Eindhoven University of Technology

MASTER

Development of an Automatic Tool to Detect the SD/SE Mix-Up Error in Meta-Analyses

Feng, Sikai

Award date:
2024

Link to publication

Disclaimer
This document contains a student thesis (bachelor's or master's), as authored by a student at Eindhoven University of Technology. Student theses are made available in the TU/e repository upon obtaining the required degree. The grade received is not published on the document as presented in the repository. The required complexity or quality of research of student theses may vary by program, and the required minimum study period may vary in duration.

General rights
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain

Take down policy
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 03. Nov. 2024

Master's Degree in Human-Techonolgy Interaction

2022-2024

Master's Thesis

"Development of an Automatic Tool to Detect the SD/SE Mix-Up Error in Meta-Analyses"

Sikai Feng

1st Tutor Daniël Lakens

2nd Tutor Cristian Mesquida

EINDHOVEN, JULY 2024

ABSTRACT

This study presents an automatic detection tool for the SD/SE error in meta-analyses. Previous studies showed the prevalence of SD/SE mix-up errors in meta-analyses, in which researchers misused standard error of mean instead of standard deviation to calculate the effect size for primary studies. Such error affects the calculation for pooled effective size and thus misleading conclusions. The automatic tool can trace and download the PDF files for primary studies based on the used data and meta-analysis articles in the form of PDF. It can automatically extract existing SD/SE values from the tables in those primary study PDF files. By matching the used SD values in meta-analyses with the extracted SD/SE values from PDF files, it can detect the potential SD/SE mix-up error. In this paper, the detailed development of the automatic tool is illustrated. The tool is also validated with two meta-analyses for the accuracy of each sub-function. Based on the validation, it is discussed the main challenges of error automatic detection are the technology failure and the inconsistency of academic reporting. In the end, the limitations and recommendations are illustrated for future development and better reporting standards.

Keywords: Meta-Analysis, Automatic Detection, Standard Deviation SD, Standard Error SE

The datasets and code generated and analysed during the study are available in the Github repository: *https://github.com/Uranusikai/SESD_Error_Meta*

DEDICATION

A lot of thanks to anyone giving support and help during the final stage of my Master's program of Human-Technology Interaction at Eindhoven University of Technology. I found it extremely challenging and satisfying during the whole process of this project. Limited by my coding skills, I go back and forth between the frustration of failure and the pleasure of success. But thanks to the help from many people, the study was successfully completed.

First, my heartfelt thanks to my tutors Daniël Lakens and Cristian Mesquida. Reremered I barely knew anything about meta-analysis before this project and neither did I have experience with automation tool development. Without their help, there is no chance I could complete this study to the current progress. They always welcomed me with inspiring ideas, helpful feedback, and most importantly, heartwarming support. Their knowledge and help are essential for this study and the thesis.

Next, I wanted to thank my friends and family. Thanks to them for putting up with me weekly yelling about "meta-analysis" at random times. They always comforted and encouraged me when I felt stuck and isolated. And they help me a lot with moving, and let me stay. I'm grateful for their understanding, and especially, for all the good cooking.

I am grateful to have you during my final project, and I sincerely wish you all the best.

CONTENTS

1. INTRODUCTION

Meta-analysis is a powerful methodology in research, enabling the synthesis of findings from multiple studies to derive more robust conclusions. It enhances statistical power and offers a more comprehensive understanding of research questions by integrating data from multiple primary studies. However, the accuracy and reliability of meta-analyses are heavily dependent on the quality of data reported in primary studies and the researcher's analytical decisions. One of the most significant challenges in conducting meta-analyses is the presence of the SD/SE mix-up error. Such errors can lead to incorrect calculations and misleading conclusions, undermining the validity of the research findings. There is evidence that SD/SE mix-up is common in certain research fields such as sports and exercise/medicine research (Sandercock, 2024). However, due to issues such as the transparency and extensive volume of data inherent to meta-analysis, these common errors are often difficult to detect. Considering the accessibility of new technology and the massive amount of data involved in meta-analysis, automated tools can make it possible for these errors to be detected and prevented. It can not only be used to evaluate the published meta-analysis but also allow researchers to screen their work before submission, enhancing the overall quality of meta-analytic research.

1.1. Development and Application of Meta-Analysis

Meta-analyses is a statistical method that integrates information from multiple studies by computing an effect size to generate an overall conclusion. It has gained significant recognition in many research fields (Borenstein et al., 2021). It requires the researchers to select relevant primary papers, extract the required data, and analyze and synthesize the results. In a meta-analysis, the included study is the basic unit of analysis. Researchers calculate an effect size for each primary study using statistics such as means, SD, sample sizes and test statistics and then calculate a pooled effect size to summarize across all studies. Compared to subjective judgement, meta-analysis is considered a more effective and efficient approach to summarizing the study results, leading to its continuous development (Lee, 2018). The use of meta-analysis has continued to increase since its introduction, and large-scale meta-analyses have often been valued and cited heavily because they play a crucial role in guiding future research and new policies.

In recent years, the importance of meta-analysis has been increasingly recognized in various scientific disciplines. In evidence-based medicine, meta-analysis is becoming popular for resolving discrepancies in clinical research and in health sciences (Lee, 2018; Mak et al., 2010). In psychology, meta-analysis has also been studied and widely used (Cooper et al., 2019). Not only in medical and social sciences, meta-analysis also affected the views of the literature for researchers in biological sciences. And

biological meta-analysis holds different methodological considerations compared to medical and social sciences (Nakagawa and Santos, 2012). In the last two decades, the number of published meta-analyses of plant ecology has increased dramatically (Koricheva and Gurevitch, 2014). Additionally, meta-analysis has also been discussed in conservation science that it holds a different epistemic role in conservation science than in biomedical sciences (Kovaka, 2022). It allows plant ecologists to make sense of and generalize from large amounts of data to draw useful conclusions for theoretical advancement. By analyzing the number of meta-analysis papers published in PubMed between 1980-2014, Fuhr and Hellmich (2015) concluded the annual exponential growth in published meta-analyses was fairly steady. And it reached a 1.17-fold increase, substantially exceeding the 1.04-fold increase seen for all publications. Over 312,000 publications on PubMed have utilized meta-analysis keywords as of now (National Center for Biotechnology Information, 2024).

The broad use and great influence of meta-analysis in the academic also raise higher quality demands. Different guidebooks and handbooks are published for different domains to introduce and develop meta-analysis methods and standards (Cooper et al., 2019; Koricheva et al., 2013; Borenstein et al., 2021). Moreover, to standardize reporting and enhance the transparency and applicability of meta-analyses, several reporting guidelines have been published. For example, the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) (Moher et al., 2010) and the MARS (Meta-analysis Reporting Standards) (Cooper, 2011) are two primary reporting guidelines. Those guidelines provide a checklist and flow diagram that researchers can check to ensure comprehensive reporting of meta-analyses. These resources aim to enhance the quality of meta-analytical articles by providing methodologies, reporting standards, and domain-specific guidance, ensuring the conclusions are reliable.

1.2. Common Statistical Errors in Meta-Analysis

However, given the broad application of meta-analysis and the development of methodology and published guidelines, the quality of published meta-analyses is not as expected based on some preliminary research. Ioannidis (2016) concluded that currently there were a large number of systematic reviews and meta-analyses that produced unnecessary, misleading and contradictory information. misleading conclusions can be generated by errors in meta-analysis that are far more common than one might think. For instance, Kadlec et al. (2023) found that errors in physical activity meta-analyses were so prevalent that he summarized and categorized these common errors. A review found the study containing 42 reviews from the Cochrane Cystic Fibrosis and Inherited Diseases Group found that nearly half of the SRs had at least one data processing or reporting error (Jones et al., 2005). These common errors have also led to the retraction or correction of some meta-analysis literature (Guo et al., 2019; Roba et al., 2019; Metaxa and Clarke, 2024).

Among these summarized common errors, one crucial error is caused by the mix-up of standard deviation (SD) and standard error (SE). The SD measures the variability within a sample. In contrast, the SE describe how close the sample mean is to the population mean. The confusion between SD and SE has been studied by many authors, with the main argument that describing sample variability in terms of SE rather than SD is mistaken (Nagele, 2003; Wullschleger et al., 2014;Ko et al., 2014). Andrade (2020) discussed that including SEM values can mislead readers into believing that the sample data is more accurate and suggested eliminating SEM reporting altogether. As for meta-analysis, the ES should be calculated from SD. The misuse of SE instead of SD is also a common error. A meta-analysis review by Sandercock (Sandercock, 2024) focusing on the common error of calculation of effect sizes based on standard error rather than standard deviation, found that this error was very common and affected the results of the analysis.

