Original Article

Research Synthesis Methods

Received 10 June 2013, Revised 29 June 2014, Accepted 11 July 2014 Published online 14 August 2014 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/jrsm.1125

Finding your way out of the forest without a trail of bread crumbs: development and evaluation of two novel displays of forest plots

Anne H. E. Schild[a,b*†] and Martin Voracek[a,c]

Research has shown that forest plots are a gold standard in the visualization of meta-analytic results. However, research on the general interpretation of forest plots and the role of researchers' meta-analysis experience and field of study is still unavailable. Additionally, the traditional display of effect sizes, confidence intervals, and weights have repeatedly been criticized. The current work presents an online statistical cognition experiment in which a total of 279 researchers with experience in meta-analysis from 36 countries evaluated conventional forest plots and two novel versions of forest plots, namely, thick forest plots and rainforest plots. The results indicate certain biases in the interpretation of forest plots, especially with regard to heterogeneity, the distribution of weights, and the theoretical concept of confidence intervals. Although the two novel displays (thick forest plots and rainforest plots) are associated with slightly longer viewing times, they are at least as well-suited and esthetically and perceptively pleasing as the conventional displays while facilitating the correct and exhaustive interpretation of the meta-analytic information. Furthermore, it is advisable to combine conventional forest plots with distribution information of the individual effects, make confidence lines more visually striking, and to display a background grid in the graph. Copyright © 2014 John Wiley & Sons, Ltd.

Keywords: meta-analysis; research synthesis; evaluation; statistical cognition experiment; thick forest plot; rainforest plot

1. Introduction

A recent systematic review in which graph use in meta-analyses between three different disciplines was compared (medicine, psychology, and business) revealed that forest plots are the most prominent and widely used graphical displays for meta-analysis (Schild and Voracek, 2013). Among the three disciplines, graph use was highest in medical journals. In medicine alone, forest plots accounted for nearly 70% of all graphs. In psychology journals, graph use was much less prevalent, and forest plots were much rarer, accounting for only 13.2% of all graphs. Graph use in business was low in general, and no forest plots were identified. Overall, more than half of all observed graphs were forest plots.

The first forest plots date back to the early 1970s, and they found their way into meta-analysis in 1982 (Lewis and Ellis, 1982; Lewis and Clarke, 2001). As they are highly versatile plots that show both individual and pooled effect sizes and enable the viewer to appraise between-study heterogeneity, their popularity has grown steadily over the years. Today, the use of forest plots to display results from meta-analysis is recommended in the *Preferred Reporting Items for Systematic Reviews and Meta-Analyses*-Statement (PRISMA, Liberati *et al*., 2009; Moher *et al*., 2009), the most comprehensive and widely used guidelines on the preparation and publication of meta-analyses. The prominence of forest plots in published meta-analyses and their role in publication guidelines underline the status of this graph type as a gold standard.

[a]*Department of Basic Psychological Research and Research Methods, School of Psychology, University of Vienna, Vienna, Austria*
[b]*Knowledge Media Research Center, Leibniz-Institut für Wissensmedien, Tübingen, Germany*
[c]*Georg Elias Müller Institut, Georg August Universität of Göttingen, Göttingen, Germany*
**Correspondence to: Anne H. E. Schild, Department of Basic Psychological Research and Research Methods, School of Psychology, University of Vienna, Vienna, Austria.*
[†]*E-mail: anne.schild@univie.ac.at*

Considering the crucial role of forest plots in meta-analysis, it is astounding that research on the interpretation of forest plots is still lacking and has focused only on narrow and specific aspects. A recent study suggests that rater agreement is highest for the assessment of between-study heterogeneity in forest plots as compared with box plots, residual histograms, and Galbraith and l'Abbé plots (Bax *et al*., 2009). However, these findings focused only on the single aspect of heterogeneity and involved only three raters, all with considerable expertise in meta-analysis. Therefore, fundamental questions remain unanswered concerning the overall interpretation of these graphs by viewers.

1.1. Design and criticisms of conventional forest plots

In forest plots (see Anzures-Cabrera and Higgins, 2010, for details on design and usage suggestions), individual effect sizes together with their confidence intervals (usually 95%) are presented. In most cases, effect sizes are shown for each study individually, but summary plots, in which effect sizes for groups of studies are shown, are another possibility. In addition, forest plots usually also include the summary effect and corresponding confidence interval of the meta-analysis. It is common practice to display individual study effect sizes with a square with size proportional to the weight assigned to the study in the meta-analysis. Summary effects are usually presented with a diamond whose center indicates the magnitude of the effect and whose length indicates the upper and lower limit of the confidence interval. Studies are often presented in alphabetical order, although ordering by publication year or other external criteria may greatly enhance ease of interpretation (see Schriger *et al*., 2010, for suggestions).

This common design of forest plots has repeatedly been criticized (Barrowman and Myers, 2003; Jackson, 2008; Anzures-Cabrera and Higgins, 2010; Schriger *et al*., 2010; Schild and Voracek, 2013), and three points have been put forward: First, small studies may attract an undue amount of attention because of their long confidence intervals. Although small studies have less impact on the overall pooled result, they may appear more visually striking than larger ones. This is worrisome, as this might lead to an interpretational bias toward potentially questionable small study effects. Although information on study size was added to the plots through the scaling of the effect size boxes (Lewis and Clarke, 2001), this solution is only a partial one, as small studies still receive much emphasis because of the long thin confidence lines.

Second, because of the practice of plotting the size of the squares in proportion to the study's weight in the analysis, individual point estimates may be difficult to discern. As a result, viewers may make false assumptions about individual effects sizes, differences between individual studies, and heterogeneity may be misjudged. Furthermore, in the case of very large studies, variability of the true effect may be underestimated because the visual impact of the confidence intervals may be unduly diminished in comparison with the large box size.

Third, viewers may be tempted to overlook the fact that the likelihood of the values within the individual confidence interval decreases as they move toward the outer boundaries of the interval. That is, readers may believe that all points within the interval are equally likely. This assumption may impact every aspect of the interpretation of the plot, especially with regard to differences between individual studies and between-study heterogeneity.

