Accepted for publication (June 12, 2016): Zeitschrift für Psychology, Issue 1: Hotspots in Psychology 2017: Guest Editors: Edgar Erdfelder and Michael Bošnjak (University of Mannheim, Germany)

erdfelder(at)psychologie.uni-mannheim.de

bosnjak(at)uni-mannheim.de

Avoiding Methodological Biases in Meta-Analysis: Use of Online Versus Offline Individual Participant Data (IPD) in Educational Psychology

Esther Kaufmann

University of Zurich, Zurich, Switzerland

Ulf-Dietrich Reips

University of Konstanz, Konstanz, Germany

Katharina Maag Merki

University of Zurich, Zurich, Switzerland

Author Note

Esther Kaufmann, Institute of Education, University of Zurich, Zurich, Switzerland; Ulf-Dietrich Reips, Department of Psychology, University of Konstanz, Konstanz, Germany; Katharina Maag Merki, Institute of Education, University of Zurich, Zurich, Switzerland.

Correspondence concerning this article should be addressed to Esther Kaufmann, Institute of Education, University of Zurich, Freiestrasse 36, 8032 Zurich, Switzerland. Tel.: +41 (0)44 634 27 72. E-mail: esther.kaufmann@ife.uzh.ch

Abstract

Individual participant data (IPD) meta-analysis is the gold standard of meta-analyses. This paper points out several advantages of IPD meta-analysis over classical meta-analysis, such as avoiding aggregation bias (e.g., ecological fallacy or Simpson's paradox) and also shows how its two main disadvantages (time and cost) can be overcome through Internet-based research. Ideally, we recommend carrying out IPD meta-analyses that consider online vs. offline data gathering processes and examine data quality. Through a comprehensive literature search, we investigated whether IPD meta-analyses published in the field of educational psychology already follow these recommendations; this was not the case. For this reason, the paper demonstrates characteristics of ideal meta-analysis on teachers' judgment accuracy and links it to recent meta-analyses on that topic. The recommendations are important for meta-analysis researchers and for readers and reviewers of meta-analyses. Our paper is also relevant to current discussions within the psychological community on study replication.

Keywords: IPD meta-analysis, ecological fallacy, online versus offline, Simpson's paradox, study replication

Avoiding Methodological Biases in Meta-Analysis:

Use of Online Versus Offline Individual Participant Data (IPD) in Educational Psychology

Avoiding Methodological Biases in Meta-Analysis:

Use of Online Versus Offline Individual Participant Data (IPD) in Educational Psychology

Classical meta-analysis has been used to investigate questions, such as whether fat intake causes breast cancer (Caroll, 1975) or how accurately teachers judge students (Hoge & Coladarci, 1989; Kaufmann, in prep.; Südkamp, Kaiser, & Möller, 2012) by aggregated person data taken from single studies. Hence, classical meta-analysis (so-called APD, aggregated person data meta-analysis) *can be seen as an evaluation of multiple replication studies for a given topic*. Considering that the field of psychology currently faces criticism with respect to replication of scientific studies (see Open Science Collaboration, 2015), having a replication check like a classical meta-analysis approach is becoming increasingly important. However, even though classical meta-analyses are often used, they are also criticized for introducing methodological bias (e.g., aggregation bias as ecological fallacy or Simpson's paradox). For this reason, methodological experts promote individual participant data (IPD) meta-analysis (also known as mega-analysis or integrative data analysis). Unlike classical meta-analysis, IPD meta-analysis is based on a direct analysis of all of the raw, unit-level data generated from multiple studies, as opposed to analysis of the aggregated summary data. 'Units' are generally human participants, but can also refer to other types of primary research units, such as schools or hospitals (see Stewart et al., 2015, p. 1657).

Although IPD meta-analysis prevents aggregation bias, the data collection is time-consuming and costly. In this paper, we argue that the use of Internet-based research as part of the data collection phase of an IPD meta-analysis is an effective means to save time and money. Moreover, such research provides an additional replication check by analyzing the data-gathering process (online vs. offline) in detail. We have focused on studies within educational psychology, however, our study aim is also applicable to other research fields. The overall aim of this paper is to verify, by means of a review, the application of online vs. offline data-gathering processes of IPD meta-analysis within educational psychology.

In our paper, we introduce the value of IPD meta-analysis research, highlight its drawbacks in terms of time and cost, and assess the use of the IPD approach within educational psychology. We then introduce Internet-based research and demonstrate how it can overcome the drawbacks of IPD meta-analysis. We further integrate this solution with a literature review to verify whether published IPD meta-analyses already consider online and offline data-gathering approaches. Finally, we take teachers' judgment achievement as an example of an ideal study. To introduce readers unfamiliar with meta-analysis in the field of education with an actual meta-analysis on the topic, we also consider the judgment accuracy of teachers (see Südkamp et al., 2012). Hence, teachers' judgment achievement is represented in the following by the correlation index between teachers' judgments on students' abilities, for example on a mathematical test as an evaluated criterion.

Classical Meta-Analysis within Educational Psychology

When meta-analysis was first introduced to the educational field, study-level data was viewed as the unit of analysis to reach more power and to reduce uncertainty (see Glass, 1976; 2016). Since then, the success of meta-analysis on study-level units has been demonstrated in other fields like medicine (Ioannidis, 2010; Rosenthal & DiMatteo, 2001; Shadish, 2015). As one positive outcome, this fruitful expansion of meta-analysis has resulted in the new method of *mega meta-analysis*, in which meta-analysis is the aggregation unit (see Hattie, 2009; Lipsey & Wilson, 1993). On the negative side, an evaluation of the various reviews of meta-analysis within educational psychology has shown that comprehensive reporting (e.g., literature search, synthesis techniques) is often missing (see Polanin, Maynard, & Dell, 2016).

Today, meta-analysis, especially in education, has a practical impact, as it provides politicians with a decision-making instrument. For example, recently the Swiss Council for Educational Research initiated a systematic review of the impact of learning multiple languages at school (see Dyssegaard et al., 2015). Seeing that meta-analysis is used as a decision-making tool for politicians worldwide, we highlight that it is also criticized for introducing methodological bias, which could have a dramatic practical impact.

