

## Metanálises em Psicologia: Uma Introdução Conceitual e Prática

Marcos Felipe Rodrigues de Lima<sup>1</sup>

Luciano Grüdtner Buratto<sup>1</sup>

<sup>1</sup>Universidade de Brasília, Brasília, Distrito Federal, Brasil

### Resumo

Metanálise consiste em um conjunto de técnicas estatísticas que visa integrar os resultados de dois ou mais estudos primários. Ela permite produzir estimativas pontuais e intervalares de algum parâmetro populacional, geralmente uma medida de tamanho de efeito. Este artigo tem como objetivo apresentar conceitos fundamentais sobre metanálise e suas aplicações para psicólogos e estudantes de psicologia. O artigo: (1) introduz a lógica da metanálise, seus potenciais e as críticas a ela endereçadas; (2) apresenta dois modelos de metanálise comumente usados por pesquisadores; e (3) aborda dois tópicos importantes para a interpretação correta dos resultados: heterogeneidade e análise de subgrupos. Um exemplo fictício ilustra os conceitos ao longo do artigo. Os Materiais Suplementares contêm equações dos modelos apresentados no texto, resultados comentados de uma síntese metanalítica, código em linguagem R para reproduzir resultados e figuras desse artigo e uma breve lista comentada de fontes adicionais sobre metanálise.

*Palavras-chave:* metanálise, revisão sistemática, revisão de literatura, tamanho de efeito, gráfico de floresta

### Meta-Analyses in Psychology: A Conceptual and Hands-On Introduction

#### Abstract

Meta-analysis consists of a set of statistical techniques that aims to combine the results of two or more primary studies. It enables the calculation of point and interval estimates of some population parameter, usually a measure of effect size. The aim of this article is to introduce fundamental concepts of meta-analysis and its applications for psychologists and psychology students. The article: (1) introduces the logic of meta-analysis, its uses and common criticisms levied against it; (2) presents two computational models of meta-analysis commonly used by researchers; and (3) addresses two issues associated with the correct interpretation of results from meta-analyses: heterogeneity and subgroup analysis. A worked example illustrates the concepts throughout the article. The Supplementary Materials contain a worked example of the models presented in the text, a script in R language that allows the reader to reproduce the results, and a commented list of additional sources.

*Keywords:* meta-analysis, systematic review, literature review, effect size, forest plot

### Metanálisis en Psicología: Una Introducción Conceptual y Práctica

#### Resumen

El metanálisis consiste en un conjunto de técnicas estadísticas que tiene como objetivo integrar los resultados de dos o más estudios primarios. Permite producir estimaciones puntuales y de intervalo de algún parámetro de población, generalmente una medida del tamaño del efecto. Este artículo presenta conceptos fundamentales sobre el metanálisis y sus aplicaciones para psicólogos y estudiantes de psicología. El artículo: (1) introduce la lógica del metanálisis, sus potencialidades y las críticas que se le dirigen; (2) presenta dos modelos de metanálisis comúnmente utilizados por los investigadores; y (3) aborda dos temas importantes para la correcta interpretación de los resultados: heterogeneidad y análisis de subgrupos. Un ejemplo ficticio ilustra los conceptos a lo largo del artículo. Los Materiales Suplementarios contienen ecuaciones de los modelos presentados en el texto, resultados comentados de una síntesis metanalítica, código en el lenguaje R para reproducir los resultados y las figuras de este artículo, y una breve lista comentada de fuentes adicionales.

*Palabras clave:* metanálisis, revisión sistemática, revisión de la literatura, tamaño del efecto, diagrama de bosque

#### *Metanálises em Psicologia: Uma Introdução Conceitual e Prática*

Psicólogos precisam lidar com o fato de que os resultados de seus estudos são probabilísticos. Ainda que um experimento duplo cego apresente um efeito dose-dependente de um determinado alucinógeno (e.g., psilocibina) sobre o desempenho dos participantes em uma tarefa psicomotora (Barrett, Carbonaro, Hurwitz, Johnson, & Griffiths, 2018), não há garantias de que um

experimento subsequente apresentará um efeito similar. Recentemente, replicações pré-registradas e em larga escala de estudos de psicologia social e de psicologia cognitiva indicaram que os tamanhos de efeito dos estudos de replicação tiveram, em média, a metade do valor dos tamanhos de efeito dos estudos originais (Open Science Collaboration, 2015). Quando os resultados de estudos originais e de replicação divergem, possíveis

explicações para isso incluem diferenças metodológicas entre estudos, diferenças importantes nas características das amostras e erros dos Tipos I ou II (LeBel & Peters, 2011; Open Science Collaboration, 2015). Dado o caráter probabilístico do empreendimento científico, a interpretação dos resultados de um estudo deve ser feita à luz dos demais estudos sobre o mesmo fenômeno. Nesse sentido, a metanálise, entendida como um conjunto de técnicas estatísticas que visa integrar os resultados de dois ou mais estudos primários (Cumming & Calin-Jageman, 2017; Glass, 1976; Hedges & Olkin, 1985), é uma ferramenta poderosa, capaz de contextualizar o estado da ciência de uma área de pesquisa.

Na última década, uma série de artigos sobre análise quantitativa de dados foi publicada em periódicos brasileiros de psicologia (e.g., Damásio, 2013; Gomes & Buratto, 2020; Valentini, Mourão, & Franco, 2018). Embora existam publicações brasileiras recentes fornecendo orientações para a elaboração de revisões sistemáticas e metanalíticas (e.g., Carvalho, Pianowski, & Santos, 2019; Silva & Otta, 2014), nenhuma delas teve como foco a síntese estatística e a implementação computacional requeridas para a condução de uma revisão metanalítica. Paralela a essa observação, em uma inspeção de 1.076 artigos publicados em periódicos brasileiros nos últimos 5 anos, constatou-se que nenhuma das publicações inspecionadas fez uso da metanálise.<sup>1</sup> Especulamos que a completa ausência de uso da metanálise é decorrente, ao menos em parte, da falta de orientações em língua portuguesa sobre como realizar a síntese estatística das evidências disponíveis. De fato, o ensino de psicologia baseada em evidências é bem recente no Brasil (Melnik, Meyer, & Sampaio, 2019), o que parece sugerir que a formação em psicologia ainda não expõe estudantes a certos tópicos relacionados ao uso das melhores evidências científicas, como a metanálise.

Com base nessas informações, julgamos que há a necessidade de esforços vindos de dentro da própria área para fomentar o uso de ferramentas que já são de uso comum, por exemplo, nas áreas de saúde. Com o objetivo de acrescentar uma contribuição inédita às literaturas brasileiras sobre revisão sistemática e análise

quantitativa de dados em psicologia, este artigo apresenta conceitos fundamentais sobre metanálise e suas aplicações. O artigo pode ser usado como uma referência introdutória sobre o assunto, tanto por psicólogos quanto por estudantes de psicologia ainda não familiarizados com o tema, mas que precisam consultar ou realizar revisões metanalíticas.

#### *Visão Geral do Artigo*

O texto a seguir é dividido em três seções. Na primeira, introduz-se a lógica da metanálise, seus potenciais e as críticas a ela endereçadas. Na segunda, apresentam-se dois modelos de metanálise comumente usados por pesquisadores (modelos de efeito fixo e de efeitos aleatórios), bem como dois tópicos importantes para a interpretação correta dos resultados (heterogeneidade e análise de subgrupos). Por fim, na terceira seção, as contribuições e limitações do artigo são descritas. De modo a facilitar a compreensão, conceitos importantes são introduzidos por meio de um exemplo fictício e bastante simplificado, visando a didática e o foco na metanálise propriamente dita.