Apart from the prevalence of errors in published meta-analyses, many of them are often difficult to identify. For transparency, Aytug et al. (2012) examined 198 meta-analyses published between 1995 and 2008 in 11 journals, and concluded the meta-analyses provided only 52.8% on average information to reproduce the meta-analysis. The lack of transparency directly affects reproducibility and validity because it is difficult to assess whether these meta-analyses are correct or not. Similarly, another study tested the reproducibility of psychological meta-analyses by reproducing 20 published meta-analyses, and it was concluded that 96% of them did not follow the reporting guidelines and 25% of them were unable to reproduce (Lakens et al., 2017). Some reasons for it were summarized as lack of access to raw data, no detailed information on effect sizes extracted from each study, or lack of information on how effect sizes were extracted or calculated. Lakens et al. (2016) pointed out the existing problem of meta-analysis reproducibility and proposed six recommendations to improve it. It was mentioned that meta-analysis quality control relies on open meta-analytic data. And all the potentially undetected common errors affect the quality of the meta-analytic literature and may even lead to many misleading conclusions. Based on Aytug et al.'s study review (2012), it was found that transparency was not significantly related to number of citations for the reviewed 198 meta-analyses. Potential undetected errrors in these articles with high citations may cause more misinformation under the influence of the lack of transparency. However, the specific detection of these potential common errors requires further research.

Compared with other existing analyses, one of the most important features of meta-analysis is that it is based on a considerable amount of data and text derived from different cited studies. However, the report does not always provide enough transparency when interpreting meta-analyses, either about the data applied or the method. The lack of transparency makes it strenuous to identify possible existing errors, even though errors can be common. Common errors such as confounding between standard error and standard deviation or ignoring the within-study correlations could easily have been noticed, but the lack of transparency complicates their detection.

1.3. Error Detection and Automation

Some common errors in meta-analyses exhibit similarities and do not require a thorough understanding of the primary studies. Therefore, we believe that there is potential for the use of an automated detection tool.

In fact, today's widely used meta-analyses also rely on well-established and convenient analytical tools (Polanin et al., 2017; Suurmond et al., 2017; Schwarzer et al., 2007). However, these common errors are often the result of different human factors. Similarly, it is believed that automated technologies can help researchers check for possible errors and correct them in a timely and efficient manner. Especially considering that the meta-analysis usually involves a huge amount of data from different research articles, the manual checking process can be tedious and time-consuming.

For academic papers, there are already automated tools like "Statcheck" to identify and correct inconsistencies in statistical reporting (Nuijten and Polanin, 2020). "Statcheck" achieves this by automatically extracting NHST results from articles and recalculating p-values. Technological developments bring new possibilities for data and text processing. For instance, Cheng et al. (2021) explored the initial steps of automating a meta-analysis using a Natural Language Processing (NLP) system, particularly to detect potential bias in scientific publications. And, Mariscal-Harana et al. (2023) developed an artificial intelligence tool for automated analysis of clinical databases that includes error detection and correction steps to ensure data accuracy.

However, there are no automated error detection tools designed for meta-analyses yet. Focusing on the SD/SE mix-up error, this study aims to explore the possibility of automatic detection for meta-analysis. The final automation tool will be validated to confirm its effectiveness.

1.4. Reseasch Questions

This project focuses on the following research question:

Can the SD/SE mix-up error in meta-analysis be detected using the automated detection tool?

It can be divided into the following research sub-questions.

1. Can the SD/SE mix-up error be detected using an automated tool? If so, how can it be automatic detected?

2. To what extent can information and data be extracted from tables or plots in a meta-analysis, as well as from primary studies cited in meta-analyses?

3. How can these methods be implemented into a tool that can accurately detect the SD/SE mix-up error?

All sub-tasks also summarize the entire method process of this project as well as the goals. This study focuses on the SD/SE mix-up error in meta-analysis caused by human factors and attempts to summarize methods to detect the error. Then an automated tool is developed based on the findings of the exploration. The tool may systematically extract and analyze information and data from the primary studies included in the meta-analysis. And ultimately the tool can utilize this extracted data to detect potential SD/SE mix-up errors.

This tool aims to enhance the quality and efficiency of meta-analyses by automating the detection of SD/SE mix-up errors. It addresses the challenge of identifying errors in meta-analyses due to the lack of transparency, with a lower threshold of expertise and in less time. With technology assistance, researchers and readers can efficiently assess the quality of meta-analysis and be aware of potential mistakes that researchers commonly make.

Additionally, while reducing errors, this tool also helps people develop more standardized research thinking and reporting habits. People are educated and aware of the importance of transparency and reproducibility in interpreting meta-analysis. It can in turn improve the reliability of meta-analyses and potentially propose a higher standard for future meta-analysis publication.

2. METHODOLOGY

This section details the implementation of the automatic detection tool for SD/SE error in meta-analysis. It is divided into three stages: an exploration phase, a development phase, and a validation phase. Figure 1 visulizes the workflow along with utilized packages.

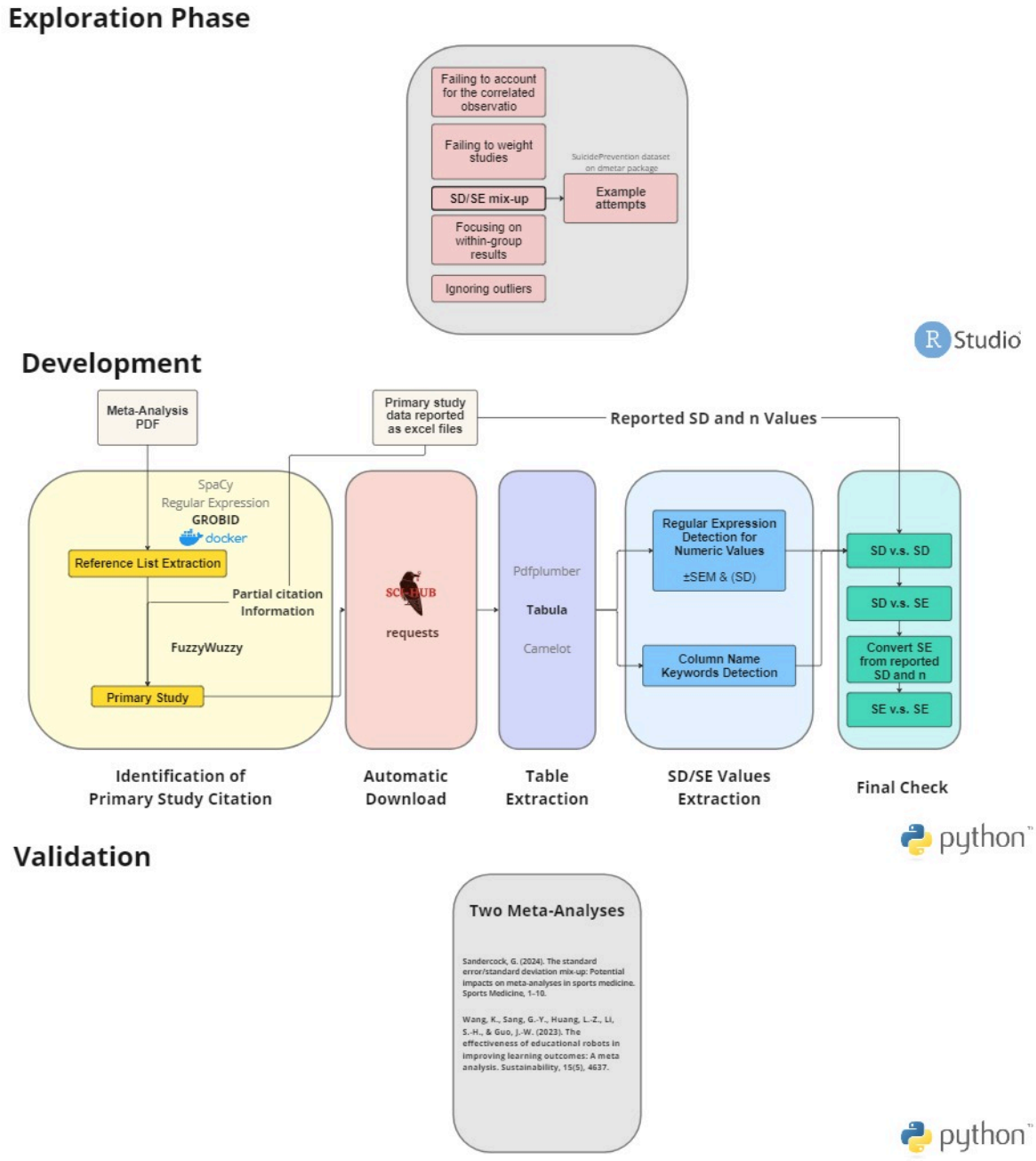

Figure 1: Entire process for the automatic detection tool

2.1. Exploration Phase

In the preliminary stage, common errors in meta-analyses are summarized mainly through literature reviews and expert meetings. Focusing on the goal of automated recognition, candidates with high feasibility are again selected from the common errors for further investigation. A total of five common errors are summarized: mixing up standard deviation (SD) with standard error (SE), failing to weight studies, failing to account for the correlated observation, focusing on within-group results, and ignoring outliers. Among them, the error of SD/SE mix-up is assessed as potential for automatic detection and further explored. The exploration concludes the necessity of obtaining cited literature data for automation implementation.

