Emerging Adulthood
2016, Vol. 4(1) 3-18

DOI: 10.1177/2167696815617076
ea.sagepub.com

Systematic Reviews With Meta-Analysis: Why, When, and How?

Elisabetta Crocetti[1]

Abstract

Systematic reviews with meta-analysis represent the gold standard for conducting reliable and transparent reviews of the literature. The purpose of this article is threefold: (a) to address why and when it is worthwhile to conduct a systematic review with meta-analysis, covering advantages of this approach in the context of the statistics reform in the behavioral sciences; (b) to explain how to conduct and publish a systematic review with meta-analysis, describe the main steps, and suggest best practices for each of them; and (c) to discuss the relevance of conducting a systematic review with meta-analysis for the emerging adulthood field, suggesting how this approach can be applied to address research questions about the specificity of this period. In addressing these issues, a fictitious systematic review with meta-analysis aimed at examining gender differences in the view of emerging adulthood as a period of exploration, instability, self-focus, feeling-in-between, and possibilities is presented. Furthermore, individual participant data systematic review with meta-analysis is proposed as an important future direction for conducting reviews within the social sciences.

Keywords

systematic review, meta-analysis, moderators, effect sizes, heterogeneity, publication bias, IPD

Systematic reviews with meta-analysis represent the gold standard for conducting reliable and transparent reviews of the literature. Specifically, a *systematic review* (or research synthesis; Cooper, Hedges, & Valentine, 2009) is a review of a clearly formulated question that uses systematic and explicit methods to identify, select, and critically appraise relevant research, and to collect and analyze data from the studies that are included in the review (Higgins & Green, 2011). *Meta-analysis* refers to the use of statistical techniques to synthesize results across multiple primary studies (Hunt, 1997). It is important to note that systematic reviews and meta-analysis can be conducted independently from each other. Indeed, a systematic review can be done without including a statistical synthesis of the results, and a meta-analysis can be applied to data not retrieved within a systematic review. In this article, it is emphasized that a best practice is to combine the advantages of systematic reviews and meta-analysis in order to provide more sophisticated and advanced reviews of a certain field. Thus, this article describes the multiple sequential steps required for conducting, according to the current state-of-art, high-quality *systematic reviews with meta-analysis*.

In line with these considerations, the purpose of this article is threefold. First, it is addressed why and when it is worthwhile to conduct a systematic review with meta-analysis, covering advantages of this approach in the context of the statistics reform in the behavioral sciences (Cumming, 2012; Kline, 2013). Second, it is explained how to conduct and publish a systematic review with meta-analysis, describing the main steps, suggesting best practices for each of them, and providing practical examples. Third, it is discussed the relevance of a systematic review with meta-analysis for the emerging adulthood field, explaining how this approach could be applied to research questions about the specificity of emerging adulthood, as compared to other periods of the life span, and about cross-cultural generalizability of topics relevant for the study of this period.

Why and When to Conduct a Systematic Review With Meta-Analysis

Why and when is it important to conduct a systematic review with meta-analysis? These related questions can be addressed by considering the strengths of this methodology. In fact, a systematic review with meta-analysis is considered to be the gold standard for conducting reliable and trustworthy synthesis of

[1]Research Centre Adolescent Development, Utrecht University, Utrecht, the Netherlands

Corresponding Author:
Elisabetta Crocetti, PhD, Research Centre Adolescent Development, Utrecht University, Martinus J. Langeveldgebouw, Heidelberglaan 1, 3584CS Utrecht, the Netherlands.
Email: e.crocetti@uu.nl

available evidence in one area of study, being superior to alternative ways of performing literature reviews. In particular, a systematic review with meta-analysis is highly preferable to conducting *narrative reviews* since a systematic review with meta-analysis implies higher transparency in study selection and in the weight assigned to each study, allows managing a large number of studies, and ensures the replicability of the conclusions that are drawn (Bushman & Wells, 2001). Additionally, a systematic review with meta-analysis is superior to the *vote-counting* method used to synthesize available studies focusing solely on the statistical significance of the results (Cooper, 2010). Specifically, in the vote-counting method, conclusions are based on counts of statistically significant results in accordance with hypothesis (positive results) versus nonstatistically significant results (null results) and statistically significant results in contrast with hypothesis (negative results). When the number of positive results is higher than the number of null or negative results, then it is concluded that the evidence supporting the hypothesis is stronger than the evidence against it. The main shortcoming of this method is its strong reliance on the results of the test of statistical significance that, as further discussed below, should not be used as the only criterion for drawing substantial conclusions (Cumming, 2012; Kline, 2013).

A systematic review with meta-analysis can be conducted to summarize and critically evaluate both inconsistent and consistent literature and be performed with a small, medium, or large number of studies. In fact, aggregation of two or more studies increases the precision of estimates and the confidence about the effect being studied (Cumming, 2012). By means of a systematic review with meta-analysis, several research questions can be addressed. In fact, scholars can perform reviews to address relevant theoretical research questions (e.g., how does personality develop over time; Roberts, Walton, & Viechtbauer, 2006) and methodological issues (e.g., which is the overall reliability of a certain instrument; Hale, Crocetti, Raaijmakers, & Meeus, 2011). In addition, another aspect addressed by several systematic reviews with meta-analysis is the efficacy of interventions and treatments (e.g., is a certain psychosocial, or medical, intervention effective; Campbell Collaboration, 2014; Higgins & Green, 2011), and the analysis of factors that can increase treatment effectiveness (e.g., can social support increase adherence to treatments; Magrin et al., 2015). In addition, systematic reviews with meta-analysis provide a context to test which factors (moderators) can explain differences in the magnitude of the effect being observed. Doing so, it is possible to thoughtfully disentangle factors that might have accounted for inconsistent findings reported in the literature or individuate a number of conditions that, also within the context of a literature that appears to be consistent, could explain an amplification or a reduction in the effect under investigation.

In this article, a fictitious systematic review with meta-analysis will be presented as an example. The aim of this review was to examine gender differences in the view of emerging adulthood as a period of exploration, instability, self-focus, feeling-in-between, and possibilities. Second, potential factors (moderators) that could magnify or attenuate these gender differences will be taken into consideration.

The main strengths of systematic reviews with meta-analysis can be further discussed considering the strong synergies between the meta-analytic approach and the statistics reform occurring in the social sciences (Cumming, 2012; Kline, 2013). The statistics reform has pointed to the shortcomings of the predominant use of the test of statistical significance for drawing conclusions about any research area. In fact, the test of statistical significance addresses only one question: "Is there an effect?" (the null hypothesis is the hypothesis of no effect) and can fail in providing meaningful answers for two main reasons. With small sample sizes, the test of statistical significance can fail in finding significant results because of issues of statistical power. In contrast, with large sample sizes, the test of statistical significance can easily yield significant findings but does not give any information about their dimension and clinical/practical relevance. For instance, with a large sample size, it is possible to obtain a difference between two groups (e.g., male and female young adults) that is statistically significant but explains less than 1% of the variance in the variable being compared. This situation clearly points to the gap that might exist between statistical significance and practical significance (Ellis, 2010).

In order to overcome these criticisms of the test of statistical significance, the statistics reform proposes the use of the new statistics (Cumming, 2012). The new statistics include effect sizes, confidence intervals, and meta-analysis itself. As can be easily seen, these statistics are not new, but it is their use that is innovative since they are proposed as a new way for drawing conclusions about research findings. In other words, scholars should not base their conclusions solely on results of the test of statistical significance but should instead focus more on the dimensions of effects being studied (this information is provided by the effect sizes) and on their precisions (this information is provided by the confidence intervals). Thus, the new statistics promote a shift from a dichotomous thinking (significant vs. nonsignificant) to an estimation thinking (focused on the dimension and precision of the effects under consideration; Cumming, 2014).

In this context, the meta-analytic section of a systematic review represents a robust method to gain more confidence in both the dimension and the precision of the effects under consideration. Indeed, meta-analysis provides an overall effect size and a confidence interval that is not based on a single study, but on cumulative evidence, yielded from the combination of two or more studies (Cumming, 2012). Thus, the statistics reform has further highlighted the importance of meta-analysis.

How to Conduct a Systematic Review With Meta-Analysis: Sequential Steps and Best Practices

Conducting a systematic review with meta-analysis requires following multiple steps, from the definition of the research

questions being addressed until the publication of the results (Crocetti, 2015). In this section, all the steps necessary for conducting a high-quality systematic review with meta-analysis are explained in light of the current state-of-art. In order to make this part as practical as possible, each step is presented providing an example represented by a fictitious systematic review with meta-analysis, conceived for instructional purposes.

Defining the Research Question

The first step for conducting a systematic review with meta-analysis is to define the object of the review and the research question being addressed. In this step, a number of issues should be taken into account. In particular, the aim of a systematic review with meta-analysis should be rooted in a clear theoretical background. This is a prerequisite for avoiding "fishing" temptations and committing a mistake that can invalidate the entire process—that is, mixing "apples and oranges" (Lipsey & Wilson, 2001). This means that it is mandatory to ascertain that the concepts investigated through a systematic review with meta-analysis are conceptually comparable.