Assume-se que o leitor tem conhecimentos de estatística frequentista em nível básico, tais como sobre teste de hipóteses, valor  $p$ , intervalos de confiança (ICs) e os principais testes descritos em livros-textos de estatística (e.g., Cumming & Calin-Jageman, 2017; Howell, 2013). O texto enfatiza aspectos conceituais da metanálise. Apesar dessa opção, é impossível ignorar completamente a matemática por trás desse conjunto de técnicas. Assim, algumas equações básicas são apresentadas no texto. O Material Suplementar A descreve as principais equações dos modelos apresentados no texto. O Material Suplementar B inclui códigos na linguagem R (R Core Team, 2021), usando os pacotes meta (Balduzzi, Rücker, & Schwarzer, 2019) e dmetar (Harrer, Cuijpers, Furukawa, & Ebert, 2019). Esses códigos permitem reproduzir a simulação dos dados, bem como os resultados e figuras do artigo. O Material Suplementar C apresenta uma breve lista comentada de fontes adicionais que poderão ser úteis a quem quiser se aprofundar no tema. Os Materiais Suplementares A, B e C podem ser acessados no *Open Science Framework* (<https://osf.io/pcxyq/>).

#### *Noções Fundamentais sobre Metanálise*

##### *Revisões Metanalíticas e Medidas de Tamanhos de Efeito*

Revisões de literatura têm o potencial de resumir uma grande quantidade de informações produzidas

<sup>1</sup> Consultamos as publicações de todos os volumes de periódicos brasileiros de psicologia classificados com Qualis A1 no quadriênio 2013–2016 (<https://sucupira.capes.gov.br/>), ocorridas de 2017 a 2021. Os seguintes periódicos foram inspecionados: *Estudos em Psicologia (Campinas)*,  $n = 212$ ; *Estudos em Psicologia (Natal)*,  $n = 182$ ; *Paidéia*,  $n = 218$ ; *Psicologia: Reflexão e Crítica*,  $n = 129$ ; e *Psicologia: Teoria e Pesquisa*,  $n = 335$ . A *Psicologia em Estudo (Maringá)* não foi inspecionada, por ter um escopo voltado a estudos qualitativos.

pela comunidade científica. A revisão metanalítica, por exemplo, consiste em uma forma particular de revisão, composta tanto pela descrição estruturada e transparente de suas questões de pesquisa, escopo e critérios de elegibilidade de estudos (Higgins & Thomas, 2019), quanto pela síntese quantitativa dos resultados dos estudos elegíveis (Cumming & Calin-Jageman, 2017).

Por meio da metanálise, agregam-se os resultados de diferentes estudos primários que investigaram uma questão comum (Howell, 2013), produzindo estimativas pontuais e intervalares. Uma maneira de agregar esses resultados é por meio de uma média aritmética ponderada das estimativas reportadas nos estudos primários (Cheung & Vijayakumar, 2016). Doravante, essa média ponderada será denominada, genericamente, de efeito geral, que estima o tamanho de efeito populacional a partir dos tamanhos de efeito amostrais.

Mas como medidas de tamanho de efeito (incluindo o efeito geral) podem ser definidas? Tamanhos de efeito podem se referir: (a) a parâmetros desconhecidos, isto é, aos verdadeiros tamanhos de efeito populacionais aos quais pesquisadores gostariam de ter acesso (Lakens, 2013); ou (b) a estatísticas usadas para fins descritivos (i.e., expressar a relação observada entre variáveis em um estudo) e/ou inferenciais (i.e., estimar aqueles parâmetros desconhecidos; Borenstein, Hedges, Higgins, & Rothstein, 2009). Desse modo, os tamanhos de efeito reportados (ou potencialmente calculáveis) em estudos primários são as estimativas usadas como pontos de dados em uma revisão metanalítica e, posteriormente, objetos da síntese quantitativa em um efeito geral.

Neste artigo, o tamanho de efeito usado como exemplo é o  $g$  de Hedges que, tal como o  $d$  de Cohen, é interpretado como a diferença média padronizada entre dois grupos em uma variável dependente ou critério. Em outras palavras,  $g$  e  $d$ , análogos ao escore  $z$ , expressam a diferença entre duas médias em unidades de desvios padrões (Hedges & Olkin, 1985). Esses índices permitem comparar efeitos observados em estudos que operacionalizaram um mesmo construto de diferentes formas (Lakens, 2013). Ambos visam estimar o verdadeiro tamanho de efeito populacional ( $\delta$ ), embora o  $d$  tenda a superestimar esse valor em amostras menores (Cooper, Hedges, & Valentine, 2019).<sup>2</sup> Uma vez que  $g$  é uma estimativa corrigida que remove

a maior parte do viés de  $d$ ,  $g$  tende a ser preferido em revisões metanalíticas.

### *Um Exemplo Fictício*

#### *O Efeito de Prática de Recuperação*

A pesquisa em memória humana tem demonstrado que recuperar uma informação da memória, comparada a diferentes condições-controle, melhora sua retenção em longo prazo e facilita a aprendizagem subsequente de novas informações (Chan, Meissner, & Davis, 2018; Rowland, 2014). Essa técnica, denominada de prática de recuperação (PR), tem recebido crescente interesse por parte de psicólogos cognitivos. Tal interesse é justificável, uma vez que a PR produz efeitos robustos (Chan et al., 2018; Rowland, 2014), podendo contribuir como técnica de aprendizagem tanto em contextos educacionais (Roediger & Karpicke, 2006) quanto em contextos clínicos (Lima, Cavendish, Deus, & Buratto, 2020).

Considere o seguinte exemplo fictício baseado no efeito de PR: Grupos de pesquisa distintos investigaram se a PR, comparada ao reestudo, melhora a retenção subsequente do material estudado. Em todos os experimentos, os participantes foram designados aleatoriamente à condição reestudo ou à condição PR. Inicialmente, os participantes leram um texto curto contendo informações factuais. Em seguida, metade dos participantes releu o texto (grupo reestudo) e a outra metade escreveu em uma folha de papel tudo o que foi lembrado do texto previamente lido (grupo PR). Após um dado intervalo de retenção (5 min ou 48 hr), os participantes fizeram um teste de retenção sobre as informações factuais previamente estudadas. Para fins didáticos, assumiremos que o teste foi similar em todos os experimentos. Os desempenhos dos grupos reestudo e PR (i.e., proporção média de acertos no teste de retenção) foram comparados. Diz-se que ocorreu um efeito de PR quando o desempenho do grupo PR é superior ao do grupo reestudo (Rowland, 2014).

#### *Simulação dos Experimentos*

Os dados apresentados neste artigo foram simulados por meio da linguagem R (R Core Team, 2021) com base em uma revisão metanalítica sobre o efeito de PR (Rowland, 2014). Embora os dados sejam simulados, eles capturam didaticamente aspectos importantes da literatura sobre o tema. Aqui, apenas a fundamentação

<sup>2</sup> Ao longo do texto, parâmetros são expressos por meio de letras do alfabeto grego (e.g., a letra delta,  $\delta$ ). Estatísticas de estudos individuais que estimam esses parâmetros são representadas por letras do alfabeto romano (e.g.,  $g$ ). Estimativas obtidas a partir da combinação de dois ou mais estudos (i.e., efeitos gerais) são expressas

por meio de letras do alfabeto grego, acompanhadas por um acento circunflexo (e.g.,  $\delta$ ).

lógica da simulação é apresentada. O leitor interessado nos detalhes da implementação deverá consultar o Material Suplementar B. Vale ressaltar que o exemplo fictício apresentado no restante do artigo consiste em um cenário um tanto homogêneo de procedimentos e de desfechos. Embora reconheçamos que a heterogeneidade de procedimentos e de desfechos deva ser muito maior em aplicações reais de pesquisa, aqui buscamos priorizar a didática, por se tratar de um artigo de caráter introdutório.