2.1.1. Common Errors

The exploration phase is aimed at identifying common errors in existing meta-analyses. According to Kadlec et al. (2023), five common five common mistakes are listed:

- SD/SE mix-up. One of the most frequent errors in meta-analysis is the confusion between standard deviation (SD) and standard error (SE) when calculating the effect. In primary studies, researchers have the option of reporting SD or SE values depending on the focus of the study. However, the calculation of effective size in the meta-analysis is strictly based on SD and miscalculated effect sizes using SE instead of SD will lead to incorrect calculations and misleading results.

- Failing to weight studies. In meta-analyses, properly weighting studies by the amount of information they provide is crucial. It is usually accomplished by weighting studies by the inverse variances of their effect estimates or by studies' validity. Failure to weight studies can result in biased outcomes, giving equal attention to all studies and skewing the overall findings.

- Failing to account for the correlated observation. In meta-analyses, some studies may contribute multiple effect sizes. This is because a study includes multiple intervention groups or multiple measurements per group. And oversighting the correlated data leads to underestimated standard errors resulting in narrower CIs and an increased type I error.

- Focusing on within-group results. Many meta-analyses incorrectly emphasize within-group results rather than between-group comparisons. This focus can obscure the true effects and interactions present in the data. However, by comparing between groups, it can remove the effects that may exist regardless of the intervention, for instance, placebo effects.

- Ignoring outliers. Outliers can significantly impact the results of a meta-analysis. And it is essential to make sure the outlier is real and not an error. If the outlier is

not an error, then the researcher should analyse the data with or without outliers to measure the overall impact.

The list was discussed and reviewed with experts, Daniël Lakens and Cristian Mesquida Caldentey, for further evaluation for feasibility. The SD/SE mix-up error was selected to be explored, focusing on how frequently it occurs and the straightforwardness of the underlying reasons for making this mistake. First, wrongly using standard errors instead of standard deviation is very common in the meta-analysis (Kadlec et al., 2023; Sandercock, 2024). Besides, as illustrated by Kadlec et al. (2023), SD/SE mix-up will lead to incorrect calculations and misleading results. Further exploration of the potential effect of this error is presented in the next section 2.1.2. Last but not the least, it was discussed to have a high potential to be automated and detectable considering the values of standard error and standard deviation. Identifying this error mainly involves extracting and comparing values, a process that is more amenable to mechanical detection than errors from research methodology or conceptual thinking. Besides that, the extraction of large amounts of data is less difficult but more tedious for manual checking, and automation can perform such tasks efficiently.

2.1.2. SD/SE Mix-Up Exploration

The reporting of SD and SE for published research does not have a strict requirement, as both are used to describe the distribution and uncertainty of the data, but they are interpreted differently. The choice of reporting SD or SE values depends on the research questions and authors. The reporting of SE rather than SD when describing sample variability appears commonplace across a number of scientific disciplines. SD are also used for within-subject or between-group designs so the error can occur in meta-analyses using within-subject designs. When it comes to meta-analysis, the calculation of standardized effect sizes requires SD. For instance, for a between-groups design, the standardized effect size is calculated as Standardized Mean Difference (SMD):

$$SMD = \frac{\overline{x}_1 - \overline{x}_2}{\sqrt{\frac{(n_1-1)S_1^2+(n_2-1)S_2^2}{n_1+n_2-2}}} \tag{2.1}$$

where x_1 is the post-test mean of the intervention group, and x_2 is the post-test mean of the control group. SD_1 is the post-test SD of the intervention group, and SD_2 is the post-test SD of the control group. n_1 and n_2 are the sample sizes of intervention and control groups, respectively.

The mistake of using standard error rather than the standard deviation from the studies cited when calculating the standardized effect size will artificially inflate the values, and it could also lead to widening or narrowing effect size confidence intervals (CIs). To illustrate the consequence of miscalculating the SMD with the SE on the pooled effect size, we used a dataset (i.e., SuicidePrevention) available on dmetar package (Harrer,

2023). This dataset consisted of 9 studies with sample sizes, mean and SD for both the control and experimental group The standard errors of the mean were calculated by dividing the standard deviation by the square root of the sample size. As shown in Equation 2.2 below:

$$SE = \frac{SD}{\sqrt{n}} \tag{2.2}$$

By utilizing 'meta' package (Schwarzer et al.) in RStudio, two meta-analyses were produced which used either only standard deviations or standard errors. Figure 2 and Figure 4 showed the meta-analysis results with standard deviation (correct) or standard error (incorrect) only, respectively. Compared with the control group(Figure 4), when misuse SE, the standardised mean differences for all studies were all numerically inflated, along with a wider range of confidence intervals. By solely reviewing Figure 4, several studies can be already identified as dubious data that the SMD values were numerically excessive (larger than 2), for instance, Devries et al., and McCarthy et al..

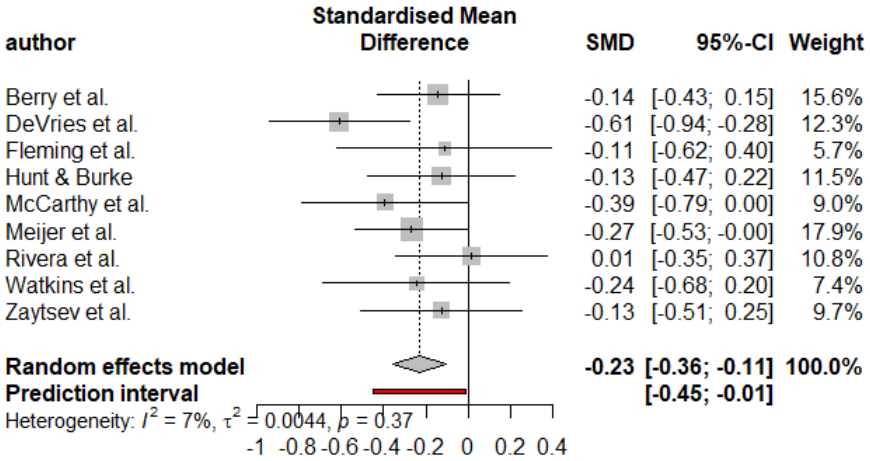

Figure 2: Correct pooled effect size calculated with SD

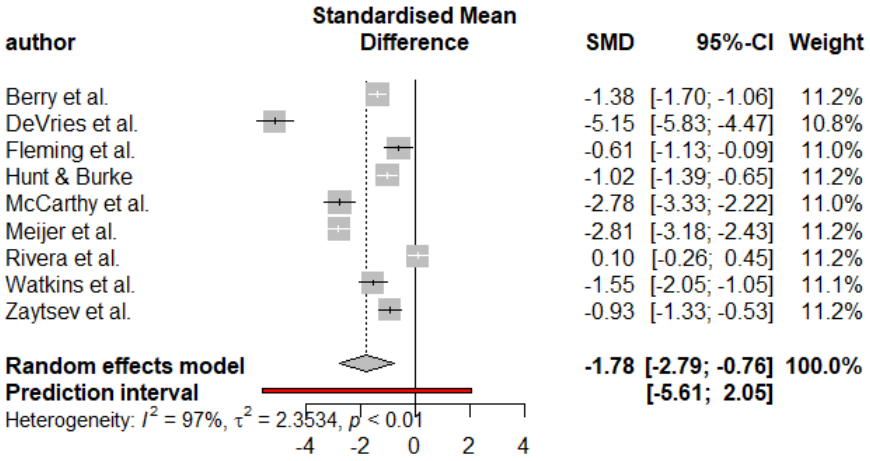

Figure 3: Deliberated incorrect pooled effect size calculated with SE

Other than that, many of the existing results would not be problematic if only Figure 2 were evaluated without informing about the error of misusing SE. For example, if a meta-analysis is performed and only those studies that incorrectly compute SMD with SE less than 1 are retained the others still compute SD correctly with SD. by only showing the resulting forest plot to the researcher and not informing him or her of the existence of the SD/SE misuse, the problem will be difficult to detect, as shown in Figure 3.