In the fictitious example presented throughout this section, the object of our systematic review with meta-analysis is to examine an aspect relevant for the emerging adulthood field. A main interest for scholars investigating emerging adulthood is to unravel how young people perceive this period of the life span. According to Arnett's (2000, 2004) theory, emerging adulthood can be considered to be the age of identity exploration, the age of instability, a self-focused period of the life stage, the age of feeling in-between, and the age of possibilities. Scholars have been interested in studying to what extent young people endorse this view of emerging adulthood and which factors can explain differences in this perception (e.g., Crocetti et al., 2015; Negru, 2012; Reifman, Arnett, & Colwell, 2007). Within this theoretical background, the aim of our fictitious systematic review with meta-analysis is to examine gender differences in the view of emerging adulthood as a period of exploration, instability, self-focus, feeling-in-between, and possibilities. Thus, our research question is "Do young females and males have a similar view of emerging adulthood?"

Specifying Inclusion and Exclusion Criteria

The second step consists of specifying inclusion and exclusion criteria. These criteria define which studies will be eligible for inclusion in the systematic review with meta-analysis. In this step, it is very important that the researcher's choices are explicit and motivated. Importantly, inclusion and exclusion criteria should logically flow from the definition of the research question (Step 1) and, thus, being rooted on a clear theoretical background. The eligibility criteria can be grouped into two main classes: eligibility criteria referring to the characteristics of the study and those referring to the characteristics of the publication.

Eligibility criteria referring to the characteristics of the study regard issues concerning the population (e.g., young people, healthy individuals, and minority groups), the effect and the related variables (e.g., anxiety and depression), and the study design (e.g., cross sectional, longitudinal, and experimental) of interest (Liberati, Altman. Tetzlaff, Mulrow, Gøtzsche, et al. 2009). Eligibility criteria referring to the characteristics of the publication concern the year (e.g., studies are eligible for inclusion if published after a certain year), the language (e.g., studies are included only if published in English), and the type (e.g., journal articles, book chapters, books, PhD dissertation, research reports, conference presentations, and unpublished manuscripts) of publication. Choices regarding the type of publication are related to the treatment of the gray literature (i.e., unpublished studies and/or studies that cannot be easily retrieved; Rothstein & Hopewell, 2009).

The choices regarding the treatment of gray literature represent a critical issue. A common practice in a systematic review with meta-analysis is to only include studies published in peer-reviewed English journals. This strategy is proposed as a way to "enhance the methodological rigor of the studies examined and the conclusions drawn" (Northouse, Katapodi, Song, Zhang, & Mood, 2010, p. 318), because it includes all studies subjected to the peer-review process and selected for publication. Although this practice is common, it has serious shortcomings, leading to a potential overestimation of the effect being studied. In fact, the gray literature might represent a portion of the literature that differs systematically from the published one. This phenomenon, known as publication bias, is strongly driven by the statistical significance obtained in primary studies: Studies with statistically significant results in accordance with hypotheses are more likely to be published in English journals, while studies with nonsignificant results or results statistically significant but contrary to expectations are more likely to remain unpublished or published in less visible outlets (e.g., non-English journals that are not indexed in main bibliographic databases). In order to overcome this criticism, main organizations whose mission is to promote evidence-informed decision making by producing regularly high-quality and accessible systematic reviews encourage the inclusion of the gray literature (e.g., Campbell Collaboration, 2014; Centre for Reviews and Dissemination, 2009), although this choice should be done with caution to avoid further bias. Importantly, systematic reviews with meta-analysis offer two main instruments to examine differences between studies published in English peer-reviewed journals and the gray literature. First, moderator analyses, as further explained below, allow for testing differences in effect sizes and can be used to statistically test whether study results published in more visible outlets differ systematically from study results unpublished or published in less visible outlets. Second, an evaluation of the study quality (Valentine, 2009) can be conducted for all studies included in the review and potential differences between studies published in different outlets or unpublished studies can be detected and controlled in further steps (e.g., conducting sensitivity analyses). Thus, the adoption of inclusion criteria aimed at including also the gray literature in combination with in-depth analyses of differences between different types of literature allow for a comprehensive review of a certain topic.

For what concern the year of publication, it is possible to choose between two options. First, do not specify any criteria related to the year of publication and search for all pertinent literature published at any time on the topic of the review. Second, to select literature published starting from a certain year. This second option can be used when (a) there is a clear starting point (e.g., the year in which a seminal contribution was published to introduce a new theory or model) or (b) the aim is to update a previous review. In this latter case, the meta-analyst conducting the updating might limit the search to all studies published in the time frame not covered by the previous review (this exact information can be found in the method section of the previous review, in which it would be specified in which month and year the literature search was conducted).

In our hypothetical example, we can set the following eligibility criteria: Primary studies must be conducted with emerging adults (aged 18–29 years), report gender differences on the five dimensions of emerging adulthood proposed by Arnett (2004), and be published after 2000 (i.e., the year in which the paper recognized as a milestone of emerging adulthood theory was published; Arnett, 2000) in journal articles published in any language.

Searching the Literature

The third step consists of searching the literature. In order to conduct a comprehensive search of all available primary studies, a good practice is to employ multiple search strategies (e.g., computerized database search, searching indexes of journals, searching in reference lists, and search strategies for the gray literature; Petticrew & Roberts, 2006) on the topic of the systematic review with meta-analysis.

The first and most relevant search strategy involves searching a number of electronic reference databases (e.g., PsycInfo, PsycArticles, ERIC, Pubmed, EMBASE, Scopus, and Web of Science). According to the topic of the review, the researcher can identify the databases specific for the area under investigation (e.g., PsycInfo for a review concerning a psychological topic; PsycInfo and Pubmed for a review concerning a topic from the psychiatric area). The search in these specific databases can be integrated with the search in multidisciplinary databases (e.g., Scopus and Web of Science). After selecting which databases to use, the researcher can decide which search strategy to employ. Commonly, a search for key words is conducted. However, this search can be combined with additional searches (e.g., searching for all the publications of leading authors in the field and searching for all the works that cite a certain publication).

A further search strategy consists of searching the websites of journals deemed most likely to publish studies on the topic of the systematic review with meta-analysis. This strategy is aimed at identifying pertinent ahead-of-print (online first) articles and checking the most recent issues of each journal to retrieve potential recent publications that could not yet be available in the electronic databases. The options ''Analyze results'' available in the multidisciplinary databases Scopus and Web of Science can be used to identify the list of 10 of the 15 journals on which most papers regarding the topic of the review have been published. Additional search strategies include checking the reference lists of (a) reviews and/or theoretical papers, available on the same topic, and/or (b) selected primary studies.

All the search strategies described above can be used (all together or in various combinations) to search for journal articles. When the researcher is interested in including gray literature, the above strategies should be supplemented with additional searches specific for retrieving the gray literature (Rothstein & Hopewell, 2009). These additional steps include searching in dissertation abstract databases, in conference programs, contacting experts in the field and/or members of scientific associations, and so on. These strategies aimed at finding the gray literature have the advantage of aiding in conducting a more comprehensive literature search (identifying not only published journal articles but also conference presentations, dissertation, unpublished materials that can be supplied by authors, etc.) but have the shortcomings of being time consuming and of potentially inducing biases that can threaten the replicability of this step (e.g., rates of authors' responses to a request for unpublished materials can depend on the status of the researcher conducting the review).

For our hypothetical systematic review with meta-analysis, we can use the following search strategies. First, we will search the electronic databases PsycINFO, ERIC, Scopus, and Web of Science using the key words (emerging adult*) and (dimension* or perception* or view*). Second, we will search in Web of Science for all the articles that cited Arnett's (2000) paper. Third, we will search in the website of journals deemed most likely to publish studies on emerging adulthood if they have additional relevant papers among articles ahead-of-print and published in the last two issues. Fourth, we will search in the references of reviews on emerging adulthood.

A good practice consists in reporting detailed information about the results of this step in the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram (Moher, Liberati, Tetzlaff, Altman, and The PRISMA Group, 2009). This diagram is part of the PRISMA guidelines (cf. paragraph Publishing a Systematic Review with Meta-Analysis). An example, filled with fictitious data regarding our hypothetical systematic review with meta-analysis, is reported in Figure 1.

Selecting Primary Studies

The fourth step consists of selecting primary studies. This step implies multiple subphases that should be all documented in the PRISMA flow diagram (Moher et al., 2009; see Figure 1). First of all, duplicates (i.e., the same reference retrieved from multiple search strategies) can be identified and deleted. Second, the remaining references are screened by checking their title and abstract. If they could potentially match the eligibility criteria they are retained, otherwise they are excluded. Third, the retained references are assessed in the full text. In