Na literatura sobre o efeito de PR, dois padrões comumente emergem: (a) os participantes têm uma maior recordação média após intervalos de retenção mais curtos; e (b) o efeito de PR é maior após intervalos de retenção mais longos (i.e., a vantagem da PR sobre o reestudo aumenta quando o intervalo de retenção é maior). A simulação buscou, portanto, gerar dados de cinco experimentos com um intervalo de retenção curto (i.e., 5 min) e, conseqüentemente, com maior proporção de recordação no teste final; e dados de outros cinco experimentos com intervalo de retenção longo (i.e., 48 hr) e, conseqüentemente, com maior efeito de PR. Essas decisões nos permitiram discorrer sobre uma síntese metanalítica geral, heterogeneidade e análise de subgrupos, tópicos que serão todos tratados mais adiante. Por fim, a simulação foi conduzida de tal modo a gerar dados oriundos de experimentos com diferentes tamanhos amostrais. Essa decisão final nos permitiu discorrer sobre como as estimativas de estudos individuais

são diferencialmente ponderadas nos dois modelos de metanálise que serão apresentados.

#### *Um Exemplo de Síntese Metanalítica*

A Tabela 1 apresenta os resultados de 10 experimentos fictícios usando o procedimento descrito na subseção “O Efeito de Prática de Recuperação.” Nela, é possível notar que metade dos experimentos (6–10) obteve efeitos de PR estatisticamente significativos (assumindo um  $\alpha = 0,05$ , bicaudal). Entre os experimentos que não atingiram significância estatística, um deles ainda mostrou uma vantagem numérica em favor de reestudo (Experimento 5). Em uma revisão narrativa, um pesquisador poderia avaliar erroneamente que a literatura sobre o efeito de PR é inconclusiva, uma vez que 50% dos experimentos tiveram resultados não significativos. No entanto, essa conclusão é inválida, pois ela se baseia na ideia incorreta de que falhar em rejeitar uma hipótese nula é evidência para aceitá-la (Cumming & Calin-Jageman, 2017). É possível, por exemplo, que a ausência de significância estatística seja decorrente de baixo poder estatístico (Howell, 2013). Além disso, essa conclusão ignora que nove experimentos tiveram resultados apontando para a mesma direção (i.e., desempenhos dos grupos PR numericamente superiores aos dos grupos reestudo) e se baseia somente no número de resultados significativos, uma abordagem sabidamente inadequada para sintetizar um conjunto de estudos (Hedges & Olkin, 1985).

Tabela 1.

*Estatísticas Descritivas e Inferenciais de Experimentos Fictícios Investigando o Efeito de Prática de Recuperação*

Experimento	Intervalo de retenção	Prática de recuperação			Reestudo			Estatísticas inferenciais		
		<i>M</i>	<i>DP</i>	<i>n</i>	<i>M</i>	<i>DP</i>	<i>n</i>	<i>t</i>	<i>gl</i>	<i>p</i>
1	5 min	0,75	0,07	50	0,73	0,08	50	1,29	98	0,20
2	5 min	0,75	0,06	45	0,73	0,06	45	1,52	88	0,13
3	5 min	0,75	0,07	15	0,73	0,06	15	1,11	28	0,28
4	5 min	0,75	0,08	20	0,72	0,09	20	1,09	38	0,28
5	5 min	0,72	0,08	25	0,73	0,06	25	-0,76	48	0,45
6	48 hr	0,66	0,08	100	0,61	0,07	100	4,94	198	< 0,001
7	48 hr	0,67	0,09	10	0,58	0,09	10	2,15	18	0,045
8	48 hr	0,66	0,08	40	0,58	0,09	40	4,83	78	< 0,001
9	48 hr	0,66	0,05	35	0,59	0,07	35	4,66	68	< 0,001
10	48 hr	0,65	0,07	30	0,61	0,06	30	2,15	58	0,04

*Nota.* *M* = Média; *DP* = Desvio padrão; *n* = Tamanho da amostra; *gl* = Graus de liberdade; *t* = *t* de Student.

De maneira alternativa, a metanálise permite combinar os resultados dos 10 experimentos, fornecendo uma conclusão que contempla os efeitos observados em cada um deles. A Figura 1A apresenta um gráfico de floresta que sumariza os tamanhos de efeito ( $g$  de Hedges) dos experimentos contidos na Tabela 1, usando um modelo de efeito fixo de metanálise (o termo será explicado depois). Esse gráfico permite que o pesquisador se desloque de uma questão cuja resposta é dicotômica (“Existe um efeito?”) para uma questão cuja resposta é contínua (“Qual é o tamanho do efeito?”; Cumming & Calin-Jageman, 2017). Note que, na Figura 1A, nenhuma menção explícita é feita ao valor  $p$  - a probabilidade de ocorrer um evento igual ou mais extremo que aquele observado nos dados amostrais se a hipótese nula for verdadeira (Cohen, 1994). Por outro lado, o  $g$  de Hedges é apresentado múltiplas vezes no gráfico de floresta, permitindo que o pesquisador avalie a importância prática dos resultados (Howell, 2013). Em nosso exemplo, o  $g$  de Hedges indica a diferença padronizada

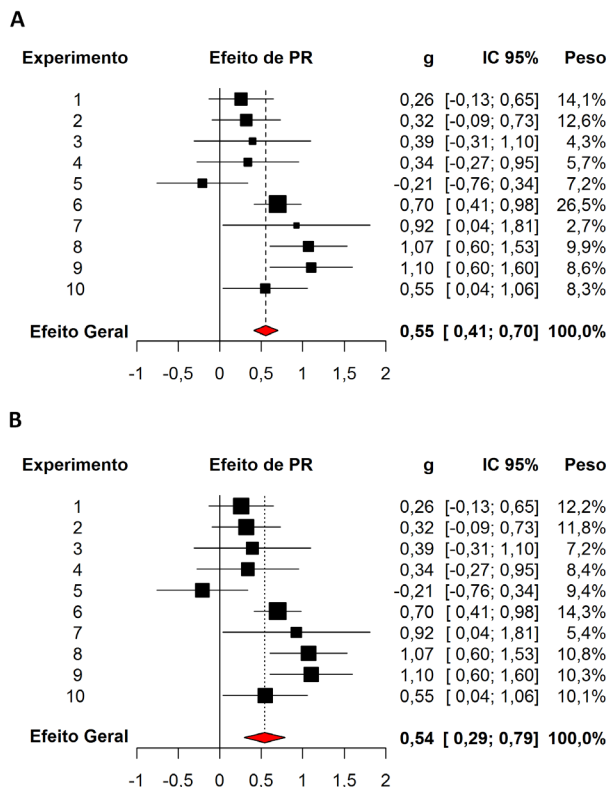


Figura 1. Gráfico de floresta sintetizando os experimentos descritos na Tabela 1 sobre o efeito de prática de recuperação. O painel A ilustra o modelo de efeito fixo. O painel B ilustra o modelo de efeitos aleatórios.  $g$  = Tamanho de efeito ( $g$  de Hedges); IC = intervalo de confiança.

de desempenho no teste de retenção dos participantes dos grupos PR e reestudo. Valores positivos indicam que o grupo PR teve um desempenho superior no teste de retenção (e.g., Experimento 8), enquanto valores negativos indicam que o grupo reestudo teve um desempenho superior (e.g., Experimento 5).