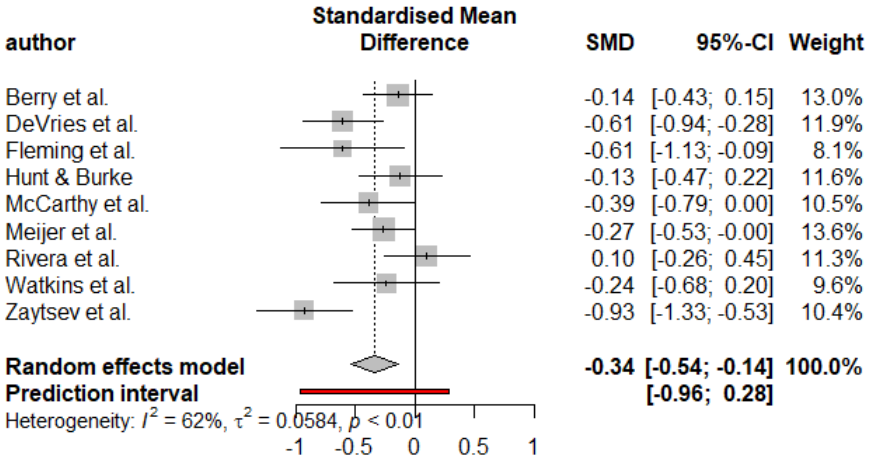

Figure 4: Partial incorrect pooled effect size that studies by Fleming et al., Rivera et al., and Zaytsev et al. used SE to calculate SMD

The effect of the misuse of SE values for single studies on the final result is also summarized. The final pooled effect size is affected by the SMD of the sign of the cited paper, and it can be exaggerated or, in some cases, diminished. For the one example that is diminished, if only the study by Rivera et al. was to misuse SE to calculate SMD, its increase in the positive value would in the end contribute to a smaller negative value for

the pooled effect size. However, it is considered a special case considering that the effect value of the study is contrary to all other experiments and can therefore be challenged as an ineligible primary paper. In summary, in most cases, the misuse of SE rather than SD exaggerates the conclusions of meta-analyses.

2.2. Identification of Primary Study Citation

To be able to identify meta-analyses that miscaculated ES by using the SE, the first step is to identify the studies included in the meta-analysis. It relies on two sources of information, one is the detailed information for the entire reference list, and the other one is the included research information that is usually in an omitted version. The task of automation is divided into two steps. The first step is to extract and parse the entire reference list from the PDF file of the meta-analysis study. The second step is to identify the primary studies in conjunction with the citation list and the partial information provided by the meta-analysis. For this step, three methods are attempted: SpaCy, regular expression, and GROBID.

2.2.1. Reference List Extraction from PDF Files

The input of the PDF files is initially pre-processed to improve mechanical readability for subsequent identification. Initially, the entire text is extracted using the PDF parsing package PyMuPDF for the PDF file as input (McKie and Liu, 2024). Then, the reference section is located through string detection of the title 'references.' The text after that is intercepted as the reference section until the 'appendix', if present, is detected. The intercepted section is processed as one continuous string with no line while removing unnecessary characters. This step is necessary due to the line-feed error from PDF parsing, in which each line is extracted separately, disregarding and breaking all the paragraph structure. Then the problem is further defined as how to recognize and divide each reference from the long continuous text.

The main difficulty of reference extraction is the readability of the text extracted from PDF files. The reference list is usually organized in a unified format within one study; and citation usually includes several pieces of information, including authors, title, year of publication, journal name, volume number, DOI, etc. However, the PDF format is complicated and makes the text file disordered, especially between lines, with the use of PDF parsing tools. Each citation breaks off between lines and does not form a paragraph, whereas it is difficult to distinguish between different citations. It makes it an unexpectedly challenging task to divide citations one from another. For example, Figure 5 shows a comparison of the extraction results for the first three citations in the PDF reference list for a sports health article (Schmidt et al., 2014). The original PDF text uses the first line indentation or not to distinguish the beginning of each quote, but the extracted text loses this structure. It is difficult to separate each citation directly even by

visual inspection for the extracted text, especially between the first and the second where the end of the first citation. It becomes challenging to summarize a general mechanical automated method for all citations.

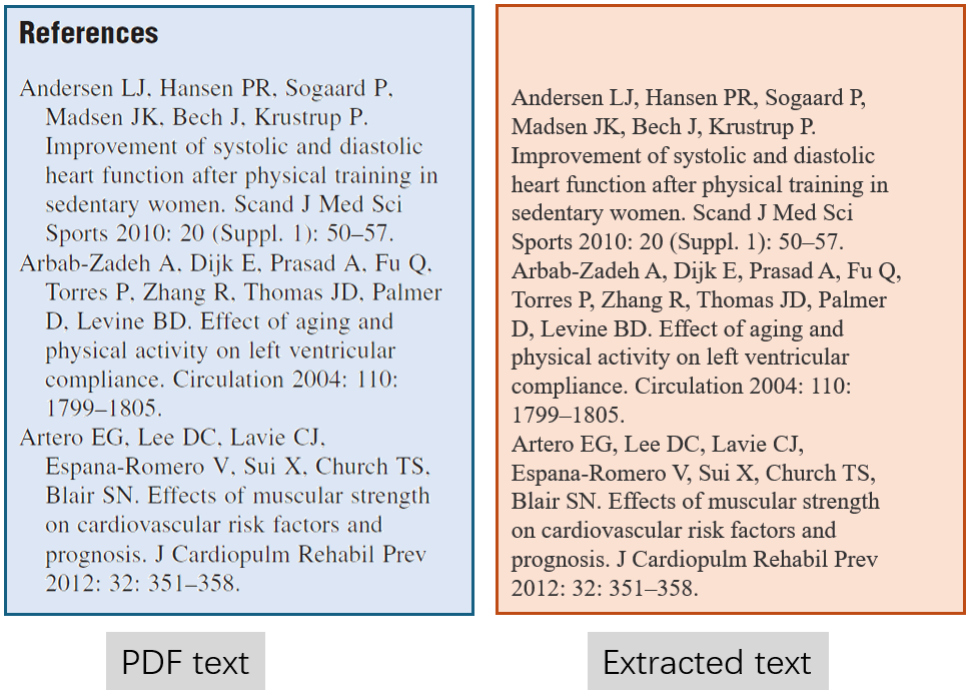

Figure 5: Citation extraction example (Schmidt et al., 2014), where the left shows the first three references to the PDF, and the right shows the corresponding extraction results of package PyMuPDF.

Distinguishing each reference can be done by identifying the beginning or end of the reference. As mentioned previously, references in published articles typically adhere to a specific structure and exhibit common patterns such as author names, publication years, journal or conference names, etc. However, the structure is inconsistent for different research paper writing formats and not all elements are mandatory. For instance, common citations may be concluded with DOI, journal, book, or access time. It makes it challenging to offer a universal solution to identify the end of the cite. Hence, in the current system, it is more efficient to prioritize the detection of the beginning of a reference, as the formatting for it is generally more consistent and predictable. Although various citation styles may differ in their formatting, they generally mandate that any citation containing an author should begin with that author's name.

SpaCy

SpaCy is an open-source Python library that provides capabilities for performing natural language processing. It can process large-scale text with multiple functions. This attempt relies upon SpaCy's function of sentence segmentation and named entity recognition (NER).

First, the sentence segmentation function of SpaCy splits the pre-processed long continuous string into sentences. Then, the named entity recognition (NER) is processed sentence by sentence. NER with the pre-trained model can recognize the existing entity, especially the name of authors in this function. Based on the result, any sentence recognized starting with the name of people will be identified as the beginning of one citation. All the next sentences are added to that citation until a new starting sentence is detected.

