Visual Inference for the Funnel Plot in Meta-Analysis

Michael Kossmeier, Ulrich S. Tran, and Martin Voracek

Department of Basic Psychological Research and Research Methods, Faculty of Psychology, University of Vienna, Austria

Abstract: The funnel plot is widely used in meta-analyses to assess potential publication bias. However, experimental evidence suggests that informal, mere visual, inspection of funnel plots is frequently prone to incorrect conclusions, and formal statistical tests (Egger regression and others) entirely focus on funnel plot asymmetry. We suggest using the visual inference framework with funnel plots routinely, including for didactic purposes. In this framework, the type I error is controlled by design, while the explorative, holistic, and open nature of visual graph inspection is preserved. Specifically, the funnel plot of the actually observed data is presented simultaneously, in a lineup, with null funnel plots showing data simulated under the null hypothesis. Only when the real data funnel plot is identifiable from all the funnel plots presented, funnel plot-based conclusions might be warranted. Software to implement visual funnel plot inference is provided via a tailored R function.

Keywords: funnel plot, meta-analysis, publication bias, small-study effects, visual inference

The funnel plot is a widely used diagnostic plot in meta-analysis to assess small-study effects and publication bias in particular (Light & Pillemer, 1984). It was one of the first genuine plots proposed to visualize meta-analytic data and, next to the forest plot (Lewis & Clarke, 2001), is the most iconic and popular display for this purpose (Schild & Voracek, 2013).

In essence, the funnel plot is a scatter plot of study effects on the abscissa and a study precision measure (preponderantly, the study standard error) on the ordinate (Sterne & Egger, 2001). Its main idea is that observed effects should scatter randomly and symmetrically around the meta-analytic summary effect. As smaller studies are displayed toward the bottom and higher effect size variability is expected for these, this gives rise to the characteristic shape of an inverted funnel for this graphical display.

Certain deviations from this expected funnel plot shape are commonly taken as suggestive for publication bias, although they similarly emerge via true effect heterogeneity or chance alone. In particular, a frequent observation is that smaller studies on average report larger effects. This so-called small-study effect, in turn, leads to asymmetric funnel plots (see Figure 1).

However, despite the popularity of the funnel plot, its suitability to detect publication bias has been questioned (Lau, Ioannidis, Terrin, Schmid, & Olkin, 2006). Indeed, empirical research suggests that subjective interpretations of whether publication bias is present or absent, based on mere visual funnel plot inspection, often times are wrong (Hunter et al., 2014; Simmonds, 2015; Tang & Liu, 2000; Terrin, Schmid, & Lau, 2005).

Formal statistical tests (e.g., Egger regression test, and others) based on funnel plot asymmetry are widely used to establish objectivity, while controlling for type I errors. All these tests have in common that they are based on funnel plot asymmetry quantified via the association of study effects with study standard errors (or respective functions of these). On the other hand, visual inspection of funnel plots allows incorporating a multitude of visually displayed statistical information in an exploratory fashion, in order to assess the presence and severity of publication bias. Important questions that can be addressed by visual funnel plot examinations include the following: Which role does statistical significance play (as indicated by significance-contour funnel plots)? Is there an abundance of just-significant studies? Is asymmetry driven by single-study outliers or clusters of studies?

As prominently outlined in the Cochrane Collaboration handbook for systematic reviews, formal tests for funnel plot asymmetry should never be interpreted in isolation, but rather always, and first of all, in the light of a visual inspection of the funnel plot (Sterne, Egger, & Moher, 2008, p. 317). Hence, visual inference might be the sought-after bridge between both worlds: it allows researchers to formally safeguard against type I errors, while still being in keeping with the more general diagnostic, open, and explorative nature of visual examination of statistical graphs.
Visual Inference for Statistical Graphs

Visual inference is a formal inferential framework (Buja et al., 2009) which allows researchers to test whether graphically displayed data support a hypothesis or not. The principal idea is that if a suitable statistical plot of actually observed data is visually distinguishable from plots of data simulated under the null hypothesis, then this constitutes evidence against the null hypothesis. In this framework, valid inferences are drawn via the so-called lineup protocol. That is, a lineup of diagnostic plots is constructed for the actually observed data and the null hypothesis a researcher wants to reject. The total lineup comprises $k$ plots, of which $k/C_0$ plots show data simulated under the null hypothesis, and wherein the single plot with the actually observed data is positioned randomly. This lineup then is inspected by a viewer unfamiliar with the actually observed data and their peculiarities with the aim to identify the real-data plot. In practice, the primary researcher and the viewer of the lineup often times will be the same person, which is valid as long as the primary researcher is unfamiliar with the shape of the real-data plot.