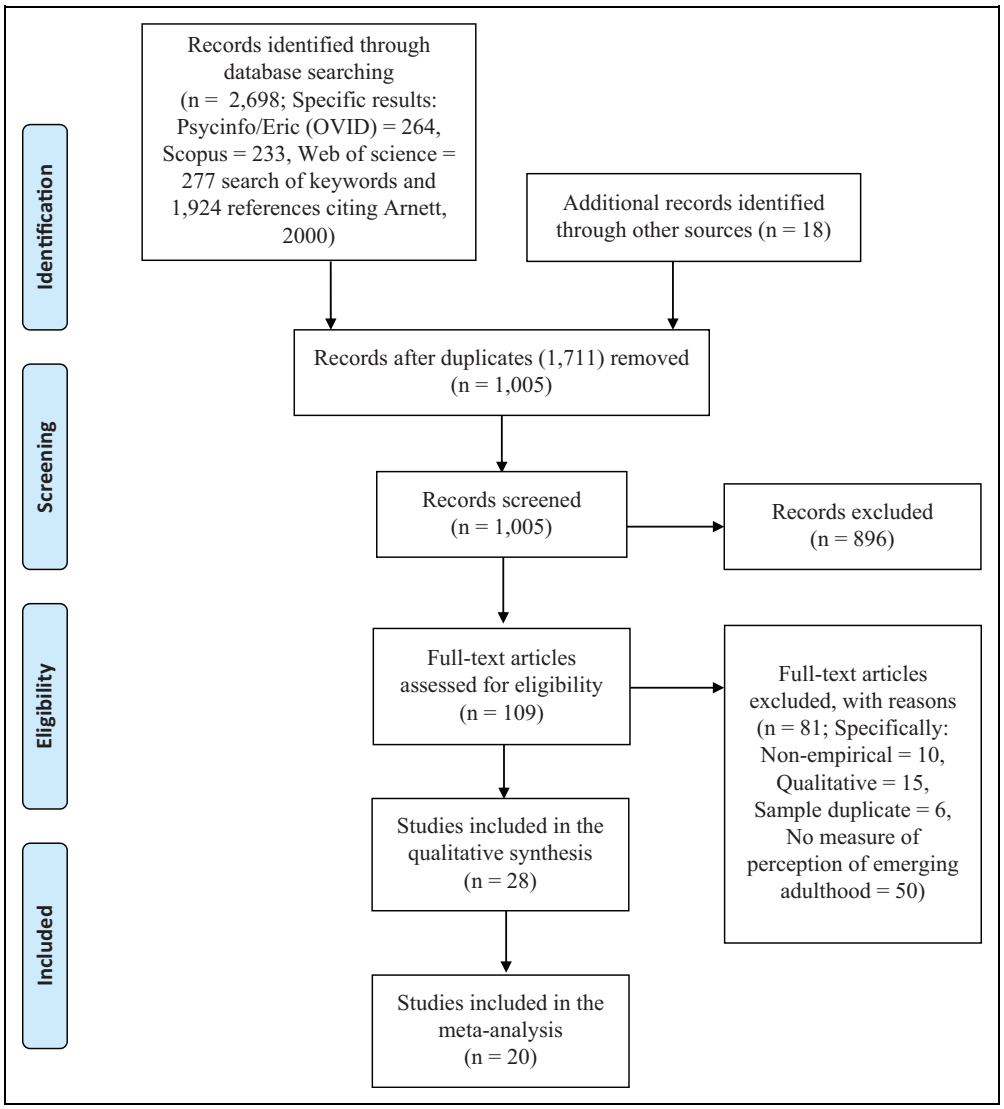

Figure 1. Example of a Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram.

this phase, if the article matches the eligibility criteria, it is included in the systematic review, otherwise it is excluded and the exclusion reason should be specified. Articles included in the systematic review can be further included in the meta-analysis if they report data required for statistical computations. To facilitate navigation through these phases, the researcher can benefit from using a reference manager (e.g., Endnote) to save search and selection results.

During the entire phase of selecting primary studies, a good practice to enhance the reliability of the selection process is to compute interrater reliability. In order to compute interrater reliability, two or more researchers independently evaluate the same number of references. Then, a percentage of agreement or a Cohen's κ is calculated on the basis of the number of agreements (both raters agree on including or excluding a publication) and disagreements (one rater would include a publication that the second rater would exclude and vice versa). Values of the Cohen's κ higher than .60 are considered acceptable and higher than .80 are very good (Landis & Koch, 1977). A best practice is to compute interrater reliability at each step (for the screening of title and abstract and for the evaluation of the full text) considering the total amount of references. However, when there is a large number of references, the authors may consider the possibility of doing it for a subsample of these references (usually 20–25% of the total).[1]

An example of this step is reported in the PRISMA diagram documenting the results of our hypothetical meta-analysis (see Figure 1). As can be seen, most references are usually excluded during the screening process (on the basis of title and abstract), drastically reducing the number of full-text references that are assessed.

It is important to further specify how to manage papers that match the eligibility criteria but do not report data for effect size computations. Three options are available. First, these papers can be excluded. In this case, when references are

checked in the full text, it is specified that these papers are excluded since they do not report data for effect size computations. Second, these papers can be included in the systematic review but excluded from the statistical synthesis. This option implies that in the PRISMA diagram, it is specified that the number of papers included in the systematic review differs from the number of papers included in the quantitative review (meta-analysis). In this case, information that can be extracted from these papers can be reported in a table descripting main study characteristics (e.g., characteristics of participants and methods), and the quality, objectives, and main conclusions of these studies can be reviewed in the text. Third, study authors can be contacted to obtain data that are not reported in the publication. If authors reply providing the requested data, then it is possible to include the paper in the statistical analyses. In contrast, when authors do not reply (e.g., they are retired) or they are not able to provide missing data (e.g., they do not have access anymore to the data file), then it is possible to choose between the two options described above (i.e., to exclude the paper or to include it in the systematic review but not in the statistical analyses).

Coding Primary Studies

After having selected the studies that would be included in the systematic review with meta-analysis, the researcher can code them. Coding is the procedure by which primary studies are examined in order to find relevant data. This step can be conducted by means of a coding protocol, detailing which data should be extracted from each study and how they should be coded (Cooper, 2010).

Also in this step, a good practice is to compute interrater reliability as explained above for the selecting primary studies phase. Thus, during the coding, two or more researchers code each study independently and they then compare their rate of agreement, reporting it as a percentage (i.e., percentage of data extracted in the same way). When disagreements occur, they can be resolved through discussion or involving a third expert in the coding procedure. Thus, reliability should be ideally computed in two phases: when selecting primary studies and when coding them.

Data coded from each primary study can be grouped into three categories: (a) characteristics of the study (e.g., age of the sample, type of design, and measures being used), (b) characteristics of the publication (e.g., year, language, and type), and (c) data for effect size computations. Importantly, a study can only be included in the statistical analyses if it reports data needed for effect size computations (Category c). As alluded to above, when these data are not available the researcher can decide to contact the study author(s) for obtaining the missing data, exclude the study from meta-analysis, since data for effect size computations are missing, or include it in the systematic review but not in the meta-analysis. In contrast, when missing data belong to Category (a) or (b), the study can be still included in meta-analysis.

During the coding, it is also possible to evaluate study quality. This procedure can be conducted by rating each study according to its quality by means of specific checklists. However, this is not an easy task. The first complexity derives from defining what quality means. In fact, study quality is a multifaceted concept that refers to various levels, including internal validity, external validity, construct validity, and statistical conclusion validity (for a discussion of these various forms of validity in the systematic review context, see Valentine, 2009). Thus, it is necessary that researchers agree, for each type of research design (e.g., experimental studies and observational studies), which aspects of validity are more important. The second complexity is a direct consequence of the first one: since study quality is a multifaceted concept that can be defined differently across disciplines and research designs several checklists exist (for a review, see Petticrew & Roberts, 2006). Each checklist has its own items and scoring criteria for evaluating quality. Importantly, it has been found that applying different checklists to the same studies included in a systematic review lead to different conclusions (Valentine, 2009). In line with these considerations, some good practice include the following aspects: (a) if the evaluation of study quality is conducted, it is important to find which can be the best tool for doing it (this can vary across disciplines and research designs); (b) it is recommended not using the results of study quality assessment to exclude studies, since this might drive to misleading conclusions; rather, results of study quality assessment can be used in moderator analyses to test how results are affected by quality. Doing so, it is possible to derive a comprehensive picture and detect whether lower quality studies found effects that differed systematically from higher quality studies (e.g., underestimating or overestimating the size of the effect under investigation).

In our hypothetical systematic review with meta-analysis, examples of information that can be coded from primary studies include (a) characteristics of the study (e.g., sample age: mean, *SD*, and range; gender composition: % girls; ethnic/racial composition: % ethnic minorities; occupation of participants: % university students, % employed, % self-employed, and % unemployed; family socioeconomic status: average family income, % fathers employed, % mothers employed, paternal educational level, and maternal educational level; context: country in which the study has been conducted; type of design: cross-sectional vs. longitudinal; measure used to assess dimensions of emerging adulthood: name of the scale, number of items, and Cronbach's α for each dimension), (b) characteristics of the publication (year of publication and language of publication), and (c) data for effect size computations (more information about this type of data are reported in the following paragraph).

From this step onward, we will present, for the sake of brevity, results regarding only one dimension of emerging adulthood (the age of identity exploration; Arnett, 2000, 2004). It is important to note that the example we are discussing shows a situation that occurs very frequently in the social sciences. To address our research question (are there gender differences in the perception of emerging adulthood?), we should conduct five meta-analyses (one meta-analysis for each dimension of

emerging adulthood) that would be then reported in the same publication. This is an example of a complex meta-analytic database, in which multiple outcomes are taken into account (Borenstein, Hedges, Higgins, & Rothstein, 2009). In the presence of complex databases, a good practice is to adopt an analytic approach (e.g., conducting statistical analyses for each outcome) in order to provide the reader with a comprehensive picture and avoid mixing ''apples and oranges.'' Other forms of complex databases involve multiple subgroups, comparisons, and/or time points within primary studies (Borenstein et al., 2009).

Computing Effect Sizes for Each Study

Data extracted during the coding are used to compute an effect size for each study. The effect size is a measure of the dimension of the effect under investigation and can represent the difference between two groups (e.g., intervention and control groups, gender groups, age-groups, etc.) or the strength of the association between two variables (Ellis, 2010). The most common effect sizes are based on means (main effect sizes include the Cohen's *d*, Hedges' *g*, raw unstandardized difference), binary data (risk ratio, odds ratio, and risk difference), correlations (Pearson's correlations, Fisher's *Z*), and survival data (hazard ratio).