However, SpaCy's NER in practice is not always effective due to the variability of citation styles and cultural diversity in author names. For example, SpaCy's model can accurately identify 'Kerstin Dautenhahn' as an entity name of persons. However, SpaCy fail to recognize when it is reported as 'Dautenhahn, K.' following the APA citation style. To improve the accuracy with the current pre-trained model, one additional check step is added. When a summary citation has more than four sentences, the additional check is initiated to set any subsequent sentences containing personal entities as the beginning of a new citation. This step is based on consideration of the information pieces contained in common references, which means that more than four sentences (authors, title, journal title, issue, etc.) inside one citation are possible to be two or more citations mixed. The additional check has higher accuracy especially when there are multiple authors for one citation. Take the pdf file of the sports health article by Schmidt et al. (2014) as an example, there are 38 citations listed in it. A total of 29 references were extracted with the introduction of the extra check, while only 16 were extracted without it. However, there is also a potential error in the introduction of an addition check. It may also lead to the misrecognition of a title that has a person's name. Overall, additional checks improved accuracy, but many citations could not be identified due to the recognition difficulties of SpaCy. Another library Flair providing similar pre-trained NER models was tested and the results improved but were still unsatisfactory (Akbik et al., 2019). For the same sample of PDF (Schmidt et al., 2014), Flair extracted 33 compared to SpaCy's 29, but one of the 33 references was checked to be incorrect. Given the purpose of our study, instead of using an existing model, building the training set to train a specialized training model could be a potential solution.

Regular Expression

Regular expressions are sequences of characters that form search patterns, primarily used for string matching and manipulation. For one specific citation format, there is a high degree of consistency in the list of references that rendering the use of regular expression an effective method to identify references.

The Vancouver style, which is a citation style widely used in the health and biomedical sciences, is primarily focused on and explored. In the Vancouver style citations, the formatting of author names is consistent and meta-analysis is very common in medical research, making it an ideal candidate for regular expression-based parsing (Moher et al.,

2010; Garg et al., 2008). The formatting of the authors' names is relatively simple as shown and explained with an example below (Nyberg et al., 2012).

Nyberg M, Blackwell JR, Damsgaard R, Jones AM, Hellsten Y, Mortensen SP. Lifelong physical activity prevents anage-related reduction in arterial and skeletal muscle nitric oxide bioavailability in humans. J Physiol 2012: 590: 5361–5370.

The author's names are listed with the last name followed by initials, and different authors are separated by commas. The regular expression implements single-person identification by recognizing a word with an initial capital letter followed by one or more capital letters. The same pattern can be utilized to identify multiple authors, separated by commas. The period is required for the end of the sentence to ensure the match is the author section. Optional 'et al' is also included before the period for the potential ending.

Some special cases were encored in the actual tests and the regular expression was adjusted for them. First, it happens to some surnames containing special punctuation like hyphens or apostrophes, for example, Beck-Nielsen and O'Connor. Second, compound surnames can have more than one part, for example, Van Gogh. Third, in addition to the 26-letter Latin alphabet, accented characters and other unique characters can be used for different cultural backgrounds, for instance, Weiß. Similarly, ligatures also exist in surnames that cannot be identified as regular characters. Because the regular expression has strict requirements to match the desired pattern, adaptations are made for all possible scenarios that arise. For the first two cases, corresponding optional parts are added to the regular expression. For special characters, instead of matching the 26-letter Latin alphabet, the Unicode character matching from the 'regex' module is used to include authors from different cultures. All these cases are universal for the surname so can be extended to all cite styles.

The advantage of regular expression is the high accuracy with the well-elaborated pattern. The final pattern has a high accuracy for some papers using the Vancouver style citations, along with some variations like the AMA (American Medical Association) style. However, in order to achieve final automation, it requires considerable time and workload to analyze and set up patterns for each citation style. In addition to this, it also needs to be coded to detect the specific citation format that the literature follows to fit the most appropriate regular expression.

GROBID

GROBID (Generation of Bibliographic Dat) is a machine learning library for extracting, parsing and restructuring raw documents such as PDFs and can transfer them into structured SML/TEI encoded documents ("GROBID", 2008–2023). It provides various functionalities for information extraction in scholarship articles and is effective for extracting and parsing references, making it especially valuable for this research task. The result of the reference extraction can be exported as a whole, and the information

for each reference (title, authors, journal title, issue, number, etc.) is tagged separately (Lopez, 2009). However, GROBID currently has limited support for Windows due to the development schedule that it is suggested to use the GROBID Docker image. The deployment and packaging of GROBID for Windows requires considerable learning costs compared to the previous two methods.

GROBID has two main advantages compared with the other two approaches above. First, GROBID demonstrates exceptional performance in handling the high variability of bibliographic formats and presentations, a challenge also outlined in the preceding section. In practice, it performs well for the extraction of different citation formats. Second, the well-structured result largely enhanced the accessibility and machine-readability of scientific information. By converting PDF documents into standardized XML/TEI encoded formats, GROBID facilitates easier citation retrieval and manipulation, especially for the following steps to identify the included studies in the meta-analysis. In conclusion, among the three methods discussed and considered, GROBID is ultimately selected for implementing the extraction of the reference list.

2.2.2. Identification of Primary Studies

With the list of references successfully exported via GROBID, the next step is to identify the included articles in the meta-analysis. Meta-analysis usually includes systematic reviews and not all studies identified in the SR end up in the meta-analysis. Also, the meta-analysis articles cite other studies and tools (which packages were used) that are not part of the meta-analysis. If the included studies are not listed entirely and separately, Meta-analytic researchers often briefly represent them by information such as researcher name, year of publication, and citation number. This step matches in all references to confirm the included studies in the meta-analysis. It mainly involves character matching between all extracted references and the partial information provided by the meta-analysis.

For literature that provides specific data for meta-analysis, the primary study is often referenced by partial information, but ensuring that specific literature can be manually identified. There is variability in reported information depending on the reporting habits of the researcher. Table 1 presented two existing examples of reporting primary studies. They are both parts taken from published tables (Sandercock, 2024; Wang et al., 2023). For both examples, they used two pieces of information to ensure that they could uniquely refer to the included literature. The example on the left is identified by the surname of only the first author and the number of citations. The example on the right, on the other hand, lists all authors and the year of publication. Notably, Casad & Jawaharlal (2012.1) and Casad & Jawaharlal (2012.2) are two studies reported in the same article, which means they refer to the same citation in this case, or the same author published two studies within the same year.

To accommodate various reporting habits, the recognition of primary studies is

Table 1: Two examples of reporting primary studies

Study	Study
Andersen [38]	Ajlouni (2023)
Andersen [32]	Al Hakim et al.(2020)
Møller [44]	Casad & Jawaharlal (2012.1)
Uth [42]	Casad & Jawaharlal (2012.2)
...	...

achieved by approximate string matching between the citation list and the partial information. For the reference list, the converted standardized XML file from GROBID is read and each part of one citation is parsed and saved. All parts of a reference are combined into one continuous string for subsequent matches; and references are grouped by if it has three or more authors. As for the partial information, the column name containing 'study' from the Excel file is detected, and all strings below that column name are utilized as matching strings. The first step is to check if 'et al' is in the partial information. If so, 'et al' will be removed from the string and this study will only match with citations with three or more authors. Then it is processed to remove all punctuation and divide the string into sub-strings of multiple keywords.

The approximate string matching is achieved using a Python library FuzzyWuzzy. It compares strings and has algorithms quantifying the similarity. For this task, among multiple algorithms, the partial ratio is chosen to identify the primary study. The partial ratio method compares the similarity of a substring with a longer string to determine the highest degree of overlap and match accuracy. And it places more emphasis on substring matching and is more tolerant of differences. The partial score ranges gives a score range from 0 to 100, where 0 indicates no similarity, and 100 indicates an exact match. By calculating the average similarity value of all keywords of an partial information string, it derives the extent to which partial information appears in each citation. Each partial information matches the entire citation list. A threshold value of 80 is set to ensure the accuracy of the matching citation and the one with the highest similarity score is saved as the final match. However, when there is no citation with a similarity score higher than 80, it will be reported and log as no citation found.

2.3. Automatic Downloading of Primary Literature

After successfully identifying primary literature citations, the next step of literature tracing is to automate the primary literature downloads. This step is crucial as it aims to procure the PDFs of primary studies for the primary studies, which contain data for SD/SE values. Although the above section mentions that the PDF format brings many or most of the difficulties in automating the text processing process, the PDF format is still the most standardized form of raw academic article documents. For the validity

of the subsequent SD/SE data and given the wide availability of PDF documents, this step explores the possibility of using the original citation as input to download a PDF automatically.

The primary literature citation information requires no additional processing for tracing. Benefiting from GROBID's expertise in processing and parsing references, each identified primary literature citation has been well-structured and the information for each reference (title, authors, journal title, issue, number, etc.) is tagged separately. In all citation information, the Digital Object Identifier (DOI) serves as a unique identification code for the academic literature that can provide direct and stable access to the original document. However, not all the references have a DOI or report the DOI in the reference list. In this process, the title of the article is used as the main input to track the links to the documents.