If the actually observed data in fact are realizations of the null hypothesis, the probability to identify the real-data plot and therefore to falsely reject the null hypothesis is $1/k$. Hence, the alpha level is controlled by design (i.e., the size of the lineup). A natural choice for the number of plots in a lineup therefore is 20, corresponding to the conventional alpha level (5%). Just as with conventional statistical tests, a test statistic (the actual plot) is compared to the null distribution (the plots showing data simulated under the null hypothesis). At the same time, visual inference differs from conventional inference, in that the test statistic is not compared to the entire null distribution, but rather to a finite number of realizations thereof (Majumder, Hofmann, & Cook, 2013).

Evaluations made by several independent viewers of a lineup, instead of by a single viewer, can be used for visual inference as well. This extension of the basic procedure has the potential to increase the power to correctly reject the null hypothesis. For either two, three, four, five, six, or seven viewers of the same lineup of 20 plots, it is sufficient that at least two viewers are able to identify the real-data plot to reject the null hypothesis with the alpha level of the procedure not exceeding 5% (for further details, see Majumder et al., 2013).

In recent studies visual inference has shown promise under scenarios of different statistical plots and data contexts (Chowdhury et al., 2015; Loy, Follett, & Hofmann, 2016; Loy, Hofmann, & Cook, 2017; Majumder et al., 2013).

Visual Inference for the Meta-Analytic Funnel Plot

We suggest evaluating meta-analytic funnel plots via visual inference for two main reasons. First, by controlling for type I errors, visual inference has the potential to increase the (often low) validity of conclusions based on mere visual
### Null-Plot Simulation for Visual Funnel-Plot Inference

Essential for visual inference is the null distribution used to simulate the data displayed in the null plots, as this directly corresponds to the null hypothesis one seeks to reject. For meta-analysis, natural choices are the fixed-effect model (FEM) and the random-effects model (REM).

The FEM assumes the observed effects $y_i$ can be modeled as

$$ y_i = \mu + u_i, \quad \text{with} \quad u_i \sim N(0, \sigma^2). $$

That is, the study effects $y_i$ are independent realizations of normal distributions with the same shared expected value $\mu$, but with study-specific variances $\sigma^2_i$, which are mainly due to different sample sizes. The FEM therefore assumes that differences between study effects entirely are due to sampling error.

The REM allows the modeling of additional (unsystematic) random variability between the study effects, which exceeds the amount of variability expected under the FEM. The REM assumes that the observed effects $y_i$ can be modeled as

$$ y_i = \mu + u_i + e_i, \quad \text{with} \quad u_i \sim N(0, \sigma^2) \quad \text{and} \quad e_i \sim N(0, \tau^2). $$

In the REM, the variance of each effect is $\sigma^2_i + \tau^2$, and therefore increased by the constant $\tau^2$, as compared to the FEM. Based on these models, two straightforward ways to construct null plots for visual funnel plot inference are as follows.

Given $n$ actually observed effects $y_{i,obs}$ with estimated standard errors $\hat{\sigma}_i$, the estimated meta-analytic summary effect $\hat{\mu}_{obs}$, and an optional estimate for the between-study variance $\tau^2_{obs}$, the effects displayed in each null plot are simulated using the following model:

$$ y_{i,simul} = \hat{\mu}_{obs} + u_i + e_i, \quad \text{with} \quad u_i \sim N(0, \hat{\sigma}^2_i) \quad \text{and} \quad e_i \sim N(0, \tau^2_{obs}). $$

That is, the effects in each null plot are randomly drawn from normal distributions with expected value equal to the actually observed meta-analytic summary effect $\hat{\mu}_{obs}$ and study-specific variances equal to the sum of the observed variance $\hat{\sigma}^2_i$, and the estimated between-study variance $\tau^2_{obs}$ from the actually observed data (REM). For the FEM, $\tau^2_{obs}$ is simply set to zero. The null dataset for one null plot is then given as the $n$ simulated effects $y_{i,simul}$ and the