Methodological Bias

According to Bernard (2014, p. 3), methodological "bias is systematic inaccuracy in data due to characteristics of the processes employed in its collection, manipulation, analysis and/or presentation." (For an overview of possible biases in meta-analysis, see Tierney et al., 2015.) To prevent any methodological bias, special guidelines are used during the journal peer-review process. For example, the leading journal for meta-analysis research within psychology, *Psychological Bulletin*, uses the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) checklist and the Meta-Analysis Reporting Standards (MARS, see Albarracín, 2015).

For meta-analyses to be published in journals, submitted papers have to report using these guidelines to check for any possible bias. However, meta-analysis guidelines do not consider possible aggregation bias resulting from the data-aggregation strategy used in classical meta-analysis (for details, see below).

Due to the absence of a quality check for aggregation bias during the publication process, critical discussion on aggregation bias may also be missing from publications. From our standpoint, this critical discussion is greatly needed, not only for (mega) meta-analysts, but even more so for readers of meta-analyses (e.g., researchers, students, politicians). For this reason, we focus on methodological bias resulting from data aggregation, or ‘aggregation bias’ in this study.

Study aggregation and possible aggregation bias. Possible aggregation bias is very important in the context of classical meta-analysis approaches, due to the fact that classical meta-analyses are based on analysis of aggregated summary data extracted from studies (e.g., the overall correlation between two variables in different studies). For example, in the meta-analysis by Südkamp et al. (2012), each study included in the meta-analysis is represented by one value, namely, the achievement of the aggregated teacher judgments. Each of these study values is comprised of a different number of teacher judgments due to the differing number of teachers in each study, which varies from 16 to 9,650. To exclude any sampling bias, each study value was weighted by the number of teachers involved and aggregated to reach the teacher judgment achievement value across studies, or, in other words, to obtain the meta-analyzed value of teacher judgment achievements. This aggregation strategy at the study level is a key feature of classical meta-analysis and may lead to misleading aggregated results, as individual teacher’s data are not properly taken into account (see below: ecological fallacy and Simpson paradox). We emphasize that the inappropriate aggregation level within classical meta-analysis approaches to draw conclusions about relationships between constructs at the individual level may result in misleading conclusions within educational psychology (see Hanushek, Rivkin, & Taylor, 1996; Moffitt, 1996; Sirin, 2005).

The Individual Participant Data (IPD) Approach

The IPD meta-analysis approach is considered to be the ‘gold standard’ of meta-analysis (Chalmers, 1993) because of its advantages over the classical meta-analytical approach. Unlike classical meta-analysis, IPD meta-analysis is based on a direct analysis of all of the raw, unit-level data generated from multiple studies. Taking teacher judgment achievement as an example, researchers need a value for each single teacher judgment achievement in the studies or need to ask the study authors for the raw data of the studies in order to perform an IPD meta-analysis. Practically, this means that 20 judgment achievement values are needed if 20 teachers are considered in one study. In Südkamp et al. (2012), researchers would have needed to compile the judgment accuracy values of 38,973 teachers, as compared to aggregating the data of 75 aggregated values as part of their classical meta-analysis. Within the IPD approach, there are two different ways of aggregating individual data (see Debray et al., 2015; Simmonds et al., 2005): (1) the so-called *one-stage approach*, in which all IPD data are analyzed simultaneously (see also multilevel model, e.g., Televantou et al., 2015); and (2) the so-called *two-stage approach*, in which data are analyzed non-simultaneously (i.e., meta-analysis). Within the one-stage approach, IPD data are analyzed as if they belong to one single study, ignoring important differences between studies, while within the two-stage approach, approaches considering sampling biases as the only reason for study differences. We highlight the psychometric Hunter and Schmidt approach (2014). The Hunter and Schmidt meta-analysis approach is the only one that includes a palette of corrections of artifacts, such as sampling and measurement error, and dichotomization of continuous variables for study differences. Hence, it is the only approach that focuses, in detail, on differences between studies. Ignoring differences between studies in detail may lead to incorrect interpretation of overall heterogeneity (see Schmidt & Hunter, 2014).

Advantages of IPD Meta-Analysis

There are several advantages of using IPD meta-analysis over classical meta-analysis; the main ones are illustrated below, starting with the prevention of two aggregation biases.

The prevention of ecological fallacy. Ecological fallacy may arise because associations between two variables at the group (or ecological) level may differ from associations between analogous variables measured at the individual level (Robinson, 1950; see Alker, 1969, for an overview and typology of ecological fallacies). Due to the fact that IPD meta-analysis relies on individual-level data as opposed to aggregated data, the IPD approach avoids potential ecological fallacy. For example, Robinson (1950) used aggregated data from the United States to show that the average correlation between the proportion of foreign-born residents and the literacy rate at state level was r = - .53, suggesting that foreign-born immigrants were less literate than their native-born peers. However, the average correlation between foreign birth and literacy at the individual level was, in fact, positive and much lower in magnitude (r = .12), suggesting that foreign-born immigrants were, on average, more literate than native citizens. The negative correlation at the state level arose because immigrants tended to settle in states where the native population was more literate (Robinson, 1950). Although Robinson first published the paper in the 1950s, the majority of meta-analytical approaches since then have neglected the potential for ecological fallacy (see also Cooper & Patall, 2009; Stewart & Parmar, 1993; Viechtbauer, 2007). A rare example considering the ecological fallacy is the meta-analysis by Berlin et al. (2002), which revealed an ecological fallacy leading to a small renal transplant patient group being overlooked for a therapy that could have had a beneficial effect. Taking our example within educational psychology of teachers' judgment achievement to show the relevance of ecological fallacy, it could be that teachers' judgment achievement accuracy at the study level contradicts teachers' judgment achievement accuracy at the individual level. These contradictory results may be due to underlying factors that influence teachers' judgment achievement accuracy. Like the classical example introduced by Robinson (1950), which differentiates in its analysis between states and across states for revealing an ecological fallacy, we recommend the same data-analyzing procedure. Particularly in Switzerland, it is necessary to analyze whether there are any differences in teachers' judgment achievement accuracy between the cantons (states) because the first nine years of education are organized differently by each canton.