For each effect size, there are different data entry formats that can be used for computational purposes (Borenstein et al., 2009). More specifically, effect sizes of interest can be exactly estimated using certain types of statistics (e.g., Cohen's *d* can be exactly estimated using means, standard deviations, and sample sizes) or can be approximated with varying levels of precision using other types of statistics (e.g., *p* values resulted from the *t*-test reported out to three decimals). For instance, in our example, we are interested in examining gender differences in the perceptions of emerging adulthood. Thus, we are dealing with effect sizes based on mean scores reported by females and males on dimensions of emerging adulthood. So, when a primary study reports means, standard deviations, and sample sizes for each group, those data can be coded and used to compute precisely the effect size for that study. In contrast, when these data are not reported, researchers can look for the best available alternative but should be aware that the effect size that they are computing is less precise. For mean scores, a common alternative (especially when the data necessary for computing the effect size were not reported to address main study goals, rather they were presented in the context of ancillary or preliminary analyses) is to use the statistical significance (*p* value or *p* range; e.g., $p = .004$ or $p < .01$) obtained from the test conducted to compare the mean scores of the two groups of interest. Additionally, when results are reported as nonsignificant with no additional data available, a commonly used conservative approach consists in assigning an effect size equal to zero. In the worst case scenario, when data for computing an effect size are not available at all, the researcher can decide to contact study authors for obtaining missing data or to exclude the study from meta-analysis (but still include it in the systematic review). In this step, the researcher can be facilitated by using software specific for meta-analysis (ProMeta and CMA) that allows selecting among different data entry formats for each study.

In the social sciences, the most used effect sizes are the Cohen's *d* (for quantifying the magnitude of the difference between two groups) and the Pearson's *r* (for reporting the strength of the association between two variables). Although some rules of thumb are used to evaluate these effect sizes (i.e., *d*s of about 0.20 are considered small effects, *d*s of about 0.50 moderate effects, and *d*s of about 0.80 large effects; similarly, correlations of about .10, .25, and .45 are considered small, medium, and large, respectively; Cohen, 1988; Lipsey & Wilson, 2001), these cutoffs should be taken with strong caution. In fact, it is highly recommended that the effect size magnitude should be interpreted within the context of the study topic and methodology. This implies that the same absolute value of the effect size can be considered trivial or meaningful in different contexts (see, e.g., Adachi & Willoughby, 2014, for a discussion of the misleading conclusions that can be derived from rigid applications of effect size cutoffs that do not take into consideration the specifics of the statistical analyses from which they are obtained).

Furthermore, for each study, the computation of the effect size is combined with a measure of its precision. Specifically, for each effect size its variance, standard error, and 95% confidence interval are also computed. As a convention, the statistical significance of the effect size is also reported. However, as discussed in the first part of this article, the researcher should be more interested in the dimension of the effect and in its precision than in its statistical significance (Cumming, 2014).

A good practice is to report effect sizes for each study in a forest plot (Moher et al., 2009). An example of a forest plot based on our hypothetical systematic review with meta-analysis is reported in Figure 2. For each study, the location of the square indicates the effect size and the line represents the confidence interval. Although not reported, from this graph, it is very easy to identify studies with significant results (Cumming, 2012): They are all the studies whose confidence intervals do not include the vertical line corresponding to the null value (in this example, the zero line).

In a complex meta-analytic database, when multiple subgroups, comparisons, time points, and/or outcomes are included within primary studies, it is possible to compute more than one effect size for each study. For instance, in our hypothetical systematic review with meta-analysis, this would mean to compute for each study one effect size for each outcome (i.e., five effect sizes documenting gender differences on views of emerging adulthood as a period of exploration, instability, self-focus, feeling-in-between, and possibilities). A good practice, to avoid mixing ''oranges and apples,'' is to conduct a meta-analysis for each outcome and publish this set of meta-analyses in the same publication. This means that one study will contribute to more than one statistical analysis. In some situations, it might be conceptually possible to combine two or more effect sizes from the same study. In this case,

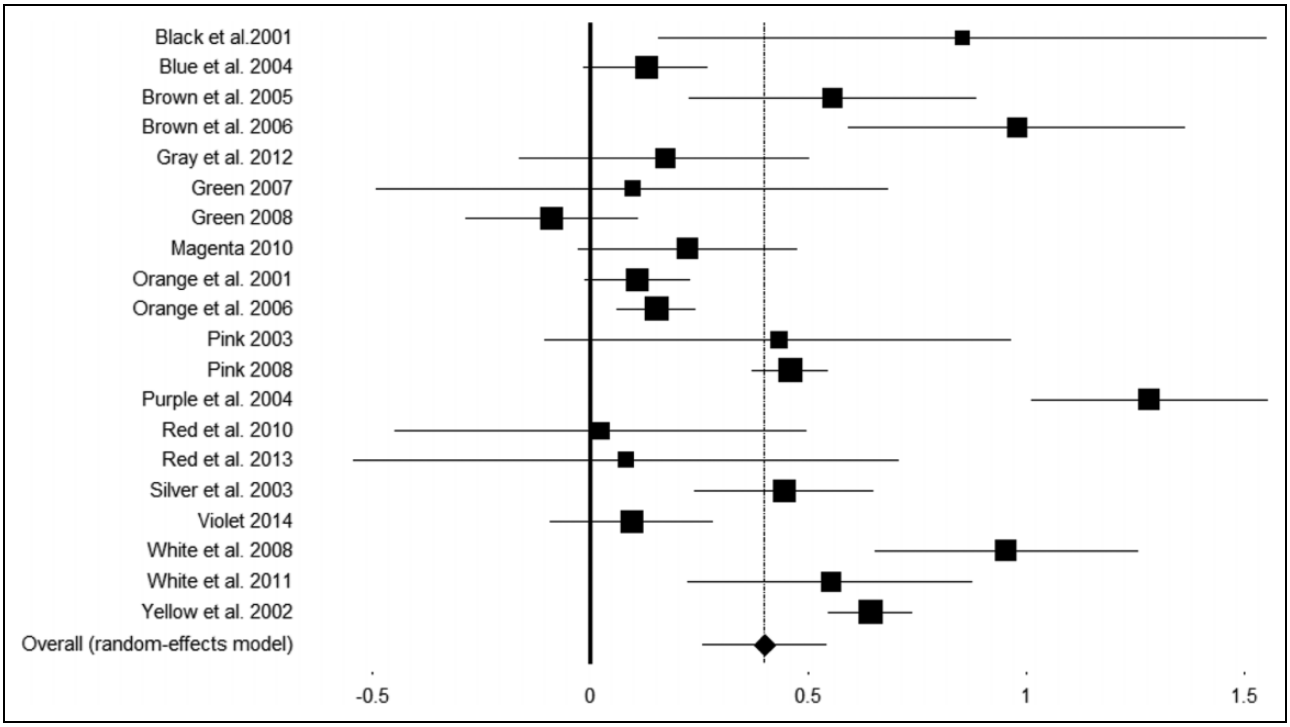

Figure 2. Example of a forest plot. This figure was generated with ProMeta 2.0.

to account for the nonindependence of data originated from the same primary data set, the combined effect size is computed controlling for the correlation among these data (e.g., for the correlation between two outcomes; for a further discussion, see Borenstein et al., 2009).

Combining Effect Sizes

After having computed an effect size for each study, the researcher can obtain the total effect size (Cooper et al., 2009). In order to reach this goal, effect sizes are pooled across studies for obtaining an overall effect size.[2] The inverse-variance method is the approach most commonly used to assign a weight to each study, with the weight being calculated as the inverse of study variance. Two different statistical models can be used to conduct this analysis: the fixed-effect model or the random-effects model (Hedges & Vevea, 1998). The fixed-effect model assumes that there is a true effect size common to all the studies. Thus, in assigning a weight to each study, it takes into account only one source of variance: the within-study variance. In contrast, the random-effects model assumes that the true effects are normally distributed. Thus, in assigning a weight to each study, it takes into account two sources of variance: the within-study variance and the between-studies variance. In line with the current state of the art, most meta-analyses are conducted with the random-effects model because accounting for these different sources of variation among studies (i.e., within-study variance and between-studies variance) is a more conservative approach that also allows generalization of the meta-analytic findings beyond the studies included in the synthesis. The only caveat is that with meta-analyses with few studies (e.g., five or less) analyses should be conducted with both models, since in this case the estimation of the variance between studies is less precise (Borenstein et al., 2009).

The inverse-variance method can be used to summarize effects across studies based on dichotomous or continuous data (Higgins & Green, 2011). An alternative approach that can be applied when effect sizes are all based on dichotomous data is the Mantel–Haenszel method (Mante & Haenszel, 1959). This method is recommended when data are sparse, either in terms of event rates being low or in terms of study size being small. In these conditions, in fact, the estimates of the standard errors that are used in the inverse-variance methods may be poor. To account for this problem, the Mantel–Haenszel method uses a different weighting scheme that depends upon which effect measure (e.g., risk ratio, odds ratio, or risk difference) is being used (Higgins & Green, 2011). So, this method is particularly suited for meta-analyses based on phenomena for which few events are being observed (e.g., a meta-analysis comparing the incidence of a rare disease or an uncommon behavior in two groups).