---

**Table 1: Data**

<table>
<thead>
<tr>
<th>ID</th>
<th>d</th>
<th>se</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.32</td>
<td>0.48</td>
</tr>
<tr>
<td>2</td>
<td>1.01</td>
<td>0.47</td>
</tr>
<tr>
<td>3</td>
<td>1.13</td>
<td>0.25</td>
</tr>
<tr>
<td>4</td>
<td>0.69</td>
<td>0.32</td>
</tr>
<tr>
<td>5</td>
<td>0.78</td>
<td>0.30</td>
</tr>
</tbody>
</table>

**Table 2: Create lineup**

- Which funnel plot stands out?
  - Typical evaluations
    - Small-study effects discernible?
    - Role of statistical significance?
    - Abundance of just-significant study outcomes?
    - Conspicuous outliers or clustering of studies?
  - Real-data funnel plot identified?
  - Reject $H_0$
    - Yes
    - Retain $H_0$

![Figure 2. Visual inference testing procedure using the lineup protocol with funnel plots. Starting from the effect sizes and their standard errors actually observed in the meta-analysis (step 1), a lineup is constructed, showing the real-data funnel plot randomly positioned among null funnel plots (step 2). A viewer visually inspects the funnel plots in the lineup and picks the one that seems most noticeably or eye-catching (step 3). If the picked funnel plot indeed is the real-data funnel plot, the null hypothesis ($H_0$) used for null plot simulation is rejected (step 4).](https://econtent.hogrefe.com/doi/pdf/10.1027/2151-2604/a000358 - Thursday, May 28, 2020 10:19:31 AM - Universidade de Sao Paulo IP Address:143.107.196.25)

---

**Figure 2.** Visual inference testing procedure using the lineup protocol with funnel plots. Starting from the effect sizes and their standard errors actually observed in the meta-analysis (step 1), a lineup is constructed, showing the real-data funnel plot randomly positioned among null funnel plots (step 2). A viewer visually inspects the funnel plots in the lineup and picks the one that seems most noticeably or eye-catching (step 3). If the picked funnel plot indeed is the real-data funnel plot, the null hypothesis ($H_0$) used for null plot simulation is rejected (step 4).
initially observed corresponding standard errors $\sigma_i$. To emphasize, effect sizes are randomly drawn from a null distribution, with actually observed standard errors regarded as fixed.

Under both the FEM and REM scenarios, the null hypothesis tested against in visual inference is that study effect sizes are independent realizations of normal distributions with the same expected value, but different (nonrandom) variances. In the context of meta-analysis, null funnel plots simulated that way are well-behaved, symmetric random noise.

Which of the two models above should be used for null-plot simulation? In most cases, the REM is a suitable default choice. This allows researchers to exclude the alternative possibility, namely, that the reason for rejecting the null hypothesis via visual inference was solely due to an excess of unsystematic between-study variation in the real-data plot. An exception to this rule would be when (paralleling the Cochran $Q$ procedure as the conventional meta-analytic test) the excess of between-study variability itself is a target for visual inference. In this case, the FEM should be used for null-plot simulation. Finally, alternative models to simulate the data for the null funnel plots, including Bayesian models, may well be proposed and used.

Quite a number of variants of the classic meta-analytic funnel plot have been proposed, which differ in the statistical information conveyed and in their diagnostic purpose (Langan, Higgins, Gregory, & Sutton, 2012). These variants include different choices for the ordinate, the display of study subgroups, confidence contours, significance
contours, or the regression line from Egger’s test, and the Duval-Tweedie trim-and-fill method. The visual inference framework can be accommodated to include all these variants. As an example, Figure 4 shows a lineup for visual inference of a meta-analytic funnel plot, incorporating subgroups, and using data from a published meta-analysis (Pietschnig, Voracek, & Formann, 2010).