In early explorations, direct downloads via literature links encountered many problems. The automated assessment of links to academic literature based on titles was conducted through Google Scholar. And with the Selenium Python library that simulates browsing actions, it is successful in automating web browsers and downloading open assessment articles with PDF links on the search results page. However, for literature links, there are two main challenges.

- **Website Accessibility**. Many studies are subscription-based, resulting in limited access even when a match is found. On top of that, different scholarly article online platforms have systems that hinder automated access to varying degrees. For example, many platforms restrict access by monitoring IP addresses and limiting the number of requests from a single IP within a short time frame. Another example is the use of login requirements and CAPTCHAs (Completely Automated Public Turing test to tell Computers and Humans Apart) to prevent automated access.
- **Structural Diversity**. Different platforms have different layouts, especially for the download function has a variety of designs. Besides the simple button download, some platforms offer downloading in the hidden menu, while some even jump to built-in online PDF reader.

These two difficulties make it no universal solution to achieve automated downloads but require adaption individually for each platform. Given the multitude of existing academic platforms and the possible changes in the structure, the feasibility of this option is highly questionable.

As an alternative, automated downloads were conducted via Sci-Hub. Sci-Hub is a widely known platform that provides unrestricted access to a vast repository of scholarly articles by bypassing paywalls. Using this approach enables a unified download solution and automation is technically accessible with a simple Sci-Hub web design. The implementation of this function relies on the Python requests library which simply

uses the requests library to construct a post request with the title of the article, then the PDF file is automatically downloaded by parsing the source address on the detail page. To facilitate subsequent comparisons of SD/SE values, the downloaded PDF files were named after previously retrieved partial information. All failed downloads of the primary cited documents were logged and need to be handled manually for subsequent full detection of all included studies.

2.4. Table Extraction

For the cited original studies, one of the most essential ways to extract data is from the table. Based on a literature review for one published meta-analysis and all cited original studies, all SE/SD values are tabulated in all sorts of forms if they were reported in the article.

To achieve the table extraction, there are three existing Python libraries: Pdfplumber, Tabula, and Camelot. However, Camelot is only accessible for Linux, which is excluded from this project. Additionally, both Pdfplumber and Tabula only work on text-based PDFs, not scanned documents, which means that they cannot extract tables in any form of images, which is encountered in practice. The following section provides the results for the comparison between PdfPlumber and Tabula.

2.4.1. Pdfplumber

Pdfplumber is an easy-to-use package that offers page-by-page table extraction. It also extracts quickly with little time, which is a feature to consider if the goal is to extract data from a large set of studies. However, after several attempts with different papers, the outcome of Pdfplumber is considered to be inefficient in that it recognises very few tables and more often than not it gives zero outcomes. Additionally, tables extracted from Pdfplumber often exhibit disorganized or inaccurate structures. For example, Table 2 shows the extracted table from PDF of the result for robot effect on kinds by Hyun et al. (2008). In the extracted table, the data structure is lost by combining all data into one column.

Table 2: Example of extracted table with incorrect structures by Pdfplumber

TABLE VIII **THE RESULTS OF THE VOCABULARY TEST (PPVT-R)**
Pretest Posttest Group t Mean (SD) Mean (SD)
Experimental Group 4.44 (.61) 5.20 (.83) 6.032* Control Group 4.38 (.38) 5.00 (.64) 4.243*
* p <.05

2.4.2. Tabula

Tabula is a JAVA-based tool, and the version used for this project is tabula-py, which is a simple Python wrapper. Tabula uses a separate algorithm to extract tables from text-based PDFs that provide output that differs significantly from those of pdfplumber. Firstly, the number of tables extracted is larger and the tables are more complete. Although tables are relatively more accurately structured, the output is still often incorrect when it comes to complex tables. For example, when a table has irregular multi-row header rows, its sub-header rows can easily be recognized as one column where several data are combined and listed in one. Secondly, text can sometimes be incorrectly recognized as a table. This could have been detected and removed as there tends to be a lot of text contained within one single column. However, there are also cases where tables are mixed with text in one single column, which makes it impractical to drop all wrongly identified tables within only one column. Thirdly, tabula-py takes a relatively longer time than Pdfplumbe to run, usually from seconds to a dozen seconds, depending on the pdf file's page.

The summarized version for comparison results can be found in Table 3. A positive sign indicates that the tool has a positive attribute for that feature, while a negative sign indicates that the tool has a negative attribute.

Overall, Tabula has the best performance of all the available packages, especially as it maximises the extraction of tables containing data. However, it has two notable flaws that impact the subsequent extraction of SD/SE values. Primarily, Tabula also encounters limitations in its capacity to extract certain tables. In addition to its inability to non-textual tables, it fails to extract some textual tables, either simple or complex structured. The reason for this failure remains unclear, and consequently, any data contained within these unextractable tables will ultimately remain undetected. Secondly, inaccurate recognition of complex tables and the mixing of body text and tables hindered the detection of values. The specific manifestation of this is that multiple data, or even one entire row of data will be recognised and merged into one. It especially affects the identification of variable names by the column name. The next section describes approaches for identifying and extracting SD/SE values in a table, while the issues arising from table extraction are further elaborated.

Table 3: Comparison results for Pdfplumber and Tabula

	Pdfplumber	Tabula
Extracted Table Number	–	+
Structure Accuracy	–	+
Complex Tables	–	–
Running Time	+	–
Misidentifying	–	–

2.5. SD/SE Detection and Extraction

There are two approaches to detecting standard deviation and standard error values from tables. The first is keyword detection, for instance, std. dev., in column names, and the second is regular expression detection for numeric values. For regular expression, there are two common ways summarised for standard deviation and standard error: '±SD/SE' and '(SD/SE)'. For each extract table, both approaches are applied and all detected data, if any, are saved together and named based on the detection results as either standard deviation or standard error.

2.5.1. Keywords Detection

The first approach to detect the standard deviation or standard error values in tables is to detect keywords in table column names. According to reviews of the literature, one common way to report SD/SE values is to report them in a table as an individual variable with specific variable names. The list of keywords is summarized in Table 4. All keywords are coded in the form of exact match patterns for regular expressions. It only detects the precise words; for example, the column 'mean sd' will be detected while the column 'PTSD' will not. A downside of this approach is that is case sensitivity, and therefore the regular expression ignores case sensitivity.

Table 4: Keywords for standard deviation and standard error

Variables	Keywords
SD	'SD', 'standard deviation', 'std dev', 'std. dev.','S.D.'
SE	'SE', 'SEM', 'standard error', 'std err', 'S.E.'

For the extracted tables, because of the extraction issues as mentioned in the last section, it is coded to detect the pre-set keywords as shown in Table 4 for all cells, instead of only the first column. Once the SE or SD keyword is retrieved, the data for that row is extracted individually until the entire table is retrieved. Eventually, on a per-table basis, where the extracted SD rows and SE rows are saved in two separate named files.

2.5.2. Regular Expression Detection for Numeric Values

Besides detection for the column names, the other approach to extract SD/SE values is conventional expression detection for numeric values. In addition to the usual reporting methods, some researchers are reporting SD/SE values with symbolic plus-minus signs (±) and enclosing them in round brackets.

Similarly, this method is applied to all cells in the extracted tables. The corresponding regular expression is used to extract any existing values behind the plus-minus signs or inside the round brackets in the table. In addition to extracting the values, it is also

necessary to know whether the values represent SDs or SEs. In order to achieve that, if a value that matches the regular expression is detected in a chart, all the text is retrieved on the same page of the PDF where that table is located. Similar to keyword search, two sets of keywords corresponding to SD and SE will be matched on a full-page text, for instance, '±SEM', and '(SD)'. The reason for detecting full-page text is to ensure accuracy. Those keywords relevant to SD/SE are frequently embedded within the comments of tables. However, table extraction often fails to include them in the table. In the end, all values extracted from the different tables will be saved and labelled with the corresponding SD/SE.

2.6. Final Check

With all the SD / SE values extracted, the final step is to check for misuse of SD / SE in the meta-analysis. This step mainly consists of matching the SD values provided by the meta-analysis for each original experiment against all the extracted SD/SE values. If a successful match is made, it can be concluded that there is or is not SD/SE misuse. However, when there is still no match, it is required to calculate the SE values using the formula (2.1) in Section 2.1.2, and then the calculated SE values are matched with the extracted SE values. This step is designed to check for the case of the researcher's correct use of the formula to calculate and use SD.