Na Figura 1A, a posição de cada quadrado em relação ao eixo  $x$  representa o tamanho de efeito observado para cada experimento. Ao redor de cada quadrado, barras indicam um IC de 95%, um intervalo construído ao redor do  $g$  de Hedges (ou de qualquer outra estimativa) por meio de um procedimento que garante que, ao longo de amostragens repetidas hipotéticas, em 95% das ocasiões, o parâmetro populacional que se busca estimar estará contido entre seus limites (Hoekstra, Morey, Rouder, & Wagenmakers, 2014). Os Experimentos 6–10 têm ICs 95% que não incluem  $g = 0$ , enquanto os demais têm ICs 95% que incluem  $g = 0$ . Esse padrão espelha os experimentos cujos resultados são estatisticamente significativos na Tabela 1. Assim, os valores  $p$  dos estudos originais são apresentados no gráfico de floresta apenas implicitamente: ICs 95% que incluem  $g = 0$  correspondem a  $ps \geq 0,05$ , enquanto ICs 95% que não incluem  $g = 0$  correspondem a  $ps < 0,05$ . No entanto, a apresentação dos resultados por meio de ICs 95% é mais informativa (Cohen, 1994), por incluir estimativas intervalares de valores plausíveis do tamanho de efeito (Cumming & Calin-Jageman, 2017).

O tamanho de cada quadrado da Figura 1A informa ao leitor o peso que cada estudo recebeu no cálculo do efeito geral. O Experimento 6 teve o maior peso no cálculo do efeito geral (26,5%) e o Experimento 7 teve o menor peso (2,7%). Isso ocorre porque o Experimento 6, com um maior tamanho amostral, foi capaz de estimar o  $g$  de Hedges com maior precisão, enquanto o oposto ocorreu no Experimento 7.

O efeito geral, representado pelo losango vermelho na Figura 1A, refere-se à estimativa do tamanho de efeito de PR populacional, com base nos experimentos metanalisados. O centro do losango representa a estimativa do efeito geral (nesse caso,  $g = 0,55$ ), enquanto sua largura representa os limites de confiança do grande efeito, IC 95% [0,41, 0,70]. Embora 50% dos experimentos revisados tenham incluído  $g = 0$  em seus ICs 95%, o efeito geral sugere que valores inferiores a 0,41 são improváveis de representar o efeito de PR populacional. Assim, o uso da metanálise permite a conclusão de que o efeito de PR é confiável. O efeito geral, representado pela “floresta” (i.e., losango) do gráfico, é visualizado apenas a partir do agrupamento de todas as

“árvores” (i.e., quadrados) que compõem essa floresta (Cumming, 2005). Esse exemplo materializa a ideia, apresentada na introdução, de que a interpretação dos resultados de um estudo deve ser feita à luz dos demais estudos sobre o mesmo fenômeno.

#### *Potenciais e Críticas do Uso da Metanálise na Psicologia*

O exemplo anterior forneceu um vislumbre inicial do que a metanálise tem a oferecer. De fato, o uso da metanálise na psicologia tem uma série de potenciais. Primeiro, seu uso aumenta a precisão das estimativas do tamanho de efeito ou do tamanho da relação entre variáveis (Cumming & Calin-Jageman, 2017), tal como indicado pelo losango da Figura 1A. Segundo, seu uso permite sintetizar e organizar o conhecimento sobre um dado tema (Cumming, 2005). Essa organização permite o estabelecimento da confiabilidade de um achado, o contraste de hipóteses concorrentes sobre um fenômeno e a indicação de áreas com poucas pesquisas e que merecem maior atenção em estudos futuros (para exemplos, ver Chan et al., 2018; Rowland, 2014).

Terceiro, as estimativas de tamanho de efeito derivadas de revisões metanalíticas podem posteriormente ser usadas em análises de poder *a priori*, que calculam o tamanho amostral necessário para observar um determinado tamanho de efeito, dados o nível de significância e o poder estatístico almejados (Lakens, 2013). Quarto, a metanálise permite investigar variáveis moderadoras que, potencialmente, afetam o tamanho ou a direção dos efeitos observados (i.e., por meio de análise de subgrupos; Cooper, 2016). Quinto, embora o uso da metanálise geralmente ocorra dentro de revisões de literatura, ela também pode ser usada fora desse contexto (Cheung & Vijayakumar, 2016). Recentemente, por exemplo, Su et al. (2021) sintetizaram seus quatro experimentos sobre o efeito de PR em uma pequena metanálise interna, de modo a aumentar o poder estatístico das relações de interação que investigaram.

Quanto às críticas, as principais delas são sumarizadas a seguir. Uma forte crítica é que a metanálise produz resultados distorcidos pelo viés de publicação, o achado de que estudos que produzem resultados estatisticamente significativos são mais prováveis de serem publicados do que aqueles que não os produzem (Rosenthal, 1979). Essa, na verdade, parece ser uma crítica mais ampla dirigida às deficientes práticas modais de pesquisa em psicologia (LeBel & Peters, 2011), incluindo o uso dogmático e mecânico da inferência estatística (Gigerenzer, 1993) e o peso que o valor *p*

tem nas decisões editoriais de vários periódicos científicos (Cumming & Calin-Jageman, 2017). Se o uso da metanálise pode levar a conclusões distorcidas pelo viés de publicação, o mesmo se aplica a outras formas de revisões (e.g., revisões narrativas). As revisões metanalíticas, ao contrário das narrativas, podem (a) adotar estratégias para minimizar a seletividade dos estudos amostrados (Cooper et al., 2019; Cooper, 2016) e (b) por meio de diferentes técnicas analíticas, investigar a extensão do viés de publicação e aplicar correções nas estimativas do efeito geral (Borenstein et al., 2009; Cooper et al., 2019).

Outras duas críticas relacionadas entre si são de que (a) ao sintetizar uma literatura com um único número, perde-se informação; e (b) é inadequado combinar estudos com qualidades e procedimentos metodológicos diversificados. Quanto à crítica (a), sintetizar os resultados de diferentes estudos não deve ser entendido, necessariamente, como sinônimo de focar exclusivamente no valor do grande efeito (Borenstein et al., 2009). A metanálise permite que o pesquisador avalie a heterogeneidade ou inconsistência dos tamanhos de efeito, de modo que as conclusões do pesquisador possam ser qualificadas à luz da presença ou da ausência de heterogeneidade, bem como de variáveis procedimentais que moderam os tamanhos de efeito (Rosenthal, 1991).

Quanto à crítica (b), suponha que os experimentos da Tabela 1 usassem textos de diferentes domínios de conhecimento (e.g., biologia, história, literatura), amstrassem distintos grupos de participantes (e.g., crianças, adultos, pacientes com afasia), ou, ainda, fossem conduzidos em diferentes contextos (e.g., laboratórios, salas de aula, plataformas *on-line*). Nesse cenário, seria legítimo comparar estudos com procedimentos que diferissem em tantas dimensões? Glass (1978) argumenta que é legítimo comparar maçãs e laranjas quando o objetivo é generalizar os achados para a categoria frutas. De fato, um dos objetivos da ciência é a identificação de princípios gerais de funcionamento do mundo (Schmidt, Oh, & Hayes, 2009). Uma vantagem de comparar estudos com diversidade metodológica que abordam uma questão conceitual comum é a possibilidade de identificar condições limítrofes para um dado fenômeno, isto é, estabelecer as fronteiras para além das quais os princípios propostos não se aplicam. Quanto à adequabilidade de se combinar estudos com níveis variados de qualidade metodológica, têm-se argumentado que a relação entre a qualidade metodológica dos estudos e as estimativas de tamanhos de efeito deve ser

investigada na própria revisão metanalítica (Cooper et al., 2019; Glass, 1978).

### Modelos Computacionais de Metanálise

Dois modelos de metanálise são tipicamente usados em revisões metanalíticas: O modelo de efeito fixo e o modelo de efeitos aleatórios. Cada modelo pressupõe distintas fontes de erros nas estimativas dos tamanhos de efeito em estudos primários. As subseções seguintes apresentam os modelos. Em seguida, é feita uma breve comparação entre eles. Por fim, dois tópicos importantes para a interpretação correta dos resultados - heterogeneidade e análise de subgrupos - são apresentados.