The prevention of Simpson's paradox. The IPD approach avoids a second type of ecological fallacy aggregation bias, namely, Simpson's paradox (Simpson, 1951), which occurs when the heterogeneity in the population is underestimated. A well-known example of Simpson's paradox is in Bickel, Hammel and O'Connell's (1975) analysis of graduate admission data from the University of California, Berkeley. Whereas aggregated data from multiple academic departments seemed to indicate that there was a higher admission rate for male than for female applicants (suggesting a gender bias in favor of men), Bickel et al.'s (1975) closer look revealed that women tended to apply to competitive departments that had higher rejection rates. Hence, if anything, there was a gender bias in favor of women.

Heterogeneity check. In addition to the aggregated values mentioned above, variances (heterogeneity of data) are also important for interpreting the results of meta-analyses. As the impact of heterogeneity is seldom discussed within medical IPD meta-analysis, we see the need to focus on it, to prevent the same mistake within educational psychology (see Simmonds et al., 2005).

Within classical meta-analysis approaches, it is often neglected that heterogeneity originates not only from differences between studies, as mentioned above, but also from differences within studies, such as differences between students' gender and age. The identification of student characteristics that are associated with heterogeneity helps identify student groups that are not accurately judged. These checks are needed to reveal any violations resulting from teachers' judgments on student equality. To identify distinctions in heterogeneity in classical meta-analysis research, meta-regression is often applied to reveal any moderator variables. Moderator variables influence the relationship between two variables. For example, teacher's judgment achievement could be influenced by students' age; younger students are possibly judged less accurately than older students. In general, there are difficulties in using summary data to represent individual participants (Lau, Ioannidis, & Schmid, 1997; Schmid, Stark, Berlin, Landais, & Lau, 2004; Schmidt & Hunter, 2014, pp. 384), IPD meta-analysis is a fruitful approach to overcome this drawbacks of classical meta-analysis.

There are more advantages to using IPD meta-analysis over classical meta-analysis, but apart from the ones we presented above, those are not vital to our argumentation (for more advantages, see Lyman & Kuderer, 2005; Tierney et al., 2015).

Disadvantages of IPD Meta-Analysis

Relative to other fields such as medical science, the IPD approach is seldom applied within the social sciences (Pigott, Williams, & Polanin, 2012). The following two disadvantages give evidence as to why IPD meta-analysis is seldom applied (see Cooper & Patall, 2009).

Time and cost. First of all, the data-gathering procedure in IPD meta-analysis is very time-consuming, as researchers require raw data for the meta-analysis. In particular, problems arise if raw data from studies published many years ago are needed and individual data are not published. It is then necessary to contact the authors of the paper, which may become quite difficult as time passes and contact information changes or is no longer valid. Researching contact information can be cumbersome and costly.

The second disadvantage of IPD meta-analysis is the increasing cost factor as a result of the work required to secure source data. Medicine is the leading field where IPD meta-analysis is employed. Medical IPD meta-analyses are often international collaborative projects organized by different teams. There are groups of researchers conducting primary research and researchers managing the IPD meta-analysis project. The management group is tasked with organizing data by asking researchers for additional data. This management group is also accompanied by a small advisory group of special knowledge experts (e.g., specialized in statistical methodology). Considering that IPD collaborative groups can be as large as 100 people (see Stewart & Tierney, 2002, p. 93; Tierney et al., 2015), factors such as time and cost vary widely. To exemplify the differences in cost and time for classical meta-analysis vs. IPD meta-analysis, a typical classic meta-analysis may cost $10,000 and last four months, while a comparable IPD meta-analysis may cost up to $200,000 and last for at least 3.5 years, as the project may still be ongoing at the time of publication (Ioannidis, Rosenberg, Goedert, & O'Brien, 2002).

Consequences. Knowing that disadvantages exist, we argue that a comparison of the results of an IPD meta-analysis with a classical meta-analysis is the optimal means of carrying out a meta-analysis. Combining both approaches overcomes any drawbacks that occur as a result of using only one approach (Debray et al., 2015). Moreover, this combined approach leads to a result validation (see also Riley, Simmonds, & Look, 2007; Simmonds et al., 2005).

Therefore, we recommend performing an IPD-only meta-analysis and comparing this with a sensitivity analysis in which the IPD meta-analysis is supplemented by a classical meta-analysis. For the classical meta-analysis, we recommend that the Hunter and Schmidt approach (2014) is the best analysis for this purpose, as it considers study heterogeneity by multiple corrections.

When conducting an IPD meta-analysis, specific individual data are needed, such as with studies on teachers' judgment achievement, or are poorly reported or even missing. Therefore, an additional solution is needed to overcome the above-mentioned disadvantages of IPD meta-analysis. As part of this paper, we recommend using Internet-based research as a solution, but we first present the state of the art of IPD meta-analysis and a general overview of its use within educational psychology. Does enough IPD meta-analyses exist for a direct comparison with classical meta-analyses?

IPD Meta-Analysis in Educational Psychology

There is currently no published review to describe the state-of-the-art in IPD meta-analysis for educational psychology. Such a review is available within the medical field (see Simmonds, Stewart, & Stewart, 2015), which also describes IPD meta-analysis characteristics (e.g., random vs. fixed models). The review by Simmonds et al. (2015) provides an ideal guideline to apply these characteristics to psychological educational IPD meta-analysis. We highlight that studies in medicine differ from studies in education, as dichotomous outcomes (healthy or not) resulting from experimental trial studies are common in medicine, while field studies and continuous outcomes of teacher responses are the norm within educational psychology. Moreover, educational psychology is often based on hierarchical data collected from different schools, staff members, teachers and students, in comparison with medical studies. These disparities warrant the need for a review in educational psychology to verify the differences between the two fields.

As a next step, we introduce suggestions for overcoming the shortcomings of IPD meta-analysis, which will lead us to more precisely determine our research questions.

Internet-Based Research as a Means to Overcome the Challenges of Conducting IPD Meta-Analysis

Comparison of online vs. offline data gathering. In recent decades Internet-based research quickly spread and was met with growing interest, not only in the educational sciences (see Batinic, Reips, & Bosnjak, 2002; Reips, 2002; Reips & Bosnjak, 2001). In the following, the online data-gathering approach is defined to studies in which participants respond via the Internet.