1.2. Design of rainforest plots

Rainforest plots (see Fig. 1C for an example) were designed as an alternative to conventional forest plots. Although the idea in this graph type is to address all three of the aforementioned criticisms, they place a strong focus on the notion of uncertainty and the display of uncertainty through the use of a combination of raindrop plots (Barrowman and Myers, 2003) and density strips (Jackson, 2008).

Raindrop plots (Barrowman and Myers, 2003), originally intended to display collections of likelihoods, and shading (Jackson, 2008) have been suggested as an alternative design to overcome the third shortcoming, namely, the issue that viewers possibly ignore distribution information and therefore misinterpret the role of the confidence intervals and the uncertainty they represent. In this graph type, a likelihood curve for a chosen confidence interval (e.g., 95%) is plotted and then mirrored, producing a shape reminiscent of a raindrop. These graphs allow the viewer to compare several likelihoods directly and to detect changes in size or possible asymmetries while providing clear distribution information and making explicit the likelihood variation within the entire length of the confidence interval.

A second possibility for conveying the idea of statistical uncertainty in one dimension are so-called density strips (Jackson, 2008) in which confidence intervals are displayed with shading, i.e., the center of the interval is saturated in color, whereas this color fades and turns to white moving along the confidence interval toward the edges. This display type makes excellent use of the very intuitive notion that if something is harder to discern, it probably is not as likely.

The novel design of rainforest plots is based on a number of theoretical assumptions and addresses all the mentioned criticisms of the conventional design: (1) individual effects are clearly marked with white ticks of the same length and thickness for all effects, (2) individual confidence intervals are clearly marked with white lines, (3) uncertainty is displayed using deliberate redundancy, i.e., through the use of raindrops and shading, and (4) individual study weights are displayed using deliberate redundancy, i.e., height and color saturation of the raindrops.

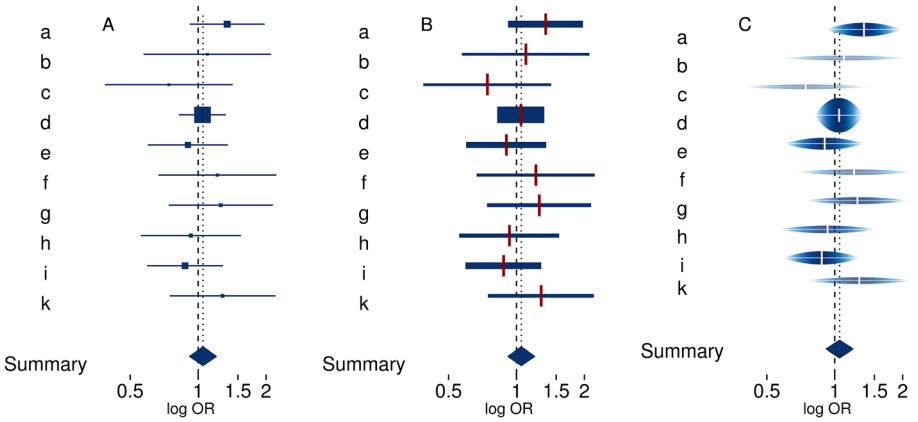

Figure 1. Examples of different displays of forest plots used in the study. Note. (A) Conventional forest plot. Squares represent effect estimates of individual studies with their 95% confidence intervals with size of squares proportional to the weight assigned to the study in the meta-analysis. The diamond represents the overall result and 95% confidence interval of the fixed-effect meta-analysis. (B) Thick forest plot. Ticks represent effect estimates of individual studies with their 95% confidence intervals with line width proportional to the weight assigned to the study in the meta-analysis. The diamond represents the overall result and 95% confidence interval of the fixed-effect meta-analysis. (C) Rainforest plot. Ticks represent effect estimates of individual studies with their 95% confidence intervals. Raindrops and shading represent the probability density for a probability of 0.95 with height of raindrop and color saturation proportional to the weight assigned to the study in the meta-analysis. The diamond represents the overall result and 95% confidence interval of the fixed-effect meta-analysis.

The name was chosen in order to reflect all the concepts on which this graph type is based. *Rainforest plot* results from a direct combination of *forest plot* and *raindrop plot*, whereas the shading of the density strips evokes the mental image of fog over a dense rainforest, such as is often found in cloud forests.

1.3. Design of thick forest plots

The name *thick forest plot* (see Fig. 1B for an example) alludes to the point that in this type of display confidence intervals are drawn with line width proportional to study weight. This novel graph type strikes a balance between the conventional design and the rainforest plot in theoretical terms. It was designed to address the first two criticisms, i.e., the following facts: (1) small studies may attract much visual attention because of the length of their confidence intervals and (2) individual effects of studies with large weights may be hard to discern because of the size of the boxes. In this type of graph, the line width of the confidence intervals of the individual studies is proportional to the weight assigned to the study in the meta-analysis to rectify the potential problem that small studies receive an undue amount of visual attention. Furthermore, individual effect estimates were clearly marked with red ticks, which are of the same thickness and length for all included studies. That is, this type of display largely corresponds to the conventional forest plot, but the line width of the confidence intervals varies with the assigned weights. In this regard, the relationship between conventional and thick forest plots is analogous to the relationship between a scatter plot and a bubble plot. The name thick forest plot was chosen to reflect the close relationship to the conventional display while alluding to the line width of the confidence intervals. On a meta-level, the name suggests the compacting and concentration of information while evoking the picture of a dense forest.

1.4. Aims of the present study

Although a number of new designs have been proposed in the last years, they remain relatively unknown and have not found their way into practice yet (Schild and Voracek, 2013). Furthermore, it is unclear whether those alternatives are effective and if they ease interpretation of forest plots. As a matter of fact, no research exists to suggest that the aforementioned criticisms are indeed problematic. Although these issues have been discussed in the literature, no large-scale studies have been published yet, which clarify if and to which extent the common design of forest plots abets misinterpretation by the viewers. To that end, two novel variants of forest plots were designed and evaluated in an online statistical cognition experiment.

A recent systematic review on graph use in meta-analyses (Schild and Voracek, 2013) revealed large differences between disciplines. Although forest plots are prominent in medical meta-analyses, they are rarely found in psychology journals and altogether absent in meta-analyses in business. It appears therefore plausible that such differences may also exist concerning the interpretation of forest plots. Researchers working in disciplines in which forest plots are less prevalent may show more interpretational difficulties. Moreover, a researchers' experience with the preparation and publication of meta-analyses may also play an important role.