Once the SD/SE values have been extracted, there are four possible outcomes:

- SD was used correctly:: reported SD matches with the extracted SD value.
- Misuse of SE instead of SD: The reported SD does not match any of the extracted SD values but it matches the extracted SE values.
- Use SD correctly after manual convert from SE: The reported SD value does not match any of the extracted SD values, and the calculated SE value matches with extracted SE value.
- No matching SD/SE value is extracted for the paper

To improve comprehension, the flow chart delineates the procedural steps involved in verification, as shown in Figure 6. The flowchart provides a visual representation of the sequential checks and all four possible results. For the four possible outcomes, the green grid indicates that the SD/SE mix-up does not occur, the red grid indicates that the SD/SE mix-up error detected and the purple grid indicates that the automated tool is inefficient for this literature.

The pdf and data used in the first of these meta-analyses on the effectiveness of soccer on improving VO_2max (Milanović et al., 2015) were taken from another article addressing its SD/SE mix-up error by Sandercock (2024)

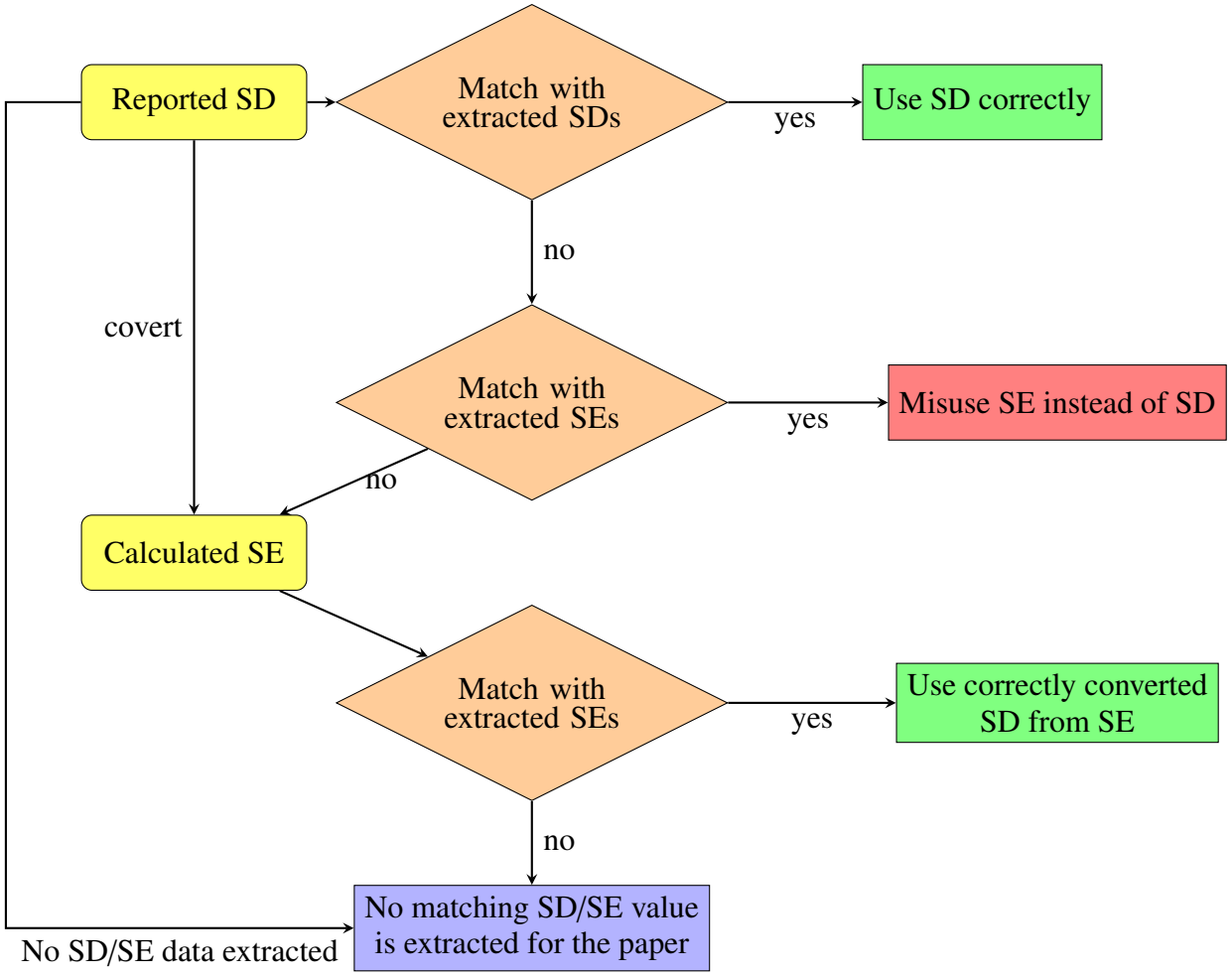

Figure 6: Flowchart illustrating the process of final check and possible outcomes

2.7. Validation

After successful tool development, the effectiveness of the tool is tested with two meta-analysis articles from different domains. For each meta-analysis, the results are examined step-by-step following the methodological process, and the automated results are manually checked against the meta-analyses and the PDF files of the main studies. Each step is run individually and the results of each step are organized and reported separately.

The two meta-analysis reports were written by Wang et al. (2023) and Sandercock (2024), and both of the reports contain all the data needed for automated identification tools. There are two main considerations when choosing the validation samples. The first one is the transparency of the data in the meta-analytic reporting. To achieve automatic detection, the automated tool requires the used SD and n values of the primary study as input, and all meta-analyses that do not meet the requirements are excluded. The second consideration is the diversity of reporting styles. The two meta-analyses are purposely selected in order to test the effectiveness of this automated tool on domains that require different writing styles.

The meta-analysis by Wang et al. (2023) investigated the effectiveness of educational robots. The entire meta-analysis incorporates a total of 34 primary studies, all of which are between-group designs. All the necessary data about the primary study are reported and manually extracted from the forest plots in the PDF file into an Excel document.

The other meta-analysis by Sandercock (2024) focused on the health domain. This study is a reproduction of a meta-analysis by Milanović et al. (2015) which investigated the football effect of exercise on adult VO_2max. The purpose of this reproduction was to assess the impact of the SD/SE mix-up error on the results of the meta-analysis by Milanović et al. It provides detailed data in supplemental documentation and summarizes the existing SD/SE mix-up error, which makes it a perfect sample for verification. In the meta-analysis, only the football vs. non-exercise is used for verification and a total of 21 primary studies are included. These twenty-one experiments are all between-group designs.

After applying the automatic tool to these two meta-analyses, the results for each step are manually checked that measure the correctness. The results of the validation are analyzed both quantitatively and qualitatively. The quantitative analysis focuses on summarizing the effectiveness of the tool, and the qualitative analysis concentrates on failure cases and special situations. Specific validation results for each section are illustrated in the next section.

3. RESULTS

This section introduces the results and findings of the automatic tool to detect SD/SE mix-up errors in meta-analyses. As described in the "Methodology" section above, automated detection requires multiple subtasks. For the results section, each subtask was validated, and the findings were illustrated correspondingly. The validation of the automated tool was conducted using two distinct meta-analyses, one from the health domain (Sandercock, 2024; Milanović et al., 2015) and the other from the human-computer interaction domain (Wang et al., 2023). The validation results for each step covered both quantitative and qualitative analyses. The quantitative analysis offered an overview of the overall performance, while the qualitative analysis delved into specific cases to provide a detailed understanding of the tool's effectiveness. The findings underscored the tool's potential to enhance the accuracy and reliability of meta-analyses by systematically identifying instances where SD and SE values have been mistakenly interchanged.

3.1. Literature Tracing

For the literature tracing section, the final outcome was the primary studies in the form of PDF files, and it involved several steps of processing. This section presented the overall validation results and the specific analysis for sub-step outcomes.

3.1.1. Overall Results for Automatic PDF Downloads

Table 5 presents the validation results for the number of primary studies that could be downloaded from the reference list of two meta-analyses. All citations that have their PDF files downloaded are tagged as save PDF correctly. There are two cases of failure, one is that the literature is not collected in Sci-Hub, and the other is that the extracted PDF download link is invalid. Based on the log report, those articles that cannot be found on Sci-Hub are tagged as not available on Sci-Hub. All links that are found on Sci-Hub but can not be downloaded were marked as automatic download failures.