Results of overall analyses are displayed in the last part of the forest plot and are usually reported in a table. An example based on our fictitious systematic review with meta-analysis is reported in Table 1. In a total of 20 studies, we found moderate gender differences (with females scoring significantly higher than males) in the perception of emerging adulthood as a period of identity exploration. After having obtained an overall effect size, conducting ancillary analyses to check the robustness of study findings is recommended. Specifically, it is a good

Table 1. Example of Summary of Hypothetical Meta-Analytic Results for Gender Differences in Perception of Emerging Adulthood as a Period of Identity Exploration.

					Heterogeneity		Publication Bias			
	k	*N* Females	*N* Males	Cohen's *d* [95% CI]	*Q*	I^2	Fail-Dafe *N*	Begg and Mazumdar's Test	Egger's Test	Trim and Fill
Overall results	20	4,682	5,678	.40*** [.26, .54]	192.78***	90.14	1,220	1.04	0.48	0

Note. k = number of studies; N = total number of participants; Cohen's d = standardized mean difference; CI = confidence interval.
***$p < .001$.

practice to conduct sensitivity analyses (Higgins & Green, 2011). For instance, researchers can compute how the overall effect size would change by removing one study at a time. This analysis is particularly useful for checking the impact of potential effect size outliers (i.e., effect sizes with standardized residuals higher than |2|). If the exclusion of a study does not substantially change the overall effect size, this is an indication of the robustness of the overall results.

Assessing Heterogeneity

The next step, after having computed the overall effect size, consists of evaluating heterogeneity across studies. This requires addressing two questions: (a) is there significant heterogeneity across studies? and (b) how large is this heterogeneity? The first question can be addressed by means of the Q statistic, with a significant Q value indicating significant heterogeneity of results among studies. The second question can be addressed by computing two indices: T^2 and I^2 (Huedo-Medina, Sánchez-Meca, Marín-Martínez, & Botella, 2006). Specifically, T^2 indicates the between-study variance and I^2 estimates the proportion of observed variance that reflects real differences in effect sizes, with values of 25%, 50%, and 75% that might be considered as low, moderate, and high, respectively (Higgins, Thompson, Deeks, & Altman, 2003).

When a meta-analysis has a small heterogeneity, it means that results of primary studies were rather similar and consistent. However, in practice, it is very common to find significant and large heterogeneity. This situation occurs especially when the meta-analysis includes 10 or more studies and each of them has addressed the main research question differently from others (e.g., in a different age-group and/or national sample, using a different measure). The result is that, although all results might be statistically significant, they can yield variations in the sizes of the effects under considerations (which can range from small to large). Thus, it is very common that meta-analyses conducted in different areas highlight significant and large heterogeneity across study findings.

In our example, the results of the heterogeneity analyses are reported in Table 1. As can be seen, these results indicate that findings regarding gender differences on the perception of emerging adulthood as a period of identity exploration were characterized by significant (as indicated by the Q statistic) and large (as indicated by the I^2) heterogeneity.

Conducting Moderator Analyses

Moderators (or predictors) are factors that are assumed to affect the magnitude of the effect sizes across the studies in which these factors are present (Cooper, 2010). Thus, moderator analyses are used to test which factors can explain the heterogeneity of study findings and they are especially suitable to clarify inconsistent results reported in the literature (Viechtbauer, 2007). In other words, the computation of the overall effect size addresses the research question being studied (e.g., are there gender differences in the perception of emerging adulthood?), while the moderator analyses examine which factors can explain the fact that some studies found gender differences whereas some other studies did not detect them.

Moderator analyses include subgroup analysis and meta-regression (Borenstein et al., 2009). Subgroup analyses are conceptually similar to the analyses of variance conducted in primary studies and they can be used to test categorical moderators (with two or more levels). In these analyses, a meta-analysis is performed for each level of the moderator (a good practice is to have at least three studies for each level of the moderator) and then these results are tested for significant differences. Meta-regressions are conceptually similar to the regression analyses used in primary studies and can be used to test both categorical (dummy coded) and/or numerical moderators.

In our example, a categorical moderator represented by the context of the study was tested. It was recoded into three macrogeographical areas: studies conducted in North America, in Europe, and in Asia. Fictitious results indicated a significant effect of the moderator, $Q(2) = 6.84$, $p = .033$, which can be easily interpreted by looking at the effect sizes for each level of this moderator. In fact, gender differences were small in studies conducted in Europe ($k = 8$, N females = 1,027, N males = 1,227, Cohen's $d = .17$ [.01, .33], $p < .05$), whereas they were moderate in studies conducted in North America ($k = 7$, N females = 2,864, N males = 3,416, Cohen's $d = .50$ [.28, .72], $p < .001$) and in Asia ($k = 5$, N females = 791, N males = 1,035, Cohen's $d = .52$ [.14, .90], $p < .01$). It can be tested also if gender differences were moderated by the age of respondents. Results of the meta-regression indicated that age was a significant moderator ($B = .08$, $p = .022$). As displayed in the scatter plot (Figure 3), the magnitude of gender differences in the perception of emerging adulthood increases with age.

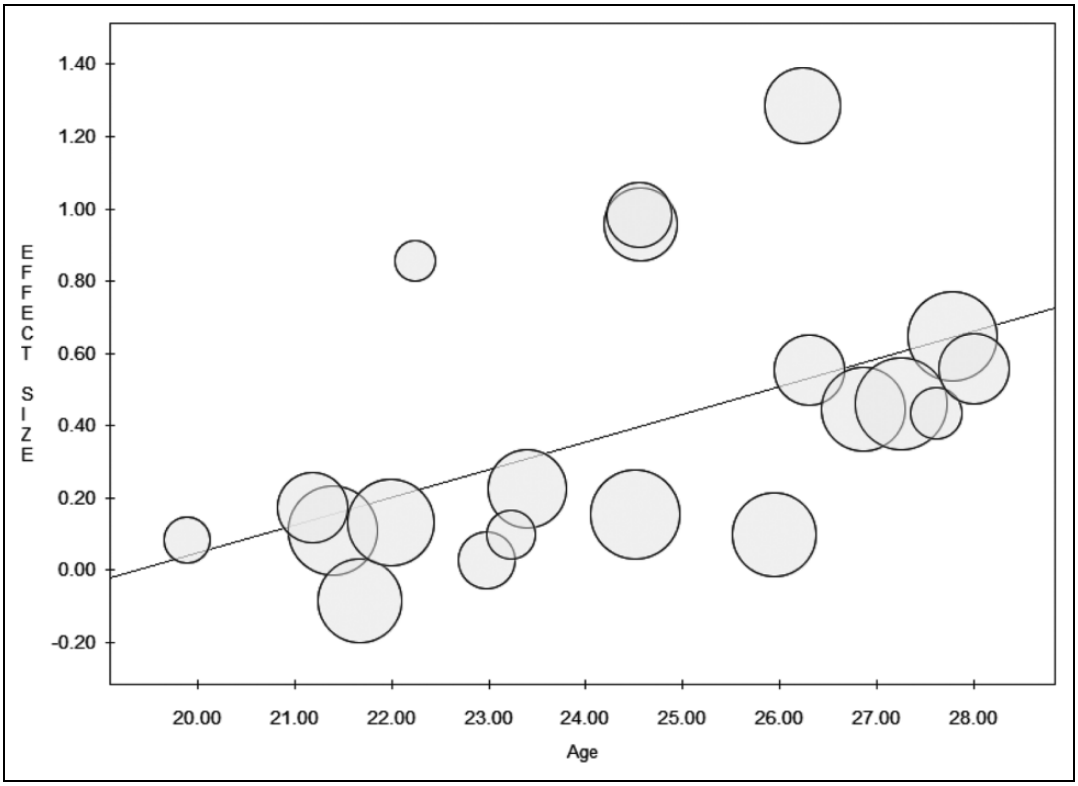

Figure 3. Example of a scatter plot reporting results of a meta-regression.

Assessing Publication Bias

The publication bias refers to the situation that occurs when published studies (those that can be easily retrieved) differ systematically from unpublished studies (gray literature; Rothstein, Sutton, & Borenstein, 2005). There are various evidences documenting that the main reason behind the publication bias is the fact that study results are published or not (or they are published in international journals that can be easily retrieved instead of being published in other channels that are less accessible) on the basis of their statistical significance (Dickersin, 2005). More specifically, results that are statistically significant and in line with hypotheses are more likely to be submitted for publication and to be accepted than study results which are nonsignificant or significant but in contrast with hypotheses (Cooper, DeNeve, & Charlton, 1997; Greenwald, 1975; Krzyzanowska, Pintilie, &Tannock, 2003; Mahoney, 1977).

The publication bias can be a threat for the conclusions of a systematic review with meta-analysis. In fact, these conclusions become less trustworthy when based on a biased literature. For this reason, an important step of meta-analysis is to assess whether the impact of the publication bias is minimal (i.e., studies that have not been included would not change the results of meta-analysis), moderate (i.e., studies that have not been included would change the results in a nonsubstantial way; e.g., a treatment is still effective but to a lesser extent), or large (i.e., studies that have not been included would change the results in a substantial way; e.g., a treatment found to be effective, in reality it is not effective). This evaluation can be performed by means of various methods (Rothstein et al., 2005).