Since the beginning of online research, many strategies have been developed and implemented to improve data quality, such as the automatic online function that alerts participants if they skip a question (for further strategies, see Reips, 2002, 2006, 2008). Today, the advantages of online functions lead many to assume that the quality of data gathered online is better than data gathered offline.

Different environments may also introduce different contexts (e.g., level of anonymity), which may lead to differences within data quality. For example, Kaufmann and Reips' (2008) online experimental study found that social desirability responding is age dependent, with younger and older people having a higher tendency to show it. Kaufmann and Reips (2008) confirmed a previous offline study by Stöber (1999), but also revealed that across and within age groups, the tendency to show social desirability responding is lower with online data-gathering approaches than with offline data-gathering approaches. On the other hand, a meta-analysis of online vs. offline data gathering (Dodou & de Winter, 2014) led to the conclusion that there are no differences in data quality introduced by the data-gathering process. However, as the participants' age was not considered in that study, we believe that the question remains open as to whether online and offline data-gathering processes lead to the same data quality in all domains.

The current state of research on the effects of online vs. offline data gathering processes on data quality is ambiguous, and we find it premature to conclude that the two approaches lead to the same data quality. Therefore, we recommend using both data-gathering approaches. Moreover, we argue that the integration of different data-gathering approaches (online vs. offline) improves the quality and validity of the database (Campbell & Fiske, 1959). Importantly, sensitivity analysis should also be used to investigate whether (and how) online and offline samples differ.

Time and cost. In addition to conducting a comprehensive data quality check with online vs. offline data-gathering processes, the two drawbacks of IPD meta-analysis, time and cost, can be overcome through Internet-based research. As mentioned by Cooper and Patall (2009), IPD meta-analysis requires many more staff for data collection, entry, and cleaning than classical meta-analysis does. We agree and would like to stress the usefulness of Internet-based research in overcoming this drawback. We also agree with Curran and Hussong (2009, p. 81), who emphasize that online data collection can overcome some of the challenges associated with collecting individual participant data, because individual-level data are entered into a data set automatically and directly via computer (Reips, 2008). The costs of online research methods are lower, as there is no need for laboratory space, personnel hours, equipment and administration. Moreover, the use of online data collection procedures can be especially advantageous if potential participants are difficult to recruit and data are hard to obtain. Especially in educational science, teachers are busy and rarely motivated to answer research questions that interrupt their daily activities. In this sense, schools likely welcome online surveys that can be filled out at any time and place as a supplementary research approach. Online research not only reduces study organization time for researchers, but more participants take part within a shorter time period, because online surveys are more accessible than traditional surveys.

Additionally, when participants are few in number, cases are rare, the database is simply too small, or information is missing in the database (as in our example of an ideal study outlined below), an Internet-based data-gathering process may be instrumental. A good example of the value of Internet-based research for a special sample is a web survey involving people suffering from sexsomnia (a rare disorder), which quickly increased the pool of data collected over 20 years through offline research by 90% (Mangan & Reips, 2007). Likewise, we argue that online data collection can quickly increase the volume of individual-level data that can be used for IPD meta-analysis.

Research Questions

We seek to answer the following research questions by conducting a review of the literature:

- How many meta-analyses within educational psychology considered the type of data-gathering approach (online vs. offline)?
- How many of these meta-analyses followed an IPD meta-analysis approach?
- What are the characteristics (e.g., random vs. fixed-effect model, see Simmonds et al., 2015) of these IPD meta-analyses within educational psychology?

Method

Literature Review

Like the study by Simmonds et al. (2015), we use the same time frame in which studies have been published. Both studies consider articles published from 2005 to 2015. We used the following databases: ERIC (Education Resources Information Center), Google Scholar, PsycInfo, Scopus and Web of Science. Our list of keywords took into account different spellings of the terms being searched (e.g., "meta-analysis" vs. meta analysis"). Complete information about the search process is available from the authors by request.

Our comprehensive literature search revealed that no IPD meta-analysis has been conducted to date that considers online and offline data-gathering approaches in educational psychology. This reinforces the need and value of this study to close this apparent gap in the literature.

Since no study is available that meets our inclusion criteria, we present here an ideal study. Our example should inspire researchers and also explain where it makes sense to launch an IPD meta-analysis.

Study Description with the Help of an Ideal Example of an Online vs. Offline IPD Meta-Analysis in Educational Psychology

When introducing our ideal study, we present the current state of research of the chosen study topic, teachers' judgment achievement accuracy, to show the need for Internet-based data collection. Next, we present our ideal study example and an ideal data comparison of different types of meta-analysis.

Teachers' judgment achievement accuracy: The current state of meta-analysis. The importance of teachers' judgment achievement is reflected by the fact that there are three reviews of it. The first review (Hoge & Coladarci, 1989) was published in the early days of meta-analysis. Based on a descriptive review of 55 different judgment tasks from 16 studies, it found a medium correlation (r = .66) between teachers' judgments of student abilities and students' scores on achievement tests. Südkamp et al. (2012) used a quantitative meta-analytical approach to review the results of 75 studies on teachers' judgment achievement accuracy published after 1989 (i.e., excluding the studies that were part of Hoge and Coladarci's review). Compared to Hoge and Coladarci (1989), Südkamp et al. (2012) found a lower estimate of teachers' judgment achievement accuracy (r = .53). These differing results indicate that there is some ambiguity with regard to how accurately teachers judge students. Südkamp et al. (2012) suggested that the difference in results between the two reviews may be due to the different meta-analytical approaches (descriptive vs. quantitative) used in the two studies. In contrast to these two reviews, our reviews (Kaufmann & Athanasou, 2009; Kaufmann, Sjödahl, & Mutz, 2007; Kaufmann et al., 2013) focused on social judgment theory

(SJT) (Hammond & Stewart, 2001; Karelaia & Hogarth, 2008; Kaufmann et al., 2013) studies. Only within SJT do we receive additional information about teacher's judgment achievement inaccuracy, revealing whether its source is due, for example, to the teacher, the task, or both. Hence, only studies within the SJT framework show exactly where to launch interventions to improve teachers' judgment achievement accuracy. We also conducted a comparison of an IPD approach with a classical meta-analysis in the framework of SJT (Kaufmann, Sjödahl, & Mutz, 2007, vs. Kaufmann & Athanasou, 2009) and a comparison of our results of a classical meta-analysis approach with a psychometric meta-analysis approach (Kaufmann et al., 2013).