#### Modelo de Efeito Fixo

O modelo de efeito fixo assume, *a priori*, que todos os estudos independentes incluídos na revisão estimam um valor comum (Schmidt et al., 2009). Fixo, nesse caso, refere-se ao pressuposto de que os parâmetros populacionais estimados têm valores constantes nos diferentes estudos (Hedges & Vevea, 1998). Tomando emprestado um exemplo didático de Howell (2013, pp. 630 - 631), em um cenário em que astrônomos estimam a luminosidade de uma determinada estrela, é razoável assumir que o valor de interesse (i.e., a real luminosidade da estrela) é constante e que diferenças em estimativas sucessivas se devam unicamente ao erro amostral. Formalmente, em uma síntese metanalítica com  $k$  estudos, o modelo de efeito fixo pode ser expresso por:

$$Y_i = \theta + \varepsilon_i, \quad (1)$$

onde  $Y_i$  representa o tamanho de efeito estimado no estudo  $i$ ,  $\theta$  representa o verdadeiro tamanho de efeito populacional que todos os estudos buscam estimar, e  $\varepsilon_i$  representa o erro amostral para o estudo  $i$ . Nesse modelo, assim como no exemplo da luminosidade da estrela, a única fonte de incerteza nas inferências de cada estudo é o erro amostral,  $Var(\varepsilon_i) = V_{Y_i}$  (Cheung, 2019). A teoria estatística indica que estudos com amostras maiores tendem a estimar o tamanho de efeito populacional com menor erro (Cooper et al., 2019).

A Equação 1 é expressa de forma genérica, de modo que  $\theta$  pode se referir a qualquer parâmetro populacional de interesse. Aqui, retomaremos o exemplo apresentado na Figura 1A (portanto,  $\theta$  e  $Y_i$  equivalem a  $\delta$  e a  $g_i$ , respectivamente). Considere que  $\theta$ , nesse exemplo, refere-se ao verdadeiro tamanho de efeito de PR de uma população de experimentos hipotéticos, que assumiremos ser  $\delta = 0,50$  (com base em uma das estimativas reportadas por Rowland, 2014).

Suponha que os experimentos da Figura 1A buscavam estimar esse tamanho de efeito. Os Experimentos 5 - 10 superestimaram o valor de  $\delta$  ( $g_i$  entre 0,70 e 1,10), enquanto os demais experimentos o subestimaram ( $g_i$  entre -0,21 e 0,39).

Uma forma simples de combinar as estimativas desses experimentos é por meio da média aritmética simples. No entanto, estimativas provenientes de estudos com amostras maiores tendem a ser mais precisas, de modo que, no cálculo do efeito geral, os estimadores são ponderados pelas inversas de suas variâncias (Hedges & Vevea, 1998). Assim, um modelo de efeito fixo: (1) estima a variância do tamanho de efeito,  $V_{Y_i}$ , para o estudo  $i$  (de 1 a  $k$ ); (2) atribui um peso,  $W_i$ , ao estudo  $i$ , igual à inversa de sua variância (i.e.,  $W_i = 1/V_{Y_i}$ ); e (3) usa esses valores no cálculo da estimativa média ponderada do efeito geral:

$$\hat{\theta} = \frac{\sum_{i=1}^k W_i Y_i}{\sum_{i=1}^k W_i}, \quad (2)$$

onde a Equação 2 expressa que a estimativa do tamanho de efeito populacional,  $\hat{\theta}$ , é uma média aritmética ponderada dos tamanhos de efeito amostrais pelos seus respectivos pesos. No exemplo da Figura 1A, usando a média ponderada, obtém-se um efeito geral que se aproxima razoavelmente ( $\hat{\delta} = 0,55$ ) do verdadeiro tamanho de efeito populacional ( $\delta = 0,50$ ; para cálculos completos, ver Material Suplementar A).

#### Modelo de Efeitos Aleatórios

O modelo de efeitos aleatórios assume, *a priori*, que o valor do parâmetro  $\theta$  varia aleatoriamente entre estudos independentes, sendo ele próprio amostrado de uma superpopulação de parâmetros verdadeiros (Borenstein et al., 2009). Aleatório, nesse caso, refere-se ao pressuposto de que os parâmetros populacionais estimados em cada estudo são genuinamente variáveis (Hedges & Vevea, 1998). É possível, por exemplo, que a eficácia de uma psicoterapia seja distinta a depender da orientação psicoterápica (e.g., psicodinâmica, cognitivo-comportamental), de seu formato (e.g., individual, grupal) ou de seu público-alvo (e.g., adolescentes, adultos). Retomando o exemplo da Tabela 1, é possível que a PR gere tamanhos de efeitos distintos a depender das características metodológicas dos estudos (e.g., conteúdos dos materiais estudados, contextos em que os estudos são conduzidos). Aqui, o ponto importante é que o modelo de efeitos aleatórios assume que há duas fontes de incerteza nas inferências, o erro amostral (tal como no modelo de efeito fixo) e a variância dos tamanhos

de efeito entre estudos (Cooper et al., 2019). Formalmente, em uma síntese metanalítica com  $k$  estudos, o modelo de efeitos aleatórios pode ser expresso por:

$$Y_i = \mu_0 + \xi_i + \varepsilon_i, \quad (3)$$

onde  $Y_i$  representa a estimativa do tamanho de efeito  $\theta_i$  no estudo  $i$  e  $\mu_0$  é um hiperparâmetro que expressa o tamanho de efeito médio da superpopulação de parâmetros verdadeiros (Hedges & Vevea, 1998). A notação  $\xi_i$  indica a diferença entre  $\theta_i$  e  $\mu_0$ , enquanto  $\varepsilon_i$  representa, mais uma vez, o erro amostral.

No modelo de efeitos aleatórios, o efeito geral também consiste em uma média ponderada dos tamanhos de efeito dos estudos incluídos na síntese. Conceitualmente, diferentes inferências são feitas em cada modelo. No modelo de efeito fixo, o efeito geral estima o parâmetro  $\theta$ , o verdadeiro tamanho de efeito populacional; ao incorporar apenas o erro amostral como fonte de incerteza, as inferências são restritas ao conjunto de estudos incluídos na revisão (Hedges & Vevea, 1998). No modelo de efeitos aleatórios, o efeito geral estima o hiperparâmetro  $\mu_0$ , a média da distribuição de verdadeiros tamanhos de efeito populacionais (i.e., a superpopulação); ao incorporar o erro amostral e a variância entre estudos como fontes de incerteza, busca-se fazer inferências para a população de estudos (observados ou não) a partir da qual os estudos incluídos na revisão são considerados uma amostra representativa (Cooper et al., 2019).

Computacionalmente, o modelo de efeitos aleatórios estima o erro amostral,  $V_{Y_i}$ , para cada estudo (similar ao modelo de efeito fixo) e a variância da distribuição populacional de tamanhos de efeito,  $Var(\xi_i) = \tau^2$  (Cheung, 2019), por meio da estatística  $T^2$  (Borenstein et al., 2009). A estatística  $T^2$ , ao contrário de  $V_{Y_i}$ , é comum a todos os estudos incluídos na revisão. Nesse modelo, a estimativa da variância total,  $V_{Y_i}^*$ , consiste na soma do erro amostral e da estimativa da variância da distribuição populacional de tamanhos de efeito,  $V_{Y_i}^* = V_{Y_i} + T^2$ . O peso de cada estudo, no modelo de efeitos aleatórios, é igual à inversa de sua variância total,  $W_i^* = 1/V_{Y_i}^*$  (Hedges & Vevea, 1998). Na prática, há diferentes maneiras de estimar  $\tau^2$  (e.g., estimação pela máxima verossimilhança), sendo o método dos momentos o mais comumente usado, devido à sua simplicidade computacional (DerSimonian & Laird, 1986).