In the social sciences, it is very common to evaluate publication bias by computing the fail-safe N (Rosenthal, 1979). This number is computed when the overall effect size is significant to know how many studies with a nonsignificant result would be required to bring the combined effect size to be nonsignificant. Rosenthal (1979) proposed a fail-safe N higher than $(5k + 10)$ as supporting findings' robustness (where k refers to the number of studies included in meta-analysis). So, for instance, if a meta-analysis includes 20 studies ($k = 20$), the fail-safe N should be higher than 110 ($5 \times 20 + 10 = 110$). Although the fail-safe N is largely used in the social sciences, its application has been subjected to severe criticisms. In fact, major shortcomings of the fail-safe N include the assumption that unpublished and omitted studies would report, on average, a null result (thus, noncontemplating the possibility that these studies might have results that are significant but in contradiction with hypothesis), and the fact that the computation of N does not take into account information about sample sizes and heterogeneity (Becker, 2005). Given these limitations of the fail-safe N, it is recommended to assess publications through multiple and more reliable methods that are now available (e.g., trim and fill approach described below).

A graphical tool for evaluating the publication bias is represented by a funnel plot (Light & Pillemer, 1984). This is a plot of the effect size estimates from individual studies against

some measure of each study's size or precision (Sterne & Egger, 2001). Commonly, in the horizontal axis is reported the effect size and in the vertical axis the standard error (see Figure 4). Each study is represented by a circle. In the absence of bias, this plot should approximately resemble a symmetrical (inverted) funnel. In contrast, in the presence of bias, for example, because smaller studies without statistically significant effects have remained unpublished, a funnel plot appears asymmetrical, with a gap in a bottom corner of the graph. However, it should be noted that publication bias is not the only reason behind the asymmetry of a funnel plot. This asymmetry could be due to small-study effects (Sterne, Becker, & Egger, 2005), occurring when studies with small samples yield effect sizes different from those obtained in studies with large samples. Thus, alternative explanations attributable to small-study effects should be considered before making claims about the presence of publication bias.

Since the visual interpretation of a funnel plot is difficult and risks of being subjective, statistical tests have been developed. The Egger's linear regression method (Egger, Davey Smith, Schneider, & Minder, 1997) and the Begg and Mazumdar's (1994) rank correlation method are two tests of the asymmetry of a funnel plot. In both tests, statistical significant results are indicative of potential publication bias.

More recently, a trim and fill procedure have been proposed. This procedure is a nonparametric statistical technique that evaluates the effect of potential data censoring on the result of the meta-analyses (Duval & Tweedie, 2000). It is an approach consisting of three iterative phases. In the first phase, the asymmetrical part of the funnel plot is trimmed off as affected by publication bias. In the second phase, the average effect size is estimated on the basis of the remained (symmetrical) studies. In the third phase, the trimmed studies are reincluded in the analysis and their symmetric counterparts (based on the new average effect size) are filled. Thus, these iterative phases allow computing an adjusted effect size and its 95% confidence interval. In this method, absence of publication bias is indicated by zero trimmed studies or, in the presence of trimmed studies, by trivial differences between the observed and the adjusted effect sizes (Duval, 2005).

A good practice consists of evaluating the publication through multiple methods. In our hypothetical meta-analysis (see Table 1), the fail-safe N is largely higher than the cutoff ($1{,}220 > 110$; where 110 is the result of $5 \times k + 10$), the nonsignificant Egger's and Begg and Mazumdar's tests, and the zero asymmetric studies detected by the trim and fill method consistently suggest the absence of publication bias.

Publishing a Meta-Analysis

The final step is publishing a high-quality report. In publishing a systematic review with meta-analysis, the researcher should be as detailed as possible. The test of whether or not asystematic review with meta-analysis is well-reported is determined by the possibility of the reader to be able to replicate the entire procedure and obtain the same results. In order to make this possible, transparency and thoroughness are strongly needed.

The researcher should consider some key aspects in deciding where and how publishing a systematic review with meta-analysis. First, it is crucial to identify the right outlet. For a systematic review with meta-analysis, a core information in this respect is provided by the list of primary studies included in the review. If these studies have appeared on a specific set of journals (e.g., three journals focused on developmental psychology), then the same set of journals is likely to be interested in publishing a systematic review with meta-analysis on the same topic. Second, it is necessary to check whether the potential journals accept systematic reviews with meta-analysis. This information can be easily retrieved from the journal website. Regarding this aspect, it is worth noting that some journals allow systematic reviews with meta-analysis to be longer than original studies (e.g., 5,000 words for a regular paper, 10,000 words for a meta-analysis).

After having selected the proper outlet, it is important to write a systematic review with meta-analysis reporting all the relevant information. In this respect, the author is strongly supported by following available guidelines, which provide useful tools for preparing high-quality reports of systematic review with meta-analysis. Most important guidelines include PRISMA (Liberati et al. 2009; Moher et al., 2009), Meta-Analysis Reporting Standards (MARS; American Psychological Association, 2010), and Meta-analysis Of Observational Studies in Epidemiology (MOOSE; Stroup et al., 2000). Importantly, MARS guidelines have been published for the first time in the 6th edition of the Publication Manual of the American Psychological Association (2010) and provide a good tool for psychology researchers. In addition, PRISMA guidelines, although focused on randomized trials, can be used as a basis for reporting meta-analysis of other types of research, particularly evaluations of interventions (Liberati et al., 2009; Moher et al., 2009). They are very useful since they include a number of materials (the Statement, Explanation, Checklist, and Flow diagram freely accessible online at http://www.prisma-statement.org) that guide the researcher step by step.

All these guidelines explain, starting from the title and covering all the sections of the paper (abstract, introduction, method, results, and discussion), how to report a systematic review with meta-analysis. The PRISMA, MARS, and MOOSE checklists can be filled by the author of the review to show that he or she adhered to the guideline and reported all relevant information (some journals might require the submission of the filled checklist). The same checklists can be also be used by journal reviewers and editors to evaluate the quality of the submitted review, suggest revisions, and decide for its acceptance. Thus, adherence to these specific guidelines enhances the quality of the publication of the systematic review and meta-analysis and its chances of being accepted.

Improving Systematic Reviews With Meta-Analysis in the Social Sciences: The Individual Participant Data (IPD) Approach

The IPD approach represents a specific type of systematic review with meta-analysis. The main novelty of this approach

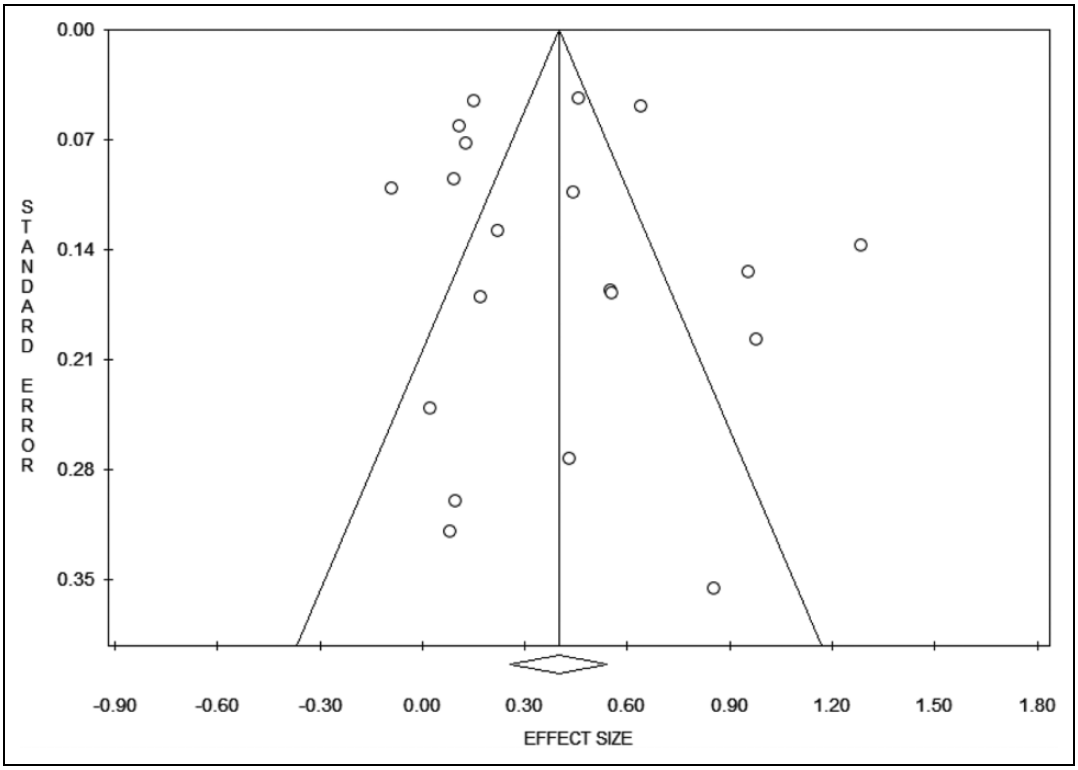

Figure 4. Example of a funnel plot.

is in the process of data extraction. In fact, rather than extracting summary (aggregate) data from study publications (as done in the traditional systematic reviews with meta-analysis), the original research data are sought directly from the researchers responsible for each study. These data can then be screened, reanalyzed centrally, and combined in meta-analyses (Stewart, Tierney, & Clarke, 2011).