Teacher's judgment achievement accuracy: Shortcomings. In the reviews by Hoge and Coladarci (1989) and Kaufmann et al. (2013), only a few studies reported data at the individual level (two by Hoge & Coladarci and three by Kaufmann et al.). An additional verification to see whether sample characteristics were reported, such as gender, was also futile (see also Südkamp et al., 2012). We conclude that although teachers' judgment achievement accuracy is a hotspot in educational psychology, with several meta-analyses conducted over the last 25 years, there is a lack of studies reporting individual teacher data in combination with student characteristics. Hence, collecting additional data via an online gathering process can overcome this lack of information in a quick and economical way.

Another shortcoming of the studies included in the meta-analysis on teachers' judgment achievement within SJT is that none of them used an online data-gathering process. To date, there is no data-gathering (online vs. offline) check done within these types of studies, although different environments introduced different circumstances, as already mentioned. For this reason, such a data check is performed before data collected from various data-gathering processes are included in an IPD meta-analysis. We maintain that in our online study, the gathering of teachers' judgment achievement is possibly more accurate because the potential for social desirability responding is reduced (see Kaufmann & Reips, 2008).

Due to missing data and our argumentation outlined above, we see Internet-based research as a fruitful tool to easily close this research gap. Our research design is outlined in detail below.

The Online Study Design

Data collection. As mentioned above, only within the framework of SJT is additional information about teachers' judgment achievement inaccuracy available, as compared to studies outside the SJT approach. A suitable study for replication was checked via the database of Kaufmann et al. (2013). The latest study that includes individual data is that of Athanasou and Cooksey (2001), even though it was published 15 years ago.

Athanasou and Cooksey (2001) constructed 120 student profiles (i.e., vignettes), which can easily be used in an online survey. Unlike in a paper survey, an online survey enables a randomized call of each vignette, which leads us to argue that modern techniques in Internet-based research often automatically lead to improvements in research design and methodology. Practically speaking, teachers are invited to complete the survey and to judge each student's profile.

Analysis

After the online data-gathering process is finished, data quality must be verified using analysis on different levels, combine analysis on different levels and finally check them by sensitivity analysis as outlined in the following.

Data comparison at the individual level. The data collected as part of an online data-gathering process is first checked for quality by comparing the two data sets (online vs. offline). Are there any differences when plotting the data? Is the online process of gathering teachers' judgment achievement more or less accurate than the offline data-gathering process? As we discussed above (see comparison of online vs. offline data gathering), based on current research, it is unclear if online or offline data-gathering processes lead to better quality.

Moreover, as individual data are available, we can check for outlier data of persons. To check this is important, as the number of teachers' (participants) also influences the overall teachers' achievement value. Since we also gathered additional vital data on teachers and students, such as gender and experience, teachers' judgment achievement can also be checked at the individual level considering these possible moderator variables.

Data analysis at the task level. In addition to an outlier screening at the individual level, an outlier screening at the task (or study) level is the next step in the following (Viechtbauer & Chueng, 2010). These checks are dependent on the number of aggregated teachers' judgment achievement values across tasks – or of possible outliers at the individual teacher level. If no individual data is available for a comprehensive outlier check, then subsequent analysis at the task level could be misleading.

Aggregation check. Due to our data-gathering process, it is possible to compare data at the individual level with data at the task level. Are teachers' judgment achievement confirmed by both analysis levels and within subgroup analyses focusing, for example, only on experienced teachers? With this often neglected aggregation check, we prevent the premature conclusion that aggregated person-level data may introduce an aggregation bias (see ecological fallacy or Simpson's Paradox above).

After this check, we recommend supplementing the IPD meta-analysis with a sensitivity analysis using a classical meta-analysis. Of the various approaches for meta-analysis, we recommend the so-called two-stage approach, namely the Hunter and Schmidt approach, due to its uniqueness in having a rich artifact corrections palette (Schmidt & Hunter, 2014).

Discussion

In this paper, we argue that IPD meta-analysis can overcome several drawbacks of classical meta-analysis. For example, meta-analysis based on aggregated data can result in erroneous conclusions due to aggregation bias (e.g., ecological fallacy and Simpson's paradox). We also highlight that Internet-based research could successfully overcome the drawbacks of IPD meta-analysis. Despite their advantages, IPD meta-analyses have seldom been conducted in (educational) psychology. Reasons for the current lack of IPD meta-analyses may be due to the insufficient availability of individual-level data and/or because it can be very costly and time-consuming to gather individual participant data. Internet-based research could be used as a control tool for previous data-gathering approaches within a focused research topic. Although we greatly recommend this approach, our literature search revealed that such an approach has yet to be used within educational psychology. Our results are in line with the review by Pigott et al. (2012), as they revealed that whereas IPD meta-analyses are found in a wide range of studies in the field of medicine, there was only one correlational IPD study in the social sciences (see Goldstein et al., 2000). Therefore, we outlined an ideal study to show how such an IPD meta-analysis considering online and offline data-gathering processes could be conducted. Although we highlight the benefits of an IPD meta-analysis approach from a methodological viewpoint in our paper, we recommend initiating IPD meta-analysis only after carefully carrying out a data and financial resources check (for an overview of factors when an IPD meta-analysis might be worthwhile, see Stewart & Tierney, 2002). The future will likely give rise to technical improvements, which will greatly facilitate such a project. On the other hand, statistically more sophisticated analyses are also expected to be developed, thereby warranting the need for more statistical experts in the field. For example, there are many unanswered questions about the combination of IPD meta-analysis and classical meta-analysis (see Riley, Simmonds, & Look, 2007) and about the evolution of new developments and modern methods on combining them (see Sutton, Kendrick, & Coupland, 2008).