The aim of the present study was to evaluate the interpretation of forest plots under different scenarios in order to determine the following: (1) whether researchers correctly interpret the graphs, (2) there are systematic differences between scientific disciplines, (3) the aforementioned criticisms are a threat to the correct interpretation, and (4) whether alternative displays may promote ease of interpretation.

2. Materials and methods

2.1. Study design

The study was designed as a Graeco-Latin square (see Bailey, 2008, pp. 157–167 for more details). This design allows us to study the effect of one treatment factor (type of forest plot) and several blocking factors (heterogeneity between studies, statistical significance, and ratio of weights assigned to the studies). The major advantage of this design is, first, that the blocking factors can be separated and, second, that the experiment can be run economically and efficiently with a small number of stimuli. In this type of design, interactions between the treatment factor and the blocking factors cannot be tested; however, in this experiment, no interactions were assumed.

2.2. Materials

The study was designed as a Graeco-Latin square (see Table 1 for an overview and Supplement 1 for details on the construction of the design); therefore, meta-analytic data for this study had to be simulated in order to meet the constraints of the experimental design. Simulations were conducted according to well-established procedures (Peters *et al*., 2006; Bax *et al*., 2009) using the open-source software R 2.15.2 (R Development Core Team, 2013), and reproducible example R code can be found in Supplement 2. The upper part of Table 2 provides an overview of the full specifications used in the simulations, which were determined so as to obtain reproducible datasets meeting the prespecified criteria of the experimental design. The lower part provides the resulting characteristics of the final datasets, which conform to the theoretically predetermined characteristics specified in the study design.

In words, nine different reproducible datasets, each consisting of ten studies, were sampled as follows. First, a baseline risk between 0.2 and 0.4 was randomly sampled from a uniform distributions. Second, *odds ratios* were sampled from a lognormal distribution. The parameters of this lognormal distribution were systematically varied as to achieve the theoretically predetermined effects (clearly significant, clearly nonsignificant, or barely (non) significant), and third, sizes of study arms were sampled from a lognormal distribution. Variances were systematically varied in order to achieve the theoretically predetermined differences in weights between studies.

In the first step, meta-analyses with these datasets were conducted using the R package metafor (Viechtbauer, 2010). Even though heterogeneity was substantial in some of the analyses (as per design), all analyses were conducted using a fixed-effect model regardless of I^2 to keep the design sufficiently simple, as the usage of both random-effects and fixed-effect models would have necessitated the inclusion of another factor in the design leading to an exponential increase in required stimuli and thus completion time. In a second step, graphs (see Supplement 2 for an example R code) were prepared using the R package ggplot2 (Wickham, 2009). All graphs were produced in a resolution of 480×480 pixels.

Three conventional forest plots and three graphs each for the two novel groups – thick forest plots and rainforest plots – were prepared under a number of scenarios regarding statistical significance, between-study heterogeneity, and ratio of weights assigned to the individual studies (see Table 1 for an overview and Supplement 2 for reproducible R code). Full specifications for each graph and characteristics of the simulated data are presented in Table 2. An explanation on how to read and interpret the graph was placed prominently alongside the stimuli (see figure notes for Fig. 1).

The survey comprised of four parts (see Table 3 for the complete survey). First, participants were asked to provide demographic data. Second, meta-analysis experience was assessed. Third, participants rated one graph

Table 1. Graeco-Latin square design showing scenarios for each graph

	Low heterogeneity	Moderate heterogeneity	High heterogeneity
Conventional forest plot	Scenario 1 ± sig Weights*	Scenario 2 + sig Weights**	Scenario 3 − sig Weights***
Thick forest plot	Scenario 4 + sig Weights***	Scenario 5 − sig Weights*	Scenario 6 ± sig Weights**
Rainforest plot	Scenario 7 − sig Weights**	Scenario 8 ± sig Weights***	Scenario 9 + sig Weights*

−, clearly nonsignificant; ±, marginally (non)significant; +, clearly significant; sig, significant.
*Low weights.
**Moderate weights.
***Extreme weights.

Table 2. Characteristics of simulated data and specifications for graphs used in the study

			Conventional forest plot			Thick forest plot			Rainforest plot		
Parameter	Distribution of parameter		Scenario 1	Scenario 2	Scenario 3	Scenario 4	Scenario 5	Scenario 6	Scenario 7	Scenario 8	Scenario 9
Baseline risk	Uniform	Min	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2	0.2
		Max	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4	0.4
OR/het	Lognormal	μ	0	0.5	0	0.5	0	0	0	0	−0.5
		τ	0.17	0.25	0.4	0.17	0.25	0.4	0.17	0.25	0.4
Study arm	Lognormal	μ	5	5	5	5	5	5	5	5	5
size/weights		τ	0.3	0.7	1.3	1.3	0.3	0.7	0.7	1.3	0.3
No. of	—		10	10	10	10	10	10	10	10	10
studies model	—		FE	FE	FE	FE	FE	FE	FE	FE	FE
$\sum$ of k	—		K: 7	K: 5	K: 2	K: 2	K: 7	K: 5	K: 5	K: 2	K: 7
largest weights			77.41	79.78	70.64	78.26	74.98	75.72	73.83	71.24	79.72
I^2 (%)	—		0.79	60.02	89.57	0	46.18	79.35	0	67.44	92.02
OR [95% CI]	—		1.16	1.75	0.98	1.88	1.02	1.08	1.05	0.90	0.42
			[0.99, 1.35]	[1.56, 1.98]	[0.89, 1.08]	[1.76, 2.07]	[0.86, 1.21]	[0.96, 1.22]	[0.92, 1.20]	[0.83, 0.98]	[0.36, 0.49]

FE, fixed effect; het, heterogeneity; *OR*, odds ratio.