The IPD approach has been introduced in the medical sciences (for this reason, IPD can be also be referred to Individual *Patient* Data; Cooper & Patall, 2009). Initially, the IPD approach has been employed primarily in cardiovascular disease and cancer research, where the methodology has been developing since the late 1980s (Stewart et al., 2011). More recently, this approach has also been used in systematic reviews in a number of other medical fields, such as research on Alzheimer disease, epilepsy, malaria, and HIV infection (Simmonds et al., 2005).

Is the IPD approach also employed in systematic reviews with meta-analysis conducted in the social sciences? A search in the bibliographic database Scopus revealed that 476 references with (''Individual Patient Data'' or ''Individual participant data'') and (meta-analy* or ''systematic review*'') in the Title were indexed in the Health Science section, whereas only 13 references were included in the Social Sciences and Humanities section (search conducted on March 29, 2015). A further screening of these 13 references indicated that actually only two of them (Clark, Mackay, & Holmes, 2015; Cuijpers et al., 2014) applied the IPD approach to social sciences topics, while the other articles had a methodological or medical focus.

Applications of the IPD approach can strongly improve the quality of systematic reviews with meta-analysis conducted in the social sciences. In fact, as this approach is considered a ''gold standard'' in medicine (Stewart et al., 2011), it could achieve a similar status within social sciences. In this article, advantages of IPD are discussed, along with its shortcomings, in order to broaden the understanding of the potential of this approach.

IPD systematic reviews with meta-analysis follow the same steps described in the current article with the main difference being that researchers extract data from primary (published and unpublished) studies directly from the study data files instead of coding them the study publications (Stewart & Tierney, 2002). In order to make this possible, IPD systematic reviews with meta-analysis should be conducted and published by collaborative groups, which include the project team managing the review and the researchers who contribute their study data for reanalysis. Both the advantages and disadvantages of the IPD approach are derived from this main difference.

A main advantage of the IPD approach includes the possibility of directly analyzing primary data and, in doing so, producing more reliable results (Stewart & Tierney 2002; Stewart et al., 2011). In fact, collecting, checking, and reanalyzing original data from all studies improve data quantity (both published and unpublished studies can be included) and quality, reducing the risk of bias. More specifically, participant-level data also allow more comprehensive, flexible,[3] and appropriate analyses, solving problems related to missing data (i.e., data included in

the study but not reported in the paper if not statistically significant), or data reported in different formats. Additionally, it is possible to perform analyses stratified by relevant factors (e.g., by gender, or by socioeconomic status) and to test for potential interactions with enough statistical power. When IPD are only available from some studies, then a good practice is to compare results based on IPD with results based on published reports. In this way, it is possible to ascertain whether and to what extent effects detected from these two sources are comparable. If this is the case, then the meta-analytic results using IPD are likely to be representative of all eligible participants (Stewart et al., 2011).

Another important advantage of the IPD approach is the possibility to discuss and interpret results within the collaborative group that consists of both the managing team and the researchers responsible for primary studies (Stewart & Tierney 2002). This step offers the opportunity to critically reflect and appraise the state of a field, validate conclusions, and provide a comprehensive overview of current shortcomings and future directions for theory and practice. This collaborative activity also has the potential benefit of strengthening network ties and provide a basis for future collaborations on primary research and grant applications.

The main disadvantages concern the organizational structure required to handle this type of review. In fact, IPD systematic reviews with meta-analysis are usually more time consuming than traditional systematic reviews with meta-analysis of published data (Riley, Lambert, & Abo-Zaid, 2010; Stewart & Tierney 2002). In particular, the main difference is regarding the amount of time and effort to contact study authors; to establish a collaborative group; to collect, check, and merge data files; and to organize a collaborators meeting to discuss results. However, it should be noted that nowadays these organizational aspects represent less and less of a barrier. First, researchers can take full advantage of the networking opportunities offered by membership in scientific societies. For instance, the Society for the Study of Emerging Adulthood launched the Topic Networks with the aim of bringing together researchers who have common interests in a specific area of emerging adulthood research or practice. Thus, these Topic Networks can provide an optimal context for forming a collaborative group that jointly works on a systematic review and meta-analysis. Second, the Information Communication Technology (ICT) facilities now available for communication (e.g., e-mails and online meetings) and data sharing strongly facilitate handling these organizational aspects.

In sum, although IPD systematic reviews with meta-analysis might be more time consuming than traditional systematic reviews with meta-analysis, they have various advantages and nowadays several network opportunities are available for establishing the collaborative team that will work on it. As for what this means for the study of emerging adulthood, the IPD systematic reviews with meta-analysis can provide a more reliable answer to several theoretical and methodological questions, such as "How do perceptions of emerging adulthood and/or criteria for adulthood vary within and across cultures?," "What is the average reliability of self-report scales most commonly used in this field?," and "Is an instrument more reliable in certain groups than in others?" Hopefully, the IPD approach will also spread within the social sciences similarly to how it has spread in the medical field where it is considered as the new gold standard for systematic reviews with meta-analysis.

Conclusion: Refining the Understanding of Emerging Adulthood Through Systematic Reviews With Meta-Analysis

In concluding this article, it is worthwhile to discuss how systematic reviews with meta-analysis can contribute to the advancement of the emerging adulthood field. First, systematic reviews with meta-analysis can be applied to examine aspects specific for the emerging adulthood period and to test the cross-cultural replicability of main study findings. This goal can be achieved in systematic reviews with meta-analysis organized similarly to the fictitious meta-analysis presented in this article. In this context, the period of emerging adulthood is a core inclusion criterion (i.e., only studies focused on emerging adults are included), the effect size estimates the magnitude of an effect relevant for the study of emerging adulthood (e.g., sociodemographic differences in the importance assigned to criteria for adulthood; predictors of mental health; predictors of timing of main life transitions), and the context of the study is used as a moderator (i.e., examining whether the effect under investigation differs across various cultural contexts).

Second, systematic reviews with meta-analysis may advance the understanding of emerging adulthood by clarifying the specificity of this period in a life span perspective. This aim can be achieved by conducting systematic reviews with meta-analysis to investigate relevant effects across a wider age period and using age as a moderator (recoding age in categories covering the main periods of the life span with one category corresponding to emerging adulthood, ages 18–29). Thus, in these reviews, contrary to those suggested above, age is not considered as an inclusion criterion but as a moderator of study results.

In conclusion, systematic reviews with meta-analysis are a powerful approach that can advance our understanding of emerging adulthood in multiple directions. Scholars interested in this fascinating period of the life span can conduct these reviews to address research questions relevant for the study of this age. Hopefully, this article can provide an accessible guideline to conduct high-quality systematic reviews with meta-analysis that would shed new light on emerging adulthood.

Author Contributions

Elisabetta Crocetti contributed to conception, design, and acquisition; drafted the manuscript; critically revised the manuscript; gave final approval; and agrees to be accountable for all aspects of work ensuring integrity and accuracy.

Declaration of Conflicting Interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Notes

1. *Interrater* reliability is the best practice for checking the correctness of the selection process. However, when it not possible to compute it (since only one author can work on it), *intrarater* reliability can be established with the same researcher reevaluating the same references twice (a first time and then after some weeks) and computing his or her degree of agreement between these two evaluations.
2. In this paragraph, only frequentist statistical methods are described. Bayesian statistical models can also be used for meta-analyses (e.g., Burr & Doss, 2005; Sutton & Abrams, 2001; Turner, Jackson, Wei, Thompson, & Higgins, 2015). They are especially useful in overcoming some of the assumptions in frequentist meta-analysis methods (e.g., assuming a normal distribution of effect size parameters).
3. Individual participant data (IPD) analyses can be conducted using a two-step or a one-step approach (e.g., Riley et al., 2010). Most commonly, researchers adopt *a two-step approach*, in which IPD are first analyzed in each separate study independently, by means of a statistical method appropriate for the type of data being analyzed. This step produces aggregate data for each study, which can be then synthesized in the second step using traditional statistical meta-analytic techniques for aggregate data such as those described in this article. In the one-step approach, the individual participant data from all studies are modeled simultaneously while accounting for the clustering of participants within studies.

References

Adachi, P., & Willoughby, T. (2014). Interpreting effect sizes when controlling for stability effects in longitudinal autoregressive models: Implications for psychological science. *European Journal of Developmental Psychology*, *12*, 116–128.

American Psychological Association. (2010). *Publication manual of the American Psychological Association* (6th ed.). Washington, DC: Author.

Arnett, J. J. (2000). Emerging adulthood: A theory of development from the late teens through the twenties. *American Psychologist*, *55*, 469–480.

Arnett, J. J. (2004). *Emerging adulthood: The winding road from the late teens through the twenties*. New York, NY: Oxford University Press.

Becker, B. J. (2005). Failsafe *N* or file-drawer number. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 111–125). Chichester, England: John Wiley.

Begg, C. B., & Mazumdar, M. (1994). Operating characteristics of a rank correlation test for publication bias. *Biometrics*, *50*, 1088–1101.

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, England: John Wiley.

Burr, D., & Doss, H. (2005). A Bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association*, *100*, 242–251.

Bushman, B. J., & Wells, G. L. (2001). Narrative impressions of literature: The availability bias and the corrective properties of meta-analytic approaches. *Personality and Social Psychology Bulletin*, *27*, 1123–1130.

Campbell Collaboration. (2014). *Campbell systematic reviews: Policies and guidelines*. Campbell Systematic reviews (Supplement 1). doi:10.4073/csrs.2014.1

Centre for Reviews and Dissemination. (2009). *Systematic reviews*. York, England: York Publishing Services.