Aggregation Units

Our review is also in contrast to the current development of meta-analysis in educational research, which does not focus on aggregation units. Nowadays, in the field of education, mega meta-analyses are conducted, where the aggregation unit is a meta-analysis (see Polanin et al., 2016). Neglecting possible aggregation bias in the field of education is also represented by Bernard's study (2014). Bernard (2014) focused on bias in meta-analysis, but neglected to consider aggregation units or aggregation bias. In our review, we focus on different aggregation units, starting with the individual level as an aggregation unit, and recommend comparing it to other aggregation units to check for any possible aggregation bias. Having our work support current discussion shines the spotlight on possible aggregation bias, not only in classical meta-analysis but also in mega meta-analysis.

Data-Gathering Process

In our review, we also recommend supplementing the offline data-gathering process with an online one. Considering this recommendation, we have to keep in mind that, depending on the topic, there may be differences in responding behavior that have to be controlled for. For example, Claxton, DeLuca and van Dulmen (2015) found that the association between alcohol consumption and engaging in casual sexual relationships and experiences seems to be stronger with online assessments than with paper-and-pencil assessments. As this is a sensitive topic, social desirability possibly plays a role (see also Kaufmann & Reips, 2008). This leads us to conclude that meta-analysis should include a methodological verification that considers an online vs. offline data-gathering process and checks whether any methodological bias, such as responding behavior, is introduced in the case of sensitive topics.

We do not want to miss the opportunity to critically discuss the promotion of online studies in the field of education. In this paper, we introduced an ideal online study that supplements the previous offline data-gathering process in SJT studies, as individual-based data is missing. We emphasize that conducting online studies successfully relies on a number of standards, techniques and methods and needs to be carefully checked before the study is launched (e.g., Reips, 2002; Reips, Buchanan, Krantz, & McGraw, 2015). However, we argue that different data-gathering processes and their verification improves validity and, therefore, are urgently needed.

Practical Consequences

From a meta-analysis consumer perspective, we argue that our review is also needed because, in our opinion, aggregation bias is not covered by current guidelines (e.g., PRISMA, MARS) to be followed before meta-analyses are published. Our review may, therefore, also shed light on a possibly forgotten point in meta-analysis and mega meta-analysis. Checking for bias prior to publication will lead to an improved publication process, as well as inspire discussion on aggregation units of meta-analysis and any possible resulting aggregation bias. We hope that our review also illustrates that a comparison of different aggregation units is preferred, rather than focusing on only one single type of meta-analysis. We see concentrating on different aggregation units and critically discussing them within different types of meta-analysis as an improvement in meta-analysis research, which may lead to more critical reading and practical transfer of meta-analysis results by researchers, students and even politicians.

Outlook

Archives

Besides our recommendation to supplement offline data-gathering processes with online gathering processes, the latter is easier to archive, thereby promoting data archiving. Several reviews that highlight reporting by IPD meta-analysis also demonstrate the need for archiving data (see Simmonds et al., 2005). Such an approach also avoids potential publication bias, data accessibility bias or reviewer selection bias (see also Debray et al., 2015). This is an important point, as there has been controversy and critical discussion on the attempts and methods of estimating publication bias (see Rothstein, 2008).

The success of IPD meta-analysis approaches in medical science relies heavily on archives. This data-saving strategy also decreases the time lost as a result of data organization and management for IPD meta-analysis. In this regard, we argue that the actual claim for archives within psychology (for details, see Bruder, Göritz, Reips, & Gebhard, 2014) also promotes IPD meta-analysis, and vice versa.

To make full use of archived data, both raw data at the individual level and study quality data is needed. Reliability values, such as study quality values, are seldom reported within studies, but should be included in future psychometric meta-analyses. In our psychometric reanalysis of the Hoge and Coladarci (1989) review, we obtained very little study quality data directly from publications. In this sense, the archiving of data also improves the quality of subsequent reviews, as information is often missing (see Polanin et al., 2016). Finally, as there is still uncertainty in data quality based on different data-gathering processes (online- vs. offline), study-specific information also needs to be archived as study quality data. For additional information what type of information should to be stored in archives from a meta-analytical perspective, we see meta-analysis guidelines as a useful source to check.

Taken together, efforts to archive data would also be helpful in checking for any aggregation bias as it promotes IPD meta-analysis, and also in re-checking classical meta-analyses and applying more sophisticated and newly developed meta-analytical techniques (see van den Heuvel & Griffith, 2016).

With the growth of databases, including archived databases, additional subgroup analyses become possible. For example, knowledge about children's development is of utmost importance in educational psychology; an enlarged database could make age-controlled statistical analysis possible. A first look at teachers' judgment achievement accuracy data reveals that the use of teachers' and students' age as a control factor has been completely disregarded, even though analyses considering age are needed.

In educational research, longitudinal designs are often required. A disadvantage not yet mentioned is the cross-sectional design in classical meta-analysis. Hence, we argue that the resulting archiving process as a consequence of the easier data-gathering process online than offline may also promote more longitudinal data studies.

Replication Crisis

Our request for more sophisticated data collection, reporting and archiving aligns well with the replication crisis of psychological data (see Open Science Collaboration, 2015), as meta-analysis itself is a check of replicated studies. In our opinion, meta-analysis as a replication check is not part of the current discussion within the psychological community. We highlight that this paper focuses only on a retrospective check of IPD studies; however, in medicine, there are also prospective IPD meta-analysis projects underway. To our knowledge, prospective approaches have not yet been undertaken in the educational field. Similar to retrospective IPD study evaluation, prospective IPD meta-analysis involves multi-collaborative international research teams. Such an approach would promote the initiation of a balanced data-gathering process using online vs. offline approaches, and a verification of any differences. In contrast to the actual replication studies done in psychological science, such an approach is prospective, meaning that it leads to the exclusion of possible confounding variables, such as programming changes, data transferring and time taken to achieve a fairer and more accurate comparison. Therefore, we see not only the need for improvements in meta-analysis research as outlined in our paper, but also their wider application to improve the evaluation of psychological research, overall.

References

Albarracín, D. (2015). Editorial. *Psychological Bulletin, 141*(1), 1–5.

Alker, H. S. (1969). A typology of ecological fallacies. In M. Dogan & S. Rokan (Eds.), *Quantitative Ecological Analysis in the Social Sciences* (p. 69-86). Massachusetts: MIT Press.

Athanasou, J. A., & Cooksey, R. W. (2001). Judgment of factors influencing interest: An Australian study. *Journal of Vocational Education Research, 26*(1), 1–13.