Table 3. Overview of complete survey

Demographics
Please indicate your gender.
Please indicate your academic degree(s).
Please indicate the academic discipline(s) in which the above qualification(s) was/were awarded.
Please indicate the country of your academic affiliation.
MA experience
Please indicate the number of meta-analyses you have conducted.
Please indicate which software you have used for meta-analysis.
Please indicate which software you have used to produce meta-analysis graphs.
Please indicate the number of meta-analyses you have published. [(a) as first author, (b) in other author position]
Please indicate the number of journal articles on development of meta-analytical methods you have published. [(a) as first author, (b) in other author position]
Please indicate the number of other texts relevant to meta-analysis (e.g., book chapters, letters, or editorials) you have published. [(a) as first author, (b) in other author position]
Aesthetics and ease of perception (three times: random combination including each graph type)
This type of graph is easy to perceive. (0–100)
I have seen this type of graph before. (y/n)
This type of graph is aesthetically pleasing. (0–100)
This type of graph is easy to understand. (0–100)
This type of graph is intuitive. (0–100)
Forest plots (nine times: random sequence of all nine graphs)
All values within the individual confidence intervals are equally likely. (y/n)
Please estimate the amount of heterogeneity (I^2). (0–100)
This meta-analysis shows statistically significant heterogeneity. (y/n)
The overall result of the analysis is statistically significant. (y/n)
Please indicate the number of studies which show a statistically significant effect.
Please indicate the number of studies which differ significantly from the overall effect.
Please indicate the study with the lowest weight.
Please indicate the study with the highest weight.
There are significant outliers. (y: … number)

from each group on aesthetic properties, ease of perception, and ease of interpretation. For this section, all possible combinations of graphs (any one graph from each group of displays, Supplement 1) were included in the study ($k = 54$). One unique combination was randomly presented to each participant under the constraint of a uniform distribution, i.e., this part of the survey was designed as an urn draw without replacement. Therefore, the first 54 participants each drew one random combination from the urn until it was emptied. It was then again replenished with the full set of combinations, i.e., it was emptied for a second time after 108 participants. Therefore, all combinations were presented equally often. Fourth, all nine graphs were presented, and respondents answered questions regarding the interpretation of the graphs. Out of the total number of permutations ($9! = 362\,880$), 100 were randomly chosen. Each participant was then presented with one random sequence again under the constraint of a uniform distribution of the total of 100 sequences.

2.3. *Participants*

Participants were identified in a systematic and exhaustive manner (see Fig. 2 for an overview) aimed at including a broad sample of researchers with varying experience in meta-analysis. Potential participants' e-mail addresses were retrieved from Scopus, i.e., all researchers who had acted as corresponding authors for a meta-analysis published in 2012 or early 2013 were eligible for participation. The database Scopus was given preference over ISI Web of Science, which lists only published papers and articles from journals indexed in ISI, whereas Scopus further contains in-press articles making it the largest and most up to date source of peer-reviewed citations with the broadest journal coverage. A search for articles and reviews published from 2012 up until present (end of January 2013) with the string *meta-anal* * in title was conducted. The search yielded a total of 5025 meta-analyses published in 2012 or early 2013. For 206 records, no e-mail address was provided (4.1%), and 741 e-mail addresses (14.7%) were duplicated. After exclusion of duplicates and entries without e-mail addresses, a list of 4078 persons eligible for participation was obtained.

Out of 4078 invited participants, 611 logged in to the online survey (14.9%). Of those, 279 persons (107 women) completed the whole questionnaire, which corresponds to a rate of 6.8%. The sample comprised participants from 36 countries with the USA ($n = 45$), China ($n = 36$), and UK and Ireland ($n = 35$) as the most frequent ones. The majority of participants (78.1%) indicated that they held at least a PhD, MD, or equivalent qualification. Most of the respondents had obtained their degrees in the medical field ($n = 187$, e.g., medicine), followed by social and

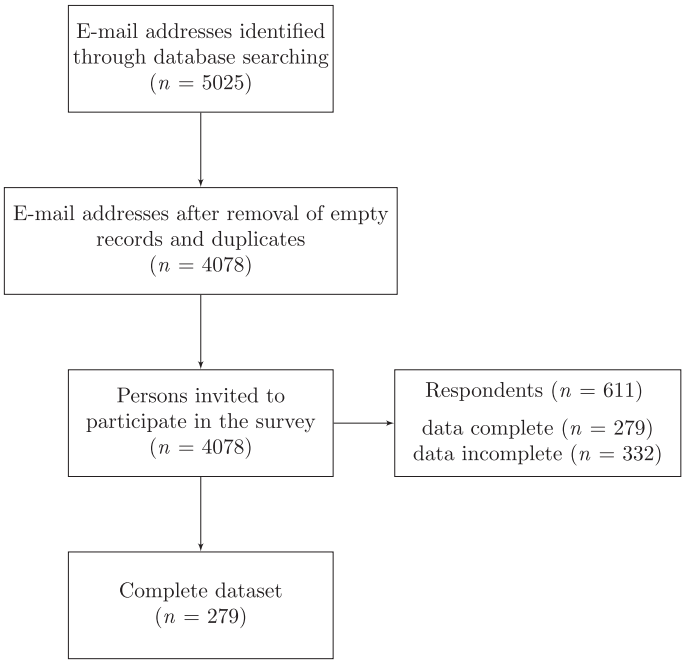

Figure 2. Flow chart of the process of data acquisition.

life sciences ($n = 39$, e.g., psychology and sociology), technical disciplines ($n = 35$, e.g., statistics and mathematics), and disciplines pertaining to the field of biology ($n = 17$, e.g., zoology and wildlife sciences). About a third ($n = 97$) of the respondents indicated more than one discipline, which, however, usually fell in the same category as the first one, which is why only the first indicated discipline was used for further analysis. Nearly all participants ($n = 277$) reported having conducted at least one meta-analysis, and the median number of conducted meta-analyses was 4 [interquartile range (IQR = 8)]. Only two participants (0.72%) indicated that they had never conducted a meta-analysis.

More than 80% ($n = 224$) of the respondents had published at least one meta-analysis as first author (*Md* number of meta-analyses published as first author = 1, IQR = 2), and 102 respondents had published at least one meta-analysis in other author positions (*Md* number of meta-analyses published in other author position = 1, IQR = 5). Participants reported having used a variety of software packages to conduct meta-analyses and to produce corresponding graphs (Table 4).