O cálculo do efeito geral, no modelo de efeitos aleatórios, é análogo à Equação 2 previamente apresentada, exceto por estimar um hiperparâmetro e por usar outra forma de ponderação para os diferentes estudos:

$$\widehat{\mu}_\theta = \frac{\sum_{i=1}^k W_i^* Y_i}{\sum_{i=1}^k W_i^*} \quad (4)$$

onde a estimativa do hiperparâmetro populacional,  $\widehat{\mu}_\theta$ , é uma média aritmética ponderada dos tamanhos de efeito amostrais pelos seus respectivos pesos, que incorporam tanto o erro amostral quanto a estimativa da variância entre estudos.

A Figura 1B apresenta um gráfico de floresta baseado em uma síntese metanalítica usando um modelo de efeitos aleatórios. Os tamanhos de efeito e ICs 95% dos experimentos são idênticos àqueles da Figura 1A. Por outro lado, os quadrados (i.e., pesos) de cada experimento divergem entre os modelos. Consequentemente, o efeito geral e o IC 95% do modelo de efeitos aleatórios difere ( $g = 0,54 [0,29, 0,79]$ ) dessas mesmas estimativas usando o modelo de efeito fixo ( $g = 0,55 [0,41, 0,70]$ ). Essa diferença é decorrente do fato de que, na medida em que cresce o valor de  $\tau^2$ , aumenta-se a homogeneidade nos pesos dos diferentes estudos incluídos na síntese. No modelo de efeito fixo, a razão entre os pesos dos Experimentos 6 e 7 foi de 9,81; no modelo de efeitos aleatórios, essa razão caiu para 2,65. O modelo de efeitos aleatórios assume que cada estudo da revisão fornece uma estimativa única de um parâmetro; assim, todos os estudos da revisão contribuem com informações relevantes para o cálculo do efeito geral.<sup>3</sup> Além disso, na medida em que a estimativa de  $\tau^2$  aumenta, as variâncias, os erros padrões e os ICs ao redor do efeito geral tendem a ser maiores no modelo de efeitos aleatórios, em comparação com o modelo de efeito fixo (Borenstein et al., 2009). Isso decorre do fato de que, ao incorporar uma segunda fonte de variabilidade dos tamanhos de efeito, o modelo de efeitos aleatórios aumenta a incerteza sobre o provável valor do hiperparâmetro (Hedges & Vevea, 1998).

#### *Modelo de Efeito Fixo ou Modelo de Efeitos Aleatórios?*

Uma questão com a qual pesquisadores se confrontam durante a realização de uma revisão metanalítica diz respeito a qual modelo de metanálise usar. Tem-se defendido que a escolha do modelo de metanálise deve ser feita com base nas inferências almeçadas

<sup>3</sup> Apenas o estudo  $i$  contribui para estimar o parâmetro  $\theta_i$ . Os demais  $k - 1$  estudos estão estimando outros  $\theta_{k-1}$  parâmetros distintos. Assim, é desejável que o peso do estudo  $i$  para o cálculo do efeito geral não seja muito pequeno, pois a informação que ele fornece é única e relevante para aquele parâmetro da distribuição de parâmetros que, em última análise, estimará  $\mu_0$ .



(Field & Gillett, 2010; Hedges & Vevea, 1998). Se o pesquisador quer fazer inferências restritas ao conjunto de estudos incluídos na revisão, deve optar pelo modelo de efeito fixo; se ele quer fazer inferências que se estendem para a população de estudos (observados ou não) a partir da qual os estudos incluídos na revisão foram amostrados, deve optar pelo modelo de efeitos aleatórios. Adicionalmente, o modelo a ser escolhido depende de quais pressupostos podem ser adotados, de forma realista, acerca das populações a partir das quais os estudos foram amostrados (Field & Gillett, 2010). Uma vez que revisões metanalíticas geralmente incluem estudos com variabilidade metodológica e que a ciência almeja, na medida do possível, o conhecimento generalizável (Schmidt et al., 2009), o modelo de efeitos aleatórios parece ser mais apropriado na ciência psicológica (Cumming & Calin-Jageman, 2017).

Algumas observações são pertinentes. Primeiro, um problema em revisões metanalíticas com modelos de efeitos aleatórios é uma definição tautológica da população, que pode ter a seguinte forma: “A população de interesse é a população a partir da qual a amostra de estudos foi retirada” (Hedges & Vevea, 1998, p. 489). Suponha que os experimentos da Figura 1B tivessem sido conduzidos com adultos e em laboratório. Embora o modelo de efeitos aleatórios permita a inferência para a população de estudos (observados ou não), seria incorreto, a partir dos resultados da Figura 1B, concluir que eles são informativos sobre o status do efeito de PR em outras populações (e.g., idosos) ou em contextos fora do laboratório (e.g., salas de aula).

Segundo, simulações indicam que a estimativa de  $\tau^2$  tende a ser imprecisa quando o número de estudos da revisão é baixo,  $k < 20$  (Hedges & Vevea, 1998). Assim como amostras pequenas em estudos primários implicam baixo poder estatístico, o mesmo pode ocorrer em revisões metanalíticas, quando poucos estudos são incluídos na revisão. Terceiro, os dois modelos previamente apresentados assumem que os estudos incluídos na síntese são independentes entre si. Em um sentido tradicional, o tamanho de efeito do estudo  $i$  é considerado não independente se ele compartilhar as mesmas unidades amostrais (e.g., participantes) em seus cálculos com pelo menos um dos demais  $k - 1$  estudos (Rosenthal, 1979, 1991). Em um sentido alternativo, dois tamanhos de efeito, embora baseados em unidades amostrais distintas, são considerados não independentes entre si se eles estiverem aninhados em uma unidade hierárquica superior (e.g., estudos reportados em um mesmo artigo ou conduzidos por um mesmo

grupo de pesquisa). Nesses cenários, é preciso usar modelos computacionais de metanálise que incorporam explicitamente a dependência das estimativas em suas formalizações, tais como modelos multivariados ou de três níveis de metanálise (para exemplos desses modelos, ver Cheung, 2019).

#### *Heterogeneidade e Análise de Subgrupos*

A homogeneidade dos tamanhos de efeito populacionais é um pressuposto do modelo de efeito fixo (Hedges & Vevea, 1998). Tal pressuposto pode ser testado por meio do teste  $Q$  de Cochran. Conceitualmente análoga à estatística  $F$  da análise de variância, a estatística  $Q$  decompõe as variâncias intra e entre estudos e avalia se a variância observada entre estudos excede o que seria esperado apenas devido ao erro amostral se todos os estudos revisados compartilhassem um mesmo tamanho de efeito populacional (Hedges & Vevea, 1998). Se o valor da estatística  $Q$  exceder o valor crítico de uma distribuição  $\chi^2$  com  $k - 1$  graus de liberdade e um dado nível de significância, assume-se que a variância dos tamanhos de efeito entre estudos é maior do que seria esperado se os tamanhos de efeito populacionais fossem homogêneos (Cooper et al., 2019).

No entanto, a estatística  $Q$  possui baixo poder estatístico quando o  $k$  é pequeno, além de rejeitar desvios sutis da homogeneidade quando o  $k$  é grande (Higgins & Thomas, 2019). Outras medidas podem ser utilizadas para comunicar diferentes informações a respeito da heterogeneidade. Derivado da própria estatística  $Q$ , o  $I^2$  é uma estatística descritiva que expressa o percentual da variância nas estimativas de tamanho de efeito observadas que se deve à variância real nos de tamanhos de efeito populacionais (Higgins & Thompson, 2002). As estatísticas  $Q$  e  $I^2$ , contudo, não são informativas do quanto variam os tamanhos de efeito populacionais. Essa informação é dada pela estatística  $T^2$  e pelo intervalo de predição ( $IP$ ). Aqui, apenas o  $IP$  é descrito (para uma discussão sobre o  $T^2$  e o método dos momentos, ver Material Suplementar A).

