see information concerning the World's major producers and consumers of primary energy

How much more energy is produced than consumed in Canada? ☐ Quadrillion BTU

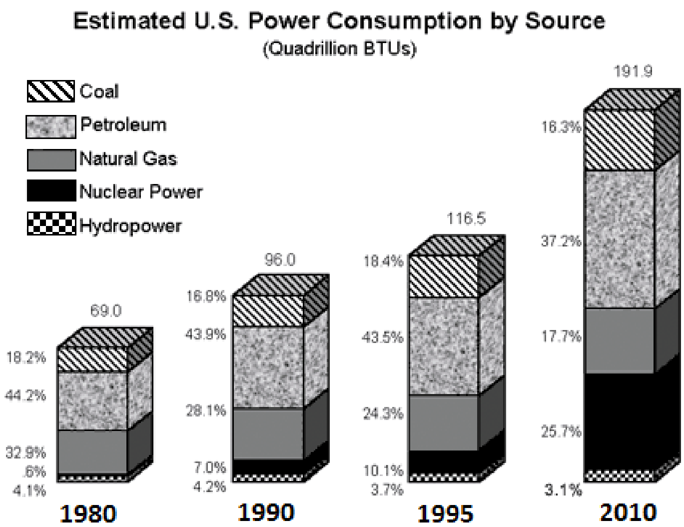


(Source: International Adult Literacy Survey, Mathematical literacy)

3. The graph below shows predictions of United States energy consumption through the year 2010. Use the graph to answer the question that follows.

In the year 2010, which energy source is predicted to supply a larger percentage of the total power than it did in 1990?

A. Coal
B. Petroleum
C. Natural gas
D. Nuclear power
E. Hydropower
F. I don´t know

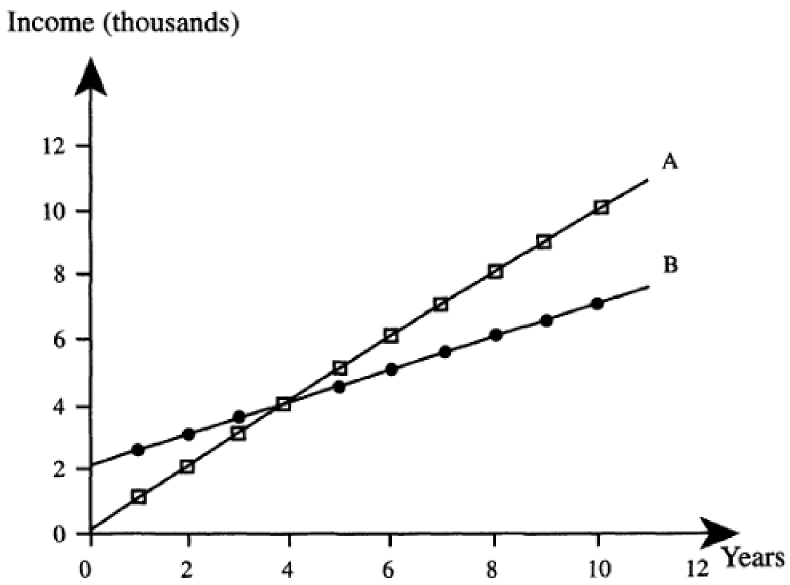


(Source: National Assesment of Adult Literacy)

4. The graph below represents the income of two companies between the years 1990 and 2000.

Until the year 1994, the change rate in the income of Company B was...

A. Greater than the change rate in the income of Company A
B. Smaller than the change rate in the income of Company A
C. Equal to the change rate in the income of Company A



(Source: Kramarski & Meravech, 2003)