2.4. Noncompleters

Out of 611 persons who logged in to the online survey, 332 (108 women) did not complete the whole questionnaire. Possible differences between completers and noncompleters were assessed using χ^2-tests, none of which showed significant results. The two groups did not differ regarding education with 74.4% of noncompleters having obtained a PhD or equivalent χ^2 (1, $N = 611$) = 0.97, $p = 0.33$. No differences between disciplines were observed (medical: $n = 185$, social: n = 59, technical: $n = 25$, biology: $n = 22$) χ^2 (4, $N = 611$) = 7.76, $p = 0.1$. Meta-analysis experience was comparable between completers and noncompleters χ^2 (1, $N = 611$) = 1.22, $p = 0.27$. Noncompleters came from 46 countries with the USA ($n = 66$), China ($n = 43$), and UK and Ireland ($n = 41$) as the most prominent groups.

Table 4. Overview of software used to conduct meta-analysis and produce meta-analysis graphs

Software	Analysis	Graphs
RevMan	140	143
Stata	100	87
CMA	54	42
R	58	53
SPSS	35	15
SAS	30	14
MIX	7	5
NCSS	2	1
other	56	66

2.5. Procedure

Participants were invited via e-mail to complete an online survey prepared with SoSciSurvey (Leiner, 2012) and accessible via www.soscisurvey.de for 2 weeks. After 1 week, a reminder was sent out to all participants. Participation was voluntary, and respondents were informed that the data were anonymous and used for research purposes only.

2.6. Statistical analyses

Only complete datasets (45.6% of the total number of persons who logged into the survey) were used for the analyses (see Supplement 3 for the complete dataset and Supplement 4 for variable names and corresponding explanations). Completion times and aesthetic and perceptual criteria (rated on a scale from 0 to 100) were analyzed with ANOVA. Then, a score of the number of correct answers per graph (max score = 7) and for the whole questionnaire (max score = 63) was calculated for every respondent. All four factors (heterogeneity, distribution of weights, type of graph, and overall (non)significance) were entered into regression models to predict the following: (1) the overall score and (2) estimates of I^2 using dummy coding for all categorical predictors. Variance inflation factors (VIFs) were calculated as regression diagnostics to check for possible multicollinearity. While there is some discussion in the literature over which cutoff values for VIF should be used (Craney and Surles, 2002; O'Brien, 2007), the maximum VIF in our analyses was 2.05, which is well beyond any of the discussed critical values. Therefore, the validity of the regression analyses is not affected by multicollinearity. All analyses were conducted using the open-source software R 2.15.2 (R Development Core Team, 2013), and graphs were produced with the R package ggplot2 (Wickham, 2009).

3. Results

Mean completion time of the whole questionnaire was 23.2 min ($SD = 5.9$). There were significant differences in completion time between the different graphs, $F(8, 1899) = 4.23$, $p < 0.05$, and the three types, $F(2, 1905) = 4.33$, $p = 0.05$. Mean completion times per page (i.e., viewing one graph and answering eight questions) ranged from 1.6 min (graph 1, conventional) to 2 min (graph 6, thick forest plot).

Regarding aesthetic and perceptual criteria, no significant differences between the three graph types were observed, all F-values (1, 837) <1.71, all p-values >0.19, and participants rated them as equally well suited to display results from meta-analysis and found them equally intuitive to understand, all F-values (1, 837) <3.64, all p-values >0.05.

On average, participants answered 41.54 questions correctly ($SD = 7.56$). All factors included in the design (type, heterogeneity, significance, and weights) were shown to be significant predictors of the total score (see Table 5 for full model specifications). Figure 3 provides a graphical overview of the frequency distribution of the scores per graph and reveals that the overall score was highest for rainforest plots, followed by thick forest plots. Overall scores were lowest for conventional forest plots.

Estimates for I^2 ranged from 0 to 100 for all graphs. In the case of low heterogeneity, I^2 was overestimated. In the cases of moderate and high heterogeneity, estimates of I^2 were generally too low (Fig. 4). Both extreme and moderate weight differences were positive predictors of the estimates. Graphs with a clearly significant summary effect were negative predictors of estimated heterogeneity, as was the use of thick forest plots (see Table 6 for full model specifications).

Nearly two thirds ($n = 177$) of the participants indicated at least once that they believed that all values within the individual confidence intervals were equally likely. The wrong answer was given significantly more often when confronted with conventional or thick forest plots as compared with rainforest plots, $F(2, 2508) = 8.76$, $p < 0.001$.

Heterogeneity, the distribution of weights, the use of rainforest plots, and nonsignificant results are all positive predictors of the number of outliers indicated by the participants (see Table 6 for full model specifications). The number of indicated outliers ranged between 0 and 9 with a median of 2 (IQR = 2).

4. Discussion

To our knowledge, the current study presents the first account of a statistical cognition experiment with regard to meta-analysis graphs. Not only is the present study the first one to examine the interpretation of forest plots, but also two novel displays constructed accordant to theoretical criticisms of conventional forest plots were presented as well.

Overall, the participants' scores were moderate indicating that the correct and exhaustive appreciation of forest plots is not necessarily a straightforward process. Meta-analytic experience as measured in the number of conducted and published meta-analyses seems to have only a minor impact. No sex differences were observed, and participants' scores did not differ according to their scientific discipline with the notable exception of technical subjects.

Table 5. Predictors of the total score

Predictor	β	*SE*	95% CI
		Model 1 ($F_{15,\ 2439}$)	
Rainforest plot	−0.39***	0.05	[0.3, 0.48]
Thick forest plot	0.34***	0.05	[0.24, 0.43]
Heterogeneity	−0.001*	0.0001	[−0.002, −0.0003]
Significant	−0.31***	0.05	[−0.4, −0.22]
Not significant	−0.52***	0.05	[−0.61, −0.43]
Extreme weight	0.21***	0.05	[0.12, 0.3]
Moderate weight	−0.26***	0.05	[−0.25, −0.17]
Discipline biological	0.32	0.32	[−0,3, −0.17]
Discipline medical	0.29	0.31	[−0.32, 0.89]
Discipline social	0.16	0.31	[−0.45, 0.76]
Discipline technical	0.79*	0.31	[0.18, 1.04]
Sex: male	0.02	0.04	[−0.06, 0.09]
No. of MA conducted	−0.0005	0.001	[−0.002, 0.0008]
No. of MA published as first author	0.001*	0.002	[0.001, 0.01]
No. of MA published in other author position	0.002	0.002	[−0.003, 0.006]
R^2		0.15	
Adjusted R^2		0.15	
F		29.37***	
Max VIF		2.05	

β, standardized regression coefficient; SE, standard error of β; max VIF, largest variance inflation factor.
*$p < 0.05$;
**$p < 0.01$;
***$p < 0.001$.