Um  $IP$  consiste em uma estimativa intervalar da região na qual uma observação futura aleatória ocorrerá com uma determinada frequência. Enquanto o  $IC$  informa a precisão do valor estimado (i.e.,  $IC$ s mais curtos indicam menor incerteza no valor do parâmetro), o  $IP$  informa a dispersão dos verdadeiros parâmetros populacionais (i.e.,  $IP$ s mais curtos indicam menor variabilidade nos parâmetros da superpopulação de parâmetros; Borenstein et al., 2009). Recomenda-se que o  $IP$  seja reportado em revisões metanalíticas, de modo

a informar ao leitor em que medida os tamanhos de efeito populacionais variam (Cooper et al., 2019).

Em nosso exemplo fictício, o pressuposto de homogeneidade dos tamanhos de efeito populacionais foi rejeitado,  $Q(9) = 22,65$ ,  $p = 0,007$ . O *IP* 95% sugere que os tamanhos de efeito variam desde valores negativos,  $-0,21$ , até valores positivos,  $1,28$ , denotando que nem sempre a PR deverá trazer benefícios mnemônicos para aprendizes. É importante notar que o *IP* usa a estatística  $T^2$  em seus cálculos e que, como previamente mencionado, a estimativa da variância da superpopulação de tamanhos de efeito tende a ser imprecisa quando o tamanho amostral é baixo (Hedges & Vevea, 1998). É possível, portanto, que a imprecisão do  $T^2$  tenha se propagado na estimativa do *IP* 95%. No entanto, dados os propósitos didáticos do exemplo, a partir dos resultados observados, seria sugerido proceder à investigação de potenciais variáveis moderadoras dos tamanhos de efeito.

Uma variável é considerada uma variável moderadora se ela modifica a magnitude ou a direção do relacionamento entre uma variável independente ou preditora,  $X$ , e uma variável dependente ou desfecho,  $Y$  (Valentini et al., 2018). Se  $X$  e  $Y$  estão positivamente correlacionadas para crianças, mas não para adultos, diz-se que a faixa etária modera a relação entre essas variáveis. Em uma revisão metanalítica, as potenciais variáveis moderadoras correspondem a características metodológicas com importância teórica ou empírica. Essas variáveis podem ser identificadas por meio de análise de subgrupos, cujo objetivo é identificar quais características metodológicas explicam parte da heterogeneidade dos tamanhos de efeito (Hedges & Olkin, 1985).

Os experimentos da Tabela 1 foram categorizados segundo seus respectivos intervalos de retenção, 5 min e 48 hr. A partir dessa categorização, é possível conduzir uma análise de subgrupos. Nela, os estudos de cada subgrupo foram sumarizados por meio de modelos de efeitos aleatórios e, posteriormente, os efeitos gerais de cada subgrupo foram sumarizados em um efeito geral combinado por meio de um modelo de efeito fixo (Borenstein et al., 2009). A estatística  $Q_B$ , que decompõe os componentes de variância entre grupos e de variância total, testa a hipótese nula de que os efeitos gerais dos diferentes subgrupos são homogêneos (Cooper et al., 2019). No caso dos dados da Tabela 1, a estatística  $Q_B$  indicou que houve diferenças significativas entre os efeitos gerais dos subgrupos 5 min e 48 hr,  $Q_B(1) = 15,11$ ,  $p < 0,001$ . O efeito geral de PR é maior em experimentos com intervalo de retenção de

48 hr ( $g = 0,82$ , *IC* 95% [0,61, 1,02]) do que em experimentos com intervalo de retenção de 5 min ( $g = 0,22$  [0,001, 0,44]). É importante notar que nenhum dos dois *ICs* 95% incluiu o 0 como um valor plausível para o tamanho do efeito de PR.

A análise de subgrupos aqui apresentada usou uma variável moderadora categórica, análoga à análise de variância, em estudos primários. Essa análise pode ser estendida para casos em que há uma ou mais variáveis moderadoras categóricas e/ou contínuas, análoga às regressões simples e múltipla, em estudos primários (denominada, na literatura sobre metanálise, de metarregressão; Borenstein et al., 2009). Em todos os casos, objetiva-se identificar moderadores que explicam parte da heterogeneidade dos tamanhos de efeito incluídos na síntese metanalítica (Hedges & Olkin, 1985).

### Considerações Finais

Este artigo apresentou conceitos fundamentais sobre metanálise e suas aplicações. Por meio de um exemplo fictício, foi possível: (a) apresentar quais são os dados usados em uma revisão metanalítica (i.e., os tamanhos de efeito e as estimativas de suas variâncias); (b) explicar como as estimativas são ponderadas pelos seus níveis de precisão; (c) diferenciar as fontes de erro nas estimativas de tamanhos de efeito assumidas nos modelos de efeito fixo e de efeitos aleatórios de metanálise; e (d) descrever como as análises de heterogeneidade e de subgrupos contribuem para a melhor caracterização dos resultados de uma metanálise. Os Materiais Suplementares A, B e C permitem que o leitor interessado reproduza nossos resultados e se aprofunde nos tópicos aqui tratados. Até onde sabemos, trata-se da primeira referência introdutória em língua portuguesa sobre o assunto voltada especialmente para o público de psicologia. A leitura de um tema novo e em outra língua pode dificultar a compreensão das informações relevantes; esse texto permite um contato inicial do falante do português com o tema em sua própria língua.

Algumas limitações do artigo devem ser reconhecidas. O artigo abordou apenas dois modelos computacionais de metanálise. Modelos alternativos foram deixados de fora dessa introdução, como os modelos multivariados e de três níveis (Cheung, 2019). Além disso, não caracterizamos a metanálise em um contexto mais amplo, que inclui as etapas necessárias para a realização de uma revisão metanalítica (para esse fim, veja Carvalho et al., 2019; Cooper, 2016; Silva & Otta, 2014). Por fim, o exemplo fictício usado

ao longo do artigo teve procedimentos e desfechos relativamente homogêneos, o que é implausível de ser observado em aplicações reais de pesquisa (e.g., Rowland, 2014). Apesar dessas limitações, julgamos que a abordagem adotada nos permitiu introduzir adequadamente conceitos importantes sobre metanálise. As escolhas feitas na simulação priorizaram a didática (e.g., possibilitar uma análise de subgrupos). Em nosso entendimento, a homogeneidade dos estudos de nosso exemplo fictício não afetou as mensagens que almejamos transmitir com o artigo.

A ciência é um empreendimento cumulativo. Resultados de estudos individuais são probabilísticos. Na medida em que o número de estudos conduzidos aumenta, torna-se mais difícil se manter atualizado em um determinado assunto. O uso da metanálise pode contribuir para solucionar o problema, bem descrito por Glass (1976), de como extrair conhecimento em meio a uma grande quantidade de informações.