Batinic, B., Reips, U.-D., & Bosnjak, M. (Eds.). (2002). *Online Social Sciences*. Seattle: Hogrefe & Huber.

Berlin, J. A., Santanna, J., Schmid, C. H., Szczech, L. A., & Feldman, H. I. (2002). Individual patient- versus group-level data meta-regressions for the investigation of treatment effect modifiers: Ecological bias rears its ugly head. *Statistics in Medicine, 31,* 371–387. doi: 10.1002/sim.1023

Bernard, R. M. (2014). Things I have learned about meta-analysis since 1990: Reducing bias in search of ''The Big Picture.'' *Canadian Journal of Learning and Instruction, 40*(3). http://www.cjlt.ca/index.php/cjlt/article/view/870

Bickel, P. J., Hammel, E. A., & O'Connell, J. W. (1975). Sex bias in graduate admission: Data from Berkeley. *Science, 187*(4175), 398–404.

Bruder, M., Göritz, A. S., Reips, U.-D., & Gebhard, R. K. (2014). Ein national gefördertes Onlinelabor als Infrastruktur für die psychologische Forschung. *Psychologische Rundschau, 65*(2), 75–85.

Campbell, D. T. & Fiske, D. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81–105.

Carroll, K. (1975). Experimental evidence of dietary factors and hormone-dependent cancers. *Cancer Research, 35*, 3374–83.

Chalmers, I. (1993). The Cochrane collaboration: Preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Annals of the New York Academy of Sciences, 703*, 156–163.

Claxton, S. E., DeLuca, H. K., & van Dulmen, M. H. (2015). The association between alcohol use and engagement in casual sexual relationships and experiences: A meta-analytical review of non-experimental studies. *Archives of Sexual Behavior, 44*(4), 837–56.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods, 14*(2), 165–176.

Curran, P. J., & Hussong, A. M. (2009). Integrative data analysis: The simultaneous analysis of multiple data sets. *Psychological Methods, 14*(2), 81–1000.

Debray, T. P. A., Moons, K. G. M., van Valkenhoef, G., Efthimiou, O., Hummel, N., Groenwold, R. H. H., Reitsma, J. B. & the getReal methods review group (2015). Get real in individual participant data (IPD) meta-analysis: A review of the methodology. *Research Synthesis Methods, 6*(4), 293–309. doi:10.1002/jrsm.1160

Dodou, D., & De Winter, J. C. F. (2014). Social desirability is the same in offline, online, and paper surveys: A meta-analysis. *Computers in Human Behavior, 36*, 487–495.

Dyssegaard, C. B., Egeberg, J. de H., Sommersel, H. B., Steenberg, K., & Vestergaard, S. (2015). *A systematic review of the impact of multiple language teaching, prior language experience and acquisition order on student's language proficiency in primary and secondary school.* Copenhagen, Denmark: Danish Clearinghouse for Educational Research, Department of Education, Aarhus University.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. *Educational Researcher, 5*(1), 3–8.

Glass, G. V. (2016). One hundred years of research: Prudent aspirations. *Educational Researcher, 45*(2), 69-72. doi: 10.3102/0013189X16639026

Goldstein, H., Yang, M., Omar, R., Turner, R., & Thompson, S. (2000). Meta-analysis using multilevel models with an application to the study of class size effects. *Applied Statistics, 49*, 399–412.

Hammond, K. R., & Stewart, T. R. (Eds.). (2001). *The essential Brunswik: Beginnings, explications, applications*. Oxford, UK: University Press.

Hanushek, E., Rivkin, S., & Taylor, L. (1996). Aggregation and the estimated effects of school resources. *Review of Economics and Statistics, 78*, 611–627.

Hattie, J. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, UK: Routledge.

Hoge, D. H., & Coladarci, T. (1989). Judgments of academic achievement: A review of literature. *Review of Educational Research, 59*(3), 297–313. doi:10.3102/00346543059003297

Ioannidis, J. P. A. (2010). Meta-research: The art of getting it wrong. *Research Synthesis Methods, 1*, 169–184. doi:10.1002/jrsm.19

Ioannidis, J. P. A., Rosenberg, P. S., Goedert, J., & O'Brien, T. R. (2002). Commentary: Meta-analysis of individual participants' data in genetic epidemiology. *American Journal of Epidemiology, 156*(3), 204–210.

Karelaia, N., & Hogarth, R. (2008). Determinants of linear judgment: A meta-analysis of lens studies. *Psychological Bulletin, 134*(3), 404–426. doi:10.1037/0033-2909.134.3.404

Kaufmann, E. (in prep.). *Teachers as judges: A psychometric (re)evaluation of teacher's judgment accuracy.* [Working paper]. Zurich, Switzerland, University of Zurich.

Kaufmann, E., & Athanasou, J. A. (2009). A meta-analysis of judgment achievement defined by the lens model equation. *Swiss Journal of Psychology, 68*(2), 99–112. doi:10.1024/1421-0185.68.2.99

Kaufmann, E., & Reips, U.-D. (2008). *Internet-basierte Messung Sozialer Erwünschtheit: Theoretische Grundlagen und Experimentelle Untersuchung [Internet-based measurement of social desirability].* Saarbrücken, Germany: VDM Verlag Dr. Müller.

Kaufmann, E., Reips, U.-D., & Wittmann, W. W. (2013). A critical meta-analysis of Lens Model Studies in human judgment and decision-making. *PLoS ONE, 8*(12): e83528. doi:10.1371/journal.pone.0083528

Kaufmann, E., Sjödahl, L., & Mutz, R. (2007). The idiographic approach in social judgment theory: A review of components of the lens model equation components. *International Journal of Idiographic Science, 2*.

Lau, J., Ioannidis, J. P. A., & Schmid, C. H. (1997). Quantitative synthesis in systematic reviews. *Annals of Internal Medicine, 127*, 820–826.

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American Psychologist, 48*, 1181–1209.

Lyman, G. H., & Kuderer, N. M. (2005). The strengths and limitations of meta-analyses based on aggregate data. *BMC Medical Research Methodology, 5,* 14. doi: 10.1186/1471-2288-5-14

Mangan, M., & Reips, U.-D. (2007). Sleep, sex, and the Web: Surveying the difficult-to-reach clinical population suffering from sexsomnia. *Behavior Research Methods, 39*, 233–236.