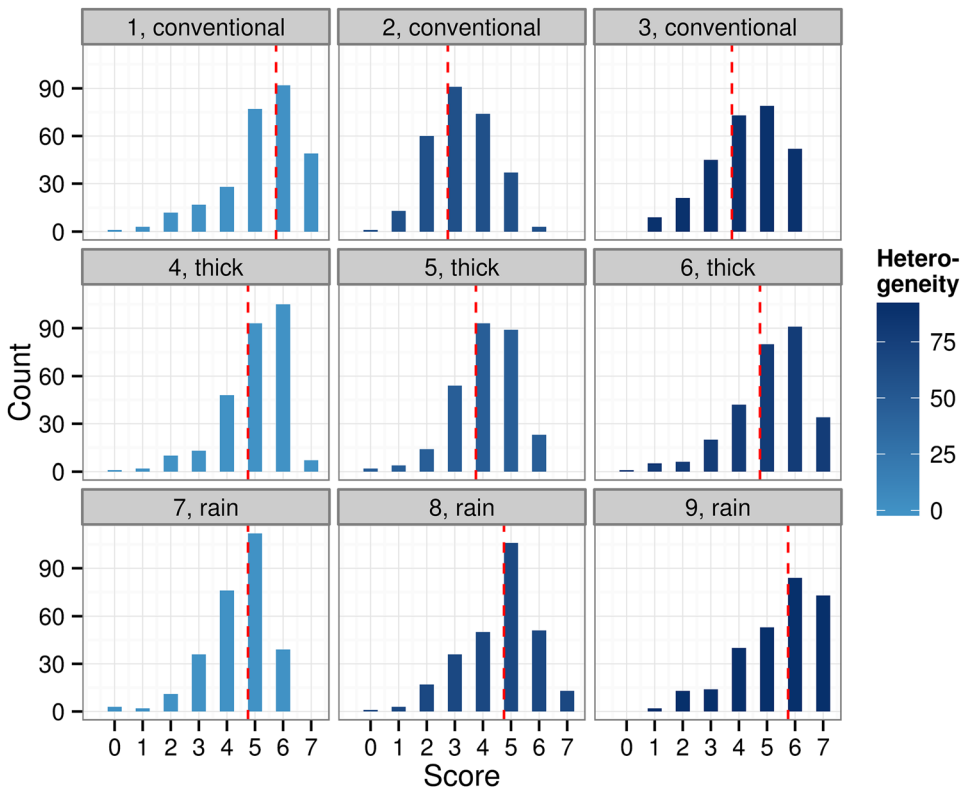

Figure 3. Frequency distributions of correct answers per graph and graph type. Note: Red dashed line indicates median number of correct answers.

Interestingly, there were no differences in aesthetic and perceptual judgements between the three graph types. However, both the thick and the rainforest plot were associated with slightly longer viewing times. Keeping in mind that these two displays are novel, not established yet, and therefore new to the viewers, this does not seem

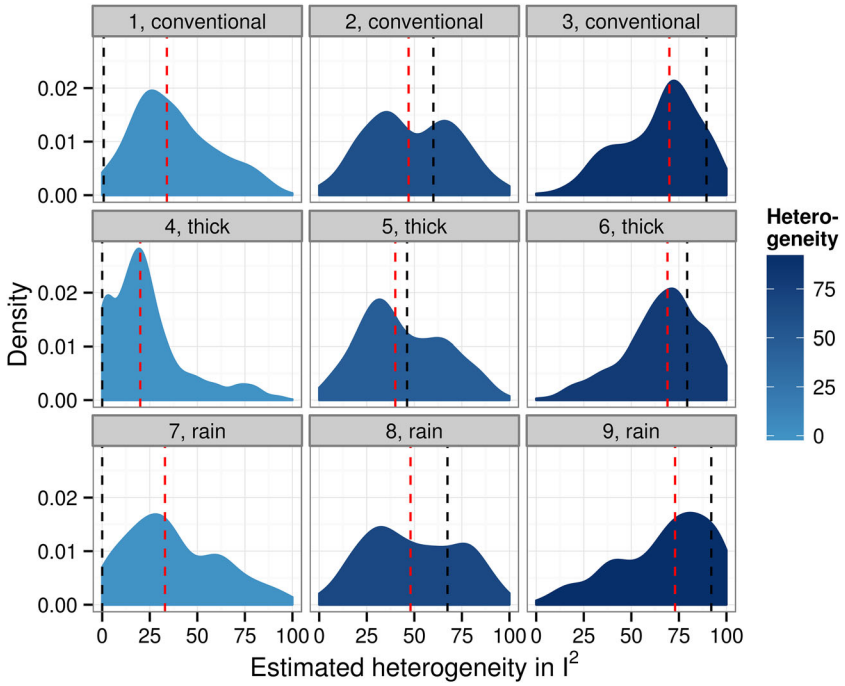

Figure 4. Distributions of heterogeneity estimates in I^2 per graph and graph type. Note. Black dashed line indicates the calculated amount of heterogeneity (I^2). Red dashed line indicates median of estimated heterogeneity.

Table 6. Predictors of estimated heterogeneity in I^2 and estimated number of outliers

Predictor	β	*SE*	95% CI
		Model heterogeneity ($F_{7,\ 961}$)	
Rainforest plot	−0.01	0.04	[−0.09, 0.074]
Thick forest plot	−0.10*	0.04	[−0.18, −0.02]
Heterogeneity	−0.51	0.02	[0.47, 0.54]
Significant	−0.2***	0.04	[−0.28, −0.11]
Not significant	−0.04	0.04	[−0.12, 0.04]
Extreme weights	0.23***	0.04	[0.15, 0.31]
Moderate weights	0.28***	0.04	[0.19, 0.36]
R^2		0.27	
F		135.3***	
Max VIF		1.03	
		Model outlier ($F_{7,\ 961}$)	
Rainforest plot	0.2*	0.1	[0.01, 0.39]
Thick forest plot	−0.146	0.09	[−0.33, 0.04]
Heterogeneity	0.01***	0.001	[0.006, 0.01]
Significant	0.2	0.1	[0.015, 0.39]
Not significant	0.16*	0.09	[−0.02, 0.34]
Extreme weights	0.35***	0.09	[0.17, 0.53]
Moderate weights	0.33**	0.1	[0.12, 0.53]
R^2		0.12	
F		18.94***	
Max VIF		1.31	

β, standardized regression coefficient; SE, standard error of β; max VIF, largest variance inflation factor.
*$p < 0.05$;
**$p < 0.01$;
***$p < 0.001$.

surprising. Moreover, completion times per page differed only by a maximum of 35 s overall; therefore, it may be concluded that the viewing times for the two novel displays may be just as economic. Furthermore, participants showed higher overall scores in the interpretation of the rainforest plots and thick forest plots. Although it cannot be ruled out

that the higher scores are associated with higher attentiveness because of the novelty of the displays, they also support the hypothesis that the interpretation of conventional forest plots may be impaired by cursory viewing behavior.