Clark, I. A., Mackay, C. E., & Holmes, E. A. (2015). Low emotional response to traumatic footage is associated with an absence of analogue flashbacks: An individual participant data meta-analysis of 16 trauma film paradigm experiments. *Cognition and Emotion*, *29*, 702–713.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. New York, NY: Academic Press.

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach* (4th ed.). Thousand Oaks, CA: Sage.

Cooper, H., DeNeve, K., & Charlton, K. (1997). Finding the missing science: The fate of studies submitted for review by a human subjects committee. *Psychological Methods*, *2*, 447–452.

Cooper, H., Hedges, L. V., & Valentine, J. (Eds.). (2009). *The handbook of research synthesis* (2nd ed.). New York, NY: Russell Sage.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, *14*, 165–176.

Crocetti, E. (2015). *Rassegne sistematiche, sintesi della ricerca e meta-analisi* [*Systematic reviews, research syntheses, and meta-analysis*]. North Charleston, SC: CreateSpace.

Crocetti, E., Tagliabue, S., Sugimura, K., Nelson, L., Takahashi, A., Niwa, T., . . . Jinno, M. (2015). Perceptions of emerging adulthood: A study with Italian and Japanese university students and young workers. *Emerging Adulthood*, *3*, 229–243.

Cuijpers, P., Weitz, E., Twisk, J., Kuehner, C., Cristea, I., David, D., . . . Hollon, S. D. (2014). Gender as predictor and moderator of outcome in cognitive behavior therapy and pharmacotherapy for adult depression: An ''individual patient data'' meta-analysis. *Depression and Anxiety*, *31*, 941–951.

Cumming, G. (2012). *Understanding the new statistics: Effect sizes, confidence intervals, and meta–analysis*. New York, NY: Routledge.

Cumming, G. (2014). The new statistics: Why and how. *Psychological Science*, *25*, 7–29.

Dickersin, K. (2005). Publication bias: Recognizing the problem, understanding its origins and scope, and preventing harm. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta–analysis* (pp. 127–144). Chichester, England: John Wiley.

Duval, S. (2005). The trim and fill method. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 11–33). Chichester, England: John Wiley.

Duval, S., & Tweedie, R. (2000). A nonparametric 'trim and fill' method of accounting for publication bias in meta-analysis. *Journal of American Statistical Association*, *95*, 89–98.

Egger, M., Davey Smith, G., Schneider, M., & Minder, C. (1997). Bias in meta-analysis detected by a simple, graphical test. *British Medical Journal*, *315*, 629–634.

Ellis, P. D. (2010). *The essential guide to effect sizes: Statistical power, meta-analysis, and the interpretation of research results*. Cambridge, England: Cambridge University Press.

Greenwald, A. G. (1975). Consequences of prejudice against the null hypothesis. *Psychological Bulletin*, *82*, 1–20.

Hale, III, W. W., Crocetti, E., Raaijmakers, Q., & Meeus, W. (2011). A meta-analysis of the cross–cultural psychometric properties of the Screen for Child Anxiety Related Emotional Disorders (SCARED). *Journal of Child Psychology and Psychiatry*, *52*, 80–90.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486–504.

Higgins, J. P. T., & Green, S. (Eds.). (2011). *Cochrane handbook for systematic reviews of interventions*, Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Retrieved from www.cochrane-handbook.org

Higgins, J., Thompson, S. G., Deeks, J. J., & Altman, D. G. (2003). Measuring inconsistency in meta-analysis. *British Medical Journal*, *327*, 557–560.

Huedo-Medina, T. B., Sánchez-Meca, J., Marin-Martinez, F., & Botella, J. (2006). Assessing heterogeneity in meta-analysis: Q Statistic or I^2 index? *Psychological Methods*, *11*, 193–206.

Hunt, M. (1997). *How science takes stock: The story of meta-analysis*. New York, NY: Russell Sage Foundation.

Kline, R. B. (2013). *Beyond significance testing: Statistics reform in the behavioral sciences* (2nd ed.). Washington, DC: American Psychological Association.

Krzyzanowska, M. K., Pintilie, M., & Tannock, I. F. (2003). Factors associated with failure to publish large randomized trials presented at an oncology meeting. *Journal of the American Medical Association*, *290*, 495–501.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, *33*, 159–174.

Liberati, A., Altman, D. G., Tetzlaff, J., Mulrow, C., Gøtzsche, P. C., Ioannidis, J. P. A., . . . 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, 6*, *e1000100*.

Light, R. J., & Pillemer, D. B. (1984). *Summing up: The science of reviewing research*. Cambridge, MA: Harvard University Press.

Lipsey, M., & Wilson, D. (2001). *Practical meta-analysis*. Thousand Oaks, CA: Sage.

Magrin, M. E., D'Addario, M., Greco, A., Miglioretti, M., Sarini, M., Scrignaro, M., . . . Crocetti, E. (2015). Social support and adherence to treatment in hypertensive patients: A meta-analysis. *Annals of Behavioral Medicine*, *49*, 307–318.

Mahoney, M. J. (1977). Publication prejudices: An experimental study of confirmatory bias in the peer review system. *Cognitive Therapy & Research*, *1*, 161–175.

Mante, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, *22*, 719–748.

Moher, D., Liberati, A., Tetzlaff, J., & Altman, D. G., & The PRISMA Group. (2009). Preferred reporting items for systematic reviews and meta–analyses: The PRISMA statement. *PLoS Medicine, 6*, *e1000097*.

Negru, O. (2012). The time of your life: Emerging adulthood characteristics in a sample of Romanian high-school and university students. *Cognition, Brain, Behavior*, *16*, 357–367.

Northouse, L. L., Katapodi, M. C., Song, L., Zhang, L., & Mood, D. W. (2010). Interventions with family caregivers of cancer patients, meta-analysis of randomized trials. *CA Cancer Journal for Clinicians*, *60*, 317–339.

Petticrew, M., & Roberts, H. (2006). *Systematic reviews in the social sciences*. Malden, MA: Blackwell Publishing.

Reifman, A., Arnett, J. J., & Colwell, M. J. (2007). Emerging adulthood: Theory, assessment, and application. *Journal of Youth Development*, *2*, 1–12.

Riley, R. D., Lambert, P. C., & Abo-Zaid, G. (2010). Meta-analysis of individual participant data: Rationale, conduct, and reporting. *British Medical Journal (Online)*, *340*, 521–525.

Roberts, B. W., Walton, K. E., & Viechtbauer, W. (2006). Patterns of mean-level change in personality traits across the life course: A meta-analysis of longitudinal studies. *Psychological Bulletin*, *132*, 1–25.

Rosenthal, R. (1979). The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, *86*, 638–641.

Rothstein, H. R., & Hopewell, S. (2009). Grey literature. In H. Cooper, L. V. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 103–125). New York, NY: Russell Sage.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.). (2005). *Publication bias in meta-analysis*. Chichester, England: John Wiley.

Simmonds, M. C., Higgins, J. P. T., Stewart, L. A., Tierney, J. F., Clarke, M. J., & Thompson, S. G. (2005). Meta-analysis of individual patient data from randomized trials: A review of methods used in practice. *Clinical Trials*, *2*, 209–217.

Sterne, J. A., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis: Guidelines on choice of axis. *Journal of Clinical Epidemiology*, *54*, 1046–1055.

Sterne, J. A., Becker, B. J., & Egger, M. (2005). The funnel plot. In H. R. Rothstein, A. J. Sutton, & M. Borenstein (Eds.), *Publication bias in meta-analysis* (pp. 75–98). Chichester, England: John Wiley.

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? Advantages and disadvantages of systematic reviews using individual patient data. *Evaluation in the Health Professions*, *25*, 76–97.

Stewart, L. A., Tierney, J. F., & Clarke, M. (2011). Reviews of individual patient data. In J. P. T. Higgins & S. Green (Eds.), *Cochrane handbook for systematic reviews of interventions*, Version 5.1.0 [updated March 2011]. The Cochrane Collaboration. Retrieved from www.cochrane-handbook.org

Stroup, D. F., Berlin, J. A., Morton, S. C., Olkin, I., Williamson, G. D., Rennie, D., . . . Thacker, S. B. (2000). Meta-analysis of observational studies in epidemiology: A proposal for reporting. *Journal of the American Medical Association*, *283*, 2008–2012.

Sutton, A. J., & Abrams, K. R. (2001). Bayesian methods in meta-analysis and evidence synthesis. *Statistical Methods in Medical Research*, *10*, 277–303.

Turner, R. M., Jackson, D., Wei, Y., Thompson, S. G., & Higgins, J. P. T. (2015). Predictive distributions for between-study heterogeneity and simple methods for their application in Bayesian meta-analysis. *Statistics in Medicine*, *34*, 984–998.

Valentine, J. C. (2009). Judging the quality of primary research. In H. Cooper, L. V. Hedges, & J. Valentine (Eds.), *The handbook of research synthesis* (2nd ed., pp. 129–146). New York, NY: Russell Sage.

Viechtbauer, W. (2007). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Journal of Psychology*, *215*, 104–121.

Author Biography

Elisabetta Crocetti, PhD, is a researcher at the Utrecht University, the Netherlands. Her major research interests include identity formation in adolescence and emerging adulthood. She is also strongly interested in methodological and statistical issues related to social research, such as cross-cultural validation of measurement instruments, longitudinal analyses, and systematic reviews and meta-analysis.