**Software for Visual Funnel-Plot Inference**

For meta-analytic practitioners, an important question is how to conveniently conduct visual funnel-plot inference. Within the statistical computing environment R (R Core Team, 2018), the package nullabor (Wickham, Chowdhury, & Cook, 2014) is available, which provides general-purpose functions to conduct visual inference with arbitrary graphical displays, including functionalities to reveal the position of the real-data funnel plot in the lineup only after inspecting a lineup. Building on this, we have developed and documented the R function funnelinf within the R package metaviz (Kossmeier, Tran, & Voracek, 2018) for specifically conducting visual funnel-plot inference. The funnelinf function provides tailored features, which currently include: (1) options for null-plot simulation under both FEM and REM meta-analysis; (2) subgroup analysis; (3) graphical options specific to the funnel plot (significance and confidence contours, and choice of the ordinate); and (4) additional options to display various statistical information (Egger’s regression line, and imputed studies by, as well...
as the adjusted summary effect from, the trim-and-fill method). For further details, example code, and example data we refer to the documentation of package metaviz

References


The magnitude likely effect of formal education in meta-analysis and research synthesis, by allowing students and users to collect experience with the manifold shapes and patterns appearing in funnel plots just by chance. For this specific purpose, plot lineups entirely comprised of null plots, also known as the Rorschach protocol of visual inference (Buja et al., 2009), might be used.

Software to conduct all these forms of visual inference with meta-analytic funnel plots is readily available in the form of a tailored function within R package metaviz (Kossmeier et al., 2018).

Conclusions and Implications

We propose to present, contemplate, and evaluate the funnel plot of the actually observed data simultaneously with null-hypothesis funnel plots. Only if the real-data funnel plot is identifiable from null-plots, the null hypothesis is formally rejected and conclusions based on visual inspection of the real-data funnel plot might be warranted. We suggest using visual funnel-plot inference routinely, as it is a convenient way to increase the validity of conclusions based on funnel plots by saving investigators from interpreting funnel-plot patterns which might be perfectly plausible by chance.

Empirical experiments are suited to examine the power of the procedure to reject the null hypothesis in different scenarios. Using datasets simulated under an alternative hypothesis of interest, the proportion of corresponding lineups leading to correctly rejecting the null hypothesis can be used as a direct estimate of the power of the lineup procedure (Majumder et al., 2013). Ideally, for this purpose larger numbers of independent viewers would work through funnel plot lineups from different experimental conditions.

A difficulty in recruiting larger numbers of viewers is the likely effect of formal education in meta-analysis and expertise with funnel plots in particular. The magnitude of this effect is unknown, but at least questions the use of online recruitment systems like Amazon’s Mechanical Turk, which has been regularly used in visual inference experiments in the past (e.g., Loy et al., 2016; Majumder et al., 2013). Hence, either innovative ways have to be found to recruit viewers with already existing funnel plot expertise or to train viewers unfamiliar with the funnel plot in an efficient way prior to the experiment.

Questions for future experimental research include:

What is the power of the procedure in different scenarios, for instance, for varying levels of publication bias or between-study heterogeneity? How does the procedure compare to conventional statistical tests for funnel plot asymmetry? Are there power differences to detect publication bias when using different graphical variants of the funnel plot for visual inference, for instance, by additionally showing Egger’s regression line? Which role do viewer characteristics and expertise with funnel plots play in successfully conducting visual inference with funnel plots? What is the power-wise benefit when basing the decision to reject the null hypothesis on more than one viewer’s evaluation per lineup? As a promising topic for future empirical inquiry, visual inference with funnel plots is yet at its beginning.

Visual funnel plot inference also holds potential to serve didactic purposes in meta-analysis and research synthesis, by allowing students and users to collect experience with the manifold shapes and patterns appearing in funnel plots just by chance. For this specific purpose, plot lineups entirely comprised of null plots, also known as the Rorschach protocol of visual inference (Buja et al., 2009), might be used.

Software to conduct all these forms of visual inference with meta-analytic funnel plots is readily available in the form of a tailored function within R package metaviz (Kossmeier et al., 2018).


History
Received February 28, 2018
Revision received September 27, 2018
Accepted October 22, 2018
Published online March 29, 2019

Michael Kossmeier
Department of Basic Psychological Research and Research Methods
Faculty of Psychology
University of Vienna
Liebiggasse 5
1010 Vienna
Austria
michael.kossmeier@univie.ac.at