Both the rainforest plot and the thick forest plot positively predicted the total score indicating that participants answered more questions correctly when confronted with these types of displays. Heterogeneity, although significant in the model, appears to play a minor role. Plots with clearly significant or clearly nonsignificant summary results were associated with lower scores, which may be rooted in cursory viewing behavior in allegedly clear cases. Moderate differences between study weights were shown to be a negative predictor, whereas extreme weight differences positively predicted the score. This suggests that information about study weight is difficult to deduce when the weights are more evenly distributed.

Estimates of heterogeneity draw an interesting picture. Not only does this appear to be a difficult task, as suggested by the fact that estimates for all graphs covered the complete range, but also in the case of moderate or large heterogeneity estimates are too low, whereas I^2 is overestimated in the case of low heterogeneity. When taking a closer look at the median estimates per graph, it becomes apparent that they match the classic benchmarks for low ($I^2 \leq 25$), medium ($25 \geq I^2 \leq 75$), and high ($I^2 \geq 75$) values of I^2 (Huedo-Medina *et al*., 2006). This observation suggests that viewers make categorical judgements about between-study heterogeneity and then choose an estimate close to the respective benchmark.

Furthermore, estimates of heterogeneity are influenced by the distribution of the study weights. When differences among weights are moderate or extreme, heterogeneity is overestimated. This result indicates that weight information is used inappropriately in judgements about between-study heterogeneity. This effect may possibly be counterbalanced when weight information is displayed differently, as suggests the fact that heterogeneity estimates were more precise in thick forest plots.

Interestingly, a nonsignificant summary effect is a negative predictor of estimated heterogeneity. It appears that there is a tendency toward a confirmation bias (Nickerson, 1998), i.e., viewers automatically assume that the overall effect should be statistically significant and therefore neglect variation between the individual effect sizes. If, however, the pooled summary is not statistically significant, heterogeneity is overestimated in the sense of confirmatory information processing.

The same process may be at work in considerations about outliers. When the overall result is clearly statistically significant, more outliers are identified. It may be that the impact of studies deviating from the overall summary is perceived as more extreme in the case of a significant pooled effect.

When asked to indicate the study with the lowest or highest weight, mistakes were more common in the conventional display, and in most cases, the study right below or above the correct answer was indicated by the participants. It appears likely that this is due both to cursory viewing and the fact that the display of the individual effects is less visually striking than in thick forest plots and rainforest plots. It may be helpful, in general, to provide a light grid in the background of the plot to facilitate the mapping of effects to the study identifier.

None of the participants answered all questions correctly. Especially the concept that values toward the center of the confidence interval are more likely than values toward its boundaries appears to be unclear to many viewers. This concept seems to be more intuitive or more easily visually accessible in thick forest plots and rainforest plots. Furthermore, estimates of heterogeneity were more accurate for these two types suggesting that salience and clear visibility of confidence intervals are beneficial in judgements about between-study heterogeneity. These facts suggest that, optimally, confidence intervals should be paired with a visual cue about the distribution. In cases where this might not be possible, using thick forest plots or increasing the thickness of the lines in conventional forest plots are easy-to-realize and effective ways to draw more attention to this detail.

The current study may be limited by the use of stripped-down versions of forest plots, which include only the graphical information. In general, forest plots are textbook examples of table graphs, i.e., they are a combination of a table and a graph, and they should be accompanied by further numerical information (e.g., study weights, confidence intervals, and heterogeneity estimates; Anzures-Cabrera and Higgins, 2010). Although a comparison between stripped-down versions of forest plots and forest plots with numerical information would have been theoretically possible in this design, it would have led to an undue increase in required stimuli and therefore completion time for the survey. Furthermore, all meta-analytic samples were limited to ten included studies, and the observed effects may not hold for larger meta-analyses. The effects of omitting versus including numerical information in forest plots and using larger samples would be interesting routes for further research.

Statistical graphs in general and forest plots in particular provide information on a number of different aspects of the underlying analysis, which makes viewing and interpreting them a holistic task. As such, it may be difficult to capture the complete essence of the graph in individual questions. However, the systematic experimental study of the interpretation of the plots requires precise questions. Therefore, although asking for correct answers may not be the ideal way to get to the heart of the forest plot as a whole, it is a suitable approach to detect systematic errors in interpretation that will provide a solid base for visual enhancements and adaptations.

Participants in this study were identified through Scopus, i.e., they had acted as corresponding author for at least one meta-analysis published between 2012 and early 2013. Hence, the sample was limited to researchers who had been strongly involved in conducting at least one meta-analysis. This systematic approach of participant selection has various strengths, namely: (1) it may easily be repeated in future research and therefore enables direct comparisons, (2) selection bias could be minimized, (3) participants had been strongly involved in the

publication of a meta-analysis according to modern standards (only meta-analyses published very recently were included), and (4) a large and scientifically diverse sample could be obtained. Furthermore, this group should be adept at the statistical and interpretational aspects of meta-analytic evidence, as they are both authors and recipients of meta-analyses. However, generalizability of the results may be limited with regard to other groups, such as recipients of meta-analyses without specific statistical knowledge. In addition, the groups for the different disciplines were not of the same size, which may limit the interpretation of differences between groups. As the study design is completely flexible, it may easily be adapted, and future studies may focus on a more general audience, such as students, practitioners, policy makers, or any other stakeholders in the meta-analytic process.