### Referências

- Balduzzi, S., Rücker, G., & Schwarzer, G. (2019). How to perform a meta-analysis with R: A practical tutorial. *Evidence Based Mental Health*, 1-8. doi:10.1136/ebmental-2019-300117
- Barrett, F. S., Carbonaro, T. M., Hurwitz, E., Johnson, M. W., & Griffiths, R. R. (2018). Double-blind comparison of the two hallucinogens psilocybin and dextromethorphan: effects on cognition. *Psychopharmacology*, 235(10), 2915-2927. doi:10.1007/s00213-018-4981-x
- Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Wiley.
- Carvalho, L. F., Pianowski, G., & Santos, M. A. (2019). Guidelines for conducting and publishing systematic reviews in psychology. *Estudos em Psicologia (Campinas)*, 36, eArticle e180144. doi:10.1590/1982-0275201936e180144
- Chan, J. C. K., Meissner, C. A., & Davis, S. D. (2018). Retrieval potentiates new learning: A theoretical and meta-analytic review. *Psychological Bulletin*, 144(11), 1114-1146. doi:10.1037/bul0000166
- Cheung, M. W.-L. (2019). A guide to conducting a meta-analysis with non-independent effect sizes. *Neuropsychology Review*, 29(4), 387-396. doi:10.1007/s11065-019-09415-6
- Cheung, M. W.-L., & Vijayakumar, R. (2016). A guide to conducting a meta-analysis. *Neuropsychology Review*, 26(2), 121-128. doi:10.1007/s11065-016-9319-z
- Cohen, J. (1994). The Earth is round ( $p < .05$ ). *American Psychology*, 49(12), 997-1003. doi:10.1037/0003-066X.49.12.997
- Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2019). *The handbook of research-synthesis and meta-analysis* (3<sup>rd</sup> ed.). Russell Sage Foundation.
- Cooper, H. M. (2016). *Research synthesis and meta-analysis: A step-by-step approach* (5<sup>th</sup> ed.). Sage.
- Cumming, G. (2005). *Understanding the new statistics: Effect sizes, confidence intervals, and meta-analysis*. Routledge.
- Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge.
- Damáσιο, B. F. (2013). Contribuições da Análise Fatorial Confirmatória Multigrupo (AFCMG) na avaliação de invariância de instrumentos psicométricos. *Psico-USF*, 18(2), 211-220. doi:10.1590/S1413-82712013000200005
- DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, 7(3), 177-188. doi:10.1016/0197-2456(86)90046-2
- Field, A. P., & Gillett, R. (2010). How to do a meta-analysis. *British Journal of Mathematical and Statistical Psychology*, 63, 665-694. doi:10.1348/000711010X502733
- Gigerenzer, G. (1993). The Superego, the Ego, and the Id in statistical reasoning. In G. Keren & C. Lewis (Eds.), *A handbook for data analysis in the behavioral sciences: Methodological issues* (pp. 311-339). Lawrence Erlbaum Associates, Publishers.
- Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher*, 5(10), 3-8. doi:10.3102/0013189X005010003
- Glass, G. V. (1978). In defense of generalization. *Behavior and Brain Sciences*, 1(3), 394-395. doi:10.1017/s0140525x00075610
- Gomes, C. F. A., & Buratto, L. G. (2020). Modelagem multinomial aplicada à pesquisa em psicologia. *Psicologia em Pesquisa*, 14(3), 44-65. doi:10.34019/1982-1247.2020.v14.29542
- Harrer, M., Cuijpers, P., Furukawa, T., & Ebert, D. D. (2019). *dmatar: Companion R package for the guide 'Doing Meta-analysis in R'*. R package version 0.0.90000. Recuperado de <https://dmatar.protectlab.org/>

- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press, Inc.
- Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, 3(4), 486-504. doi:10.1037/1082-989x.3.4.486
- Higgins, J. P. T., & Thomas, J. (Eds.). (2019). *Cochrane handbook for systematic reviews of interventions* (2<sup>nd</sup> ed.). The Cochrane Collaboration.
- Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21, 1539-1158. doi:10.1002/sim.1186
- Hoekstra, R., Morey, R. D., Rouder, J. N., & Wagenmakers, E.-J. (2014). Robust misinterpretation of confidence intervals. *Psychonomic Bulletin & Review*, 21(5), 1157-1164. doi:10.3758/s13423-013-0572-3
- Howell, D. C. (2013). *Statistical methods for psychology* (8<sup>th</sup> ed.). Cengage Learning.
- Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology*, 4, eArticle 00863. doi:10.3389/fpsyg.2013.00863
- LeBel, E. P., & Peters, K. R. (2011). Fearing the future of empirical psychology: Bem's (2011) evidence of psi as a case study of deficiencies in modal research practice. *Review of General Psychology*, 15(4), 371-379. doi:10.1037/a0025172
- Lima, M. F. R., Cavendish, B. A., Deus, J. S., & Buratto, L. G. (2020). Retrieval practice in memory- and language-impaired populations: A systematic review. *Archives of Clinical Neuropsychology*, 35(7), 1078-1093. doi:10.1093/arclin/aaa035
- Melnik, T., Meyer, S. B., & Sampaio, M. I. C. (2019). Relatório de experiência docente: A primeira disciplina no Brasil sobre a prática da psicologia baseada em evidências ministrada no Instituto de Psicologia da Universidade de São Paulo. *Psicologia: Teoria e Pesquisa*, 35, eArticle e35418. doi:10.1590/0102.3772e35418
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), 943-952. doi:10.1126/science.aac4716
- R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria. Recuperado de <https://www.R-project.org/>
- Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1(3), 181-210. doi:10.1111/j.1745-6916.2006.00012.x
- Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86(3), 638-641. doi:10.1037/0033-2909.86.3.638
- Rosenthal, R. (1991). *Meta-analytic procedures for social research - Revised edition*. SAGE Publications, Inc.
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140(6), 1432-1463. doi:10.1037/a0037559
- Schmidt, F. L., Oh, S., & Hayes, T. L. (2009). Fixed- versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, 62, 97-128. doi:10.1348/000711007X255327
- Silva, G. A., & Otta, E. (2014). Revisão sistemática e meta-análise de estudos observacionais em psicologia. *Revista Costarricense de Psicología*, 33(2), 137-153. Recuperado de <http://rcps-cr.org/>
- Su, N., Buchin, Z. L., & Mulligan, N. W. (2021). Levels of retrieval and the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 47(4), 652-670. doi:10.1037/xlm0000962
- Valentini, F., Mourão, L., & Franco, V. R. (2018). Modelos latentes e slopes randômicos para análise de moderação e mediação. *Análise Psicológica*, 17(4), 439-450. doi:10.15689/ap.2018.1704.4.04

Recebido em: 24/05/2021

Reformulado em: 20/12/2021

Aprovado em: 31/01/2022

Nota dos autores:

Os autores não possuem conflitos de interesse a declarar. O estudo foi apoiado pelo Edital N° 0004/2021 da Fundação Universidade de Brasília, Universidade de Brasília, Decanato de Pós-Graduação (FUB/UnB/DPG). Agradecemos a Lara Borges Larroyed, Carlos Biagolini-Jr e João Gonçalves Moreira Neto pelos comentários e sugestões.

Sobre os autores:

**Marcos Felipe Rodrigues de Lima** é doutorando em Ciências do Comportamento pela Universidade de Brasília. Membro do Laboratório de Aprendizagem, Memória e Neurocognição.

ORCID: <https://orcid.org/0000-0002-5922-2543>

*E-mail:* lima.piraju@gmail.com

**Luciano Grütner Buratto** é professor adjunto no Departamento de Processos Psicológicos Básicos (Instituto de Psicologia) da Universidade de Brasília. Coordenador do Laboratório de Aprendizagem, Memória e Neurocognição. Doutor em Psicologia Experimental pela Universidade de Warwick (2008; Reino Unido).

ORCID: <https://orcid.org/0000-0002-7003-7824>

*E-mail:* lburatto@unb.br

Contato com os autores:

Departamento de Processos Psicológicos Básicos, Instituto de Psicologia, Universidade de Brasília, Campus Darcy Ribeiro, ICC Sul, Sala ASS-012/5

Brasília-DF, Brasil

CEP: 70910-900