Moffitt, R. (1996). Symposium on school quality and educational outcomes. *Review of Economics and Statistics, 78*, 559–561.

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*. doi:10.1126/science.aac4716

Pigott, T., Williams, R., & Polanin, J. (2012). Combining individual participant and aggregated data in a meta-analysis with correlational studies. *Research Synthesis Methods, 3*, 257–268.

Polanin, J. R., Maynard, B. R., & Dell, N. A. (2016). Overviews in educational research: A systematic review and analysis. *Review of Educational Research*. First published on February 17, 2016 as doi: 10.3102/0034654316631117

Reips, U.-D. (2002). Standards for Internet-based experimenting. *Experimental Psychology, 49*, 243–256.

Reips, U.-D. (2006). Web-based methods. In M. Eid & E. Diener (Eds.), *Handbook of multimethod measurement in psychology* (pp. 73–85). Washington, DC: American Psychological Association. doi:10.1037/11383-006.

Reips, U.-D. (2008). How Internet-mediated research changes science. In A. Barak (Ed.), *Psychological aspects of cyberspace: Theory, research, applications* (pp. 268–294). Cambridge, UK: Cambridge University Press.

Reips, U.-D., & Bosnjak, M. (Eds.). (2001). *Dimensions of Internet Science.* Lengerich: Pabst.

Reips, U.-D., Buchanan, T., Krantz, J. H., & McGraw, K. (2015). Methodological challenges in the use of the Internet for scientific research: Ten solutions and recommendations. *Studia Psychologica.* Advance online publication.

Riley, R. D., Simmonds, M. C., & Look, M. P. (2007). Evidence synthesis combining individual patient data and aggregate data: a systematic review identified current practice and possible methods. *Journal of Clinical Epidemiology, 60*(5), 431–439. doi.org/S0895-4356(06)00403-3

Robinson, W. S. (1950). Ecological correlations and the behavior of individuals. *American Sociological Review, 15*(2), 351–357.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review of Psychology, 52*, 59–82. doi:10.1146/annurev.psych.52.1.59

Rothstein, H. R. (2008). Publication bias as a threat to the validity of meta-analytic results. *Journal of Experimental Criminology, 4,* 61–81.

Schmid, C. H., Stark, P. C., Berlin, J. A., Landais, P., & Lau, J. (2004). Meta-regression detected associations between heterogeneous treatment effects and study-level, but not patient-level, factors. *Journal of Clinical Epidemiology, 57,* 683–697.

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings*. Los Angeles, CA: Sage.

Shadish, W. R. (2015). Introduction to the special issue on the origins of modern meta-analysis. *Research Synthesis Methods, 6*(3), 219–220. doi:10.1002/jrsm.1148

Simmonds, M. C., Higgins, J. P. T., Stewart, L. S., 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*(3), 209–217. doi: 10.1191/1740774505cn087oa

Simmonds, M., Stewart, G., & Stewart, L. (2015). A decade of individual participant data meta-analyses: A review of current practice. *Contemporary Clinical Trials, 45,* 76–83. doi:10.1016/j.cct.2015.06.012

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society, Series B, 13,* 238–241.

Sirin, S. R. (2005). Socioeconomic status and academic achievement: A meta-analytic review of research. *Review of Education Research, 75*(3), 417–453.

Stewart, L. A., Clarke, M., Rovers, M., Riley, R. D., Simmonds, M., Stewart, G., & Tierney, J. F. (2015). Preferred reporting items for a systematic review and meta-analysis of individual participant data: The PRISMA-IPD statement. *Journal of the American Medical Association (JAMA), 313*(16), 1657–1665. doi:10.1001/jama.2015.3656

Stewart, L. A., & Parmar, M. K. B. (1993). Meta-analysis of the literature or of individual patient data: Is there a difference? *The Lancet, 341*(8842), 418–422.

Stewart, L. A., & Tierney, J. F. (2002). To IPD or not to IPD? *Evaluation & the health professions, 25*(1), 76–97.

Stöber, J. (1999). Die Soziale-Erwünschtheits-Skala-17 (SES-17): Entwicklung und erste Befunde zu Reliabilität und Validität [The Social Desirability Scale-17 (SDS-17): Development and first findings on reliability and validity]. *Diagnostica, 45*(4), 173–177. doi:10.1026//0012-1924.45.4.173

Südkamp, A., Kaiser, J., & Möller, J. (2012). Accuracy of teachers' judgments of students' academic achievement: A meta-analysis. *Journal of Educational Psychology, 104*(3), 743–763. doi:10.1037/a0027627

Sutton, A. J., Kendrick, D., & Coupland, C. A. C. (2008). Meta-analysis of individual- and aggregate-level data. *Statistics in Medicine, 27*, 651–669. doi: 10.1002/sim.2916

Televantou, I., Marsh, H. W., Kyriakidies, L., Nagengast, B., Fletcher, J., & Malmberg, L.-E. (2015). Phantom effects in school composition research: Consequences of failure to control biases due to measurement error in traditional multilevel models. *School Effectiveness and School Improvement, 26*(1), 75–101.

Tierney, J. F., Vale, C., Riley, R., Smith, C. T., Stewart, L., Clarke, M., et al. (2015). Individual Participant Data (IPD) meta-analysis of randomized controlled trials: Guidance on their use. *PLOS Medicine, 12*(7), e1001855. doi:10.1371/journal.pmed.1001855

Van den Heuvel, E. R., & Griffith, L. E. (2016). Statistical harmonization methods in Individual Participants Data meta-analysis are highly needed. *Biometrics & Biostatistics International Journal, 3*(3): 00064. doi: 10.15406/bbij.2016.03.00064

Viechtbauer, W. (2007). Random-effects models and moderator analyses in meta-analysis. *Zeitschrift für Psychologie / Journal of Psychology, 215*(2), 104–121.

Viechtbauer, W., & Chueng, M. W.- L. (2010). Outlier and influence diagnostics for meta-analysis. *Research Synthesis Methods, 1*(2), 112–125. doi: 10.1002/jrsm.11