Interactions between factors were not addressed in this study, as they cannot be studied in Graeco-Latin square designs, which may be a possible limitation of the study. However, it has to be admitted that such interactions are routinely neglected in natural settings. Assuming that there are interactions between, for example, the graph type *conventional forest plot* and the amount of between-study heterogeneity, the use of such graphs may be questionable altogether as long as systematic research on interaction effects is missing. Hence, it may be argued that the assumption of no interactions extends well beyond this report. Moreover, this is the first study to report systematic experimental research on the perception and interpretation of forest plots and variants thereof. In order to study a variety of factors while keeping the design economic, a full factorial design was not feasible. As the process of participant selection was conducted in a systematic and reproducible manner, further research may be conducted to elucidate possible interactions.

While the existence of carryover effects, i.e., effects of specific sequences, is theoretically conceivable, it is highly unlikely that such effects may have introduced a bias on the reported results. A large number of random sequences ($n = 100$) were administered in the study in order to counterbalance possible carryover effects across participants. For the same reason, i.e., a large number of random sequences and a small ratio of participants to sequences, a formal statistical assessment is not practicable in this case. Because of a lack of empirical evidence and the pioneering nature of this work, the option of counterbalancing carryover effects was preferred over the administration of a low number of sequences, in which case statistical assessments are feasible.

The present study revealed that there are systematic errors in the interpretation of forest plots, especially with regard to the concept of confidence intervals. Both thick and rainforest plots are aesthetically and perceptively well suited to display meta-analytic results. They allow for economic viewing while facilitating certain aspects of interpretation, such as considerations regarding the distribution of study weights and underlying heterogeneity between studies. Participants showed the highest scores in the evaluation of rainforest plots, followed by thick forest plots and conventional forest plots, respectively. Both novel graphs are suitable candidates for the display of meta-analytic information and are preferable over conventional forest plots. If, however, conventional forest plots are used, it is advisable to use background grids and thicker or more visually striking confidence intervals paired with distribution information in conventional displays.

Graph usage studies and statistical cognition experiments are scarce in general and limited mostly to very specific topics (e.g., judgements of point sizes; Chong and Treisman, 2003; Hurewitz *et al*., 2006). In the field of meta-analysis, no such studies have been published, yet. Because this area of research has received relatively little attention so far, general information on graph usage and interpretation is still rare. This means that vital opportunities for the identification of design and usage errors are passed by which hinders both the development of design and application guidelines and new graphical methods (Cleveland, 1984). The present results reveal how much information can be conveyed in one simple graph that can be perceived in a matter of seconds, making graphs powerful and adaptable tools in the communication of meta-analytic results.

Acknowledgements

We thank Nick Barrowman, PhD, for generously sharing his R code for the original raindrop plots (Barrowman and Myers, 2003).

References

Anzures-Cabrera J, Higgins JPT. 2010. Graphical displays for meta-analysis: an overview with suggestions for practice. *Research Synthesis Methods* **1**: 66–80.

Bailey RA. 2008. Design of Comparative Experiments, chap. More about Latin squares, 157–167. Cambridge University Press: Cambridge, UK.

Barrowman NJ, Myers RA. 2003. Raindrop plots: a new way to display collections of likelihoods and distributions. *American Statistician* **57**: 1–6.

Bax L, Ikeda N, Fukui N, Yaju Y, Tsuruta H, Moons KGM. 2009. More than numbers: the power of graphs in meta-analysis. *American Journal of Epidemiology* **169**: 249–255.

Chong SC, Treisman A. 2003. Representation of statistical properties. *Vision Research* **43**: 393–404.

Cleveland WS. 1984. Graphs in scientific publications. *American Statistician* **38**: 261–269.
Craney TA, Surles G J 2002. Model-dependent variance inflation factor cutoff values. *Quality Engineering* **14**: 391–403.
Huedo-Medina TB, Sanchez-Meca J, Marin-Martinez F, Botella J. 2006. Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? *Psychological Methods* **11**: 193–206.
Hurewitz F, Gelman R, Schnitzer B. 2006. Sometimes area counts more than number. *Proceedings of the National Academy of Sciences of the United States of America* **103**: 19599–19604.
Jackson CH. 2008. Displaying uncertainty with shading. *American Statistician* **62**: 1–8.
Leiner DJ 2012. SoSci survey (version 2.3.04) [computer software]. Available at: https://www.soscisurvey.de. Accessed July 31, 2014.
Lewis JA, Ellis SH. 1982. A statistical appraisal of post-infarction beta blocker trials. *Primary Cardiology* **S1**: 31–37.
Lewis S, Clarke M. 2001. Forest plots: trying to see the wood and the trees. *BMJ* **322**: 1479–1480.
Liberati A, Altman DG, Tetzlaff J, Mulrow C, Gøtzsche PC, Ioannidis JPA, Clarke M, Devereaux PJ, Kleijnen J, Moher D. 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. *PLoS Medicine* **7**: e1000100.
Moher D, Liberati A, Tetzlaff J, Altman D 2009. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *PLoS Medicine* **6**: e1000097.
Nickerson RS. 1998. Confirmation bias: a ubiquitous phenomenon in many guises. *Review of General Psychology* **2**: 175–220.
O'Brien RM. 2007. A caution regarding rules of thumb for variance inflation factors. *Quality & Quantity* **41**: 673–690.
Peters JL, Sutton AJ, Jones DR, Abrams KR, Rushton L. 2006. Comparison of two methods to detect publication bias in meta-analysis. *Journal of the American Medical Association* **295**: 676–680.
R Development Core Team. 2013. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available at: http://www.R-project.org. Accessed December 31, 2013.
Schild AHE, Voracek M. 2013. Less is less: a systematic review of graph use in meta-analyses. *Research Synthesis Methods*. Advance online publication. DOI: 10.1002/jrsm.1076
Schriger DL, Altmann DG, Vetter JA, Heafner T, Moher D. 2010. Forest plots in reports of systematic reviews: a cross-sectional study reviewing current practice. *International Journal of Clinical Epidemiology* **39**: 421–429.
Viechtbauer W. 2010. Conducting meta-analyses in R with the metafor package. *Journal of Statistical Software* **36**: 1–48.
Wickham H. 2009. ggplot2: Elegant Graphics for Data Analysis. Springer: New York.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.