Table 5: Validation results for automatic downloads

Meta-Analyses	Number of Primary Study	Save PDF Correctly	Not Available on Sci-Hub	Automatic Download Failure	Success Rate
Wang et al., 2023	34	17	16	1	50.00%
Sandercock, 2024	21	18	1	2	85.71%
Overall	55	35	17	3	63.64%

Based on Table 5, the results varied widely in success rates for the two meta-analyses. For meta-analysis by Wang et al.(2023), the success rate of download was only 50.00% (17 out of 34), while the main failure reason was that desired articles were not available on Sci-Hub. However, for the meta-analysis of the effects of exercise on VO_2max in adults (Sandercock, 2024, the success rate of 85.71% was considered to be effective that 18 primary articles out of 21 were successfully traced and downloaded as PDFs.

However, during the analysis, it was found that several studies provided multiple effect sizes in the meta-analysis of the effectiveness of educational robots, and thus much of the literature traced back was repetitive. For example, Z.-W. Hong et al. (2016) studied the effect of the designed robots using descriptive statistics that have multiple items. 5 items such as speaking, listening, and reading were included in the meta-analysis as 5 separate primary studies (Wang et al., 2023). The 34 included studies were drawn from 16 academic articles. For better validation analysis, after manually correcting the number of articles to be retrieved, a modified result is displayed in Table 6.

Table 6: Modified validation results for automatic downloads

Meta-Analysis	Modified Number of Primary Study	Save PDF Correctly	Not Available on Sci-Hub	Automatic Download Failure	Success Rate
Wang et al., 2023	16	10	5	1	62.50%
Sandercock, 2024	21	18	1	2	85.71%
Overall	37	28	6	3	75.68%

The corrected results have positively improved the success rate of the meta-analysis by Wang et al. (2023) from 50% to 62.5%. In contrast, the meta-analysis reproduced by Sandercock (2024) did not address the situation where a single article provided multiple effects, so the results were unchanged. With the decrease in the total number of primary articles and the increase in the success rate of the robot meta-analysis, the total correctness rate also increased from 63.64% to 75.68%.

3.1.2. Correctness of Citation Information Extraction

In addition to the successful download of the PDF file, the match between the PDF file and the citation information was verified. For both meta-analyses, all downloaded PDFs and the list of citations in the meta-analysis were manually checked to verify the correctness of the literature tracing. Only 2 of the 55 primary studies was verified to be inconsistent, with a 96.36% correct rate. It presented the high accuracy and effectiveness of the reference list and the function of the fuzzy matching, given that the two meta-analyses followed different citation rules. The meta-analysis by Wang et al. (2023) used APA style, while the meta-analysis reproduced by Sandercock (2024) used Vancouver style.

Upon review, one of the two incorrect citations was due to Meta-analysis reporting errors, while the other was due to a program matching error. First, the study of Ortiz et al. (2016) was incorrectly matched because the year of publication was written incorrectly as 2017 in the meta-analysis forest plot (Wang et al., 2023). Second, the matching information for the other false one was 'Hong et al. (2016.1)', where 2016 is the year of publication and 1 represents the first study included in this article. It incorrectly matched another citation published in 2011, and it has the first author who shares the same surname (J.-C. Hong et al., 2011). However, the other four ionformations for the same paper from 'Hong et al.(2016.2)' to 'Hong et al.(2016.5)' were all correctly matched (Z.-W. Hong et al., 2016).

Additionally, a review of GROBID's exported citation lists in TEI/XML documents revealed a problem with the identification/counting of primary studies. For both meta-analyses the derived coded reference list maintained a high level of formatting and correctness, but the extracted list from Wang et al. (2023) showed that GROBID incorrectly identified the forty-eighth citation as two separate citations during the parsing process. This error affected both the citation at number forty-eight and subsequent citations. Firstly, the citation at number forty-eight was incorrectly parsed, resulting in incomplete citation information. Secondly, all subsequent citations in the TEI document had incremented by one value. The error in reference list extraction had no effect on the matching of the primary studies in this validation because that meta-analysis did not use citation labels to refer to included learning. Further explanation will be given in the subsequent discussion chapter.

3.2. SD/SE Mix-Up Error Detection

All data provided by the meta-analysis were matched to the extracted SD/SE to validate the effectiveness of the automated tool in identifying SD/SE mix-up errors. These PDF files were automatically downloaded as illustrated above, except for one which required manual download to complete. The two incorrectly matched primary studies mentioned above were also manually replaced with correct PDF files.

3.2.1. Final Check Results

Figure 7 illustrates results between both meta-analyses. Failure to extract a value was classified as a failure of automatic detection, while the other three cases were classified as achieving detection. For the meta-analysis by Wang et al. (2023), out of 34 effect sizes, 17 were identified as correctly using SD values, while the other 17 failed to extract any matching SD/SE values from the PDF files. The detection rate of the automated tool for this meta-analysis was 50.00%. As for the reproduced meta-analysis by Sandercock, 15 out of 21 ES coud be matched to the extracted SD/SE. The detection rate for this meta-analysis was 71.43%. For the 15 successful detections, one effect size was tagged

as SE/SD mix-up that both reported SD values (for the experimental and control groups) matched with extracted SE values. There were 9 effect sizes tagged as using SDs correctly. By matching the calculated SE values with extracted SE values, five studies were tagged as use the correctly converted SD from SE. The overall detection rate was 58.18%, with 32 out of 55 effect sizes being successfully checked for SD/SE mix-up errors with the automatic tool.

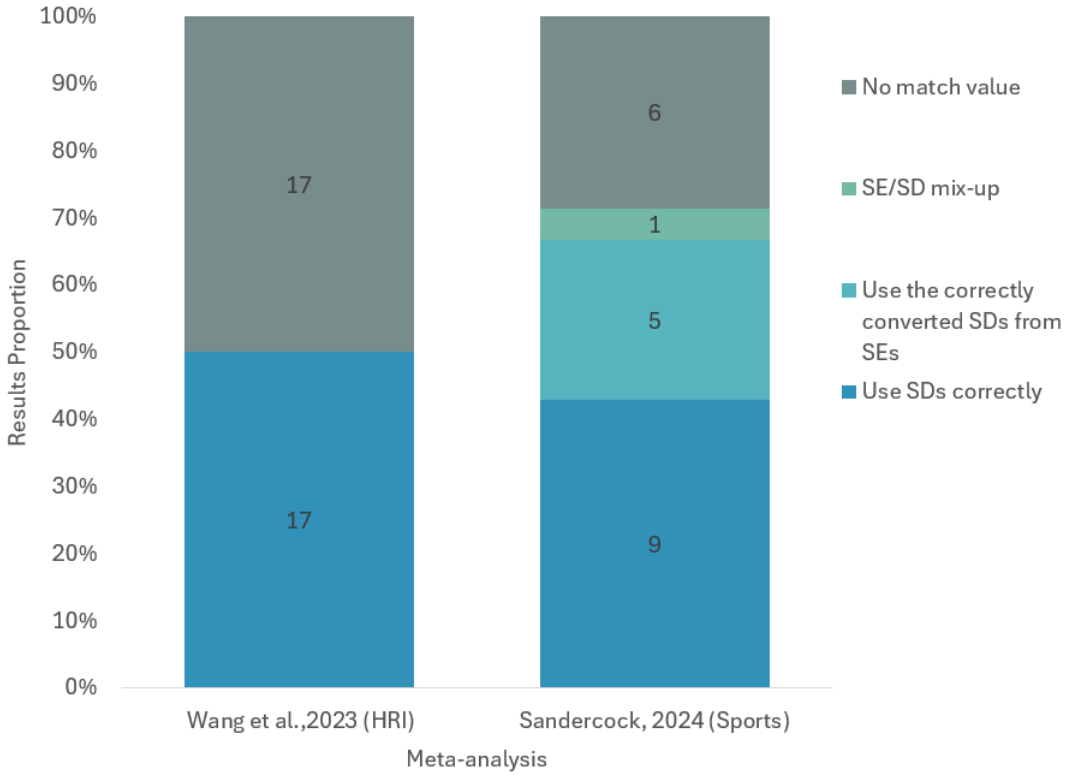

Figure 7: SD/SE check analysis across different results

The detailed results of the final check validation for each specific primary study are documented in Table 7 and Table 8. Both tables contain data used and released by the researchers in the meta-analysis (Sandercock, 2024; Wang et al., 2023). It includes the mean, SD and N for the experimental and control groups, where the former being the experimental group and the latter the control group in the tables. The column of Study is the partial information also provided by the meta-analyses, which were also used for automatic primary study tracing in the previous step. The results column summarizes the results of the automatic detection. Mix-up means SD/SE mix-up error, converted SD means using the correctly converted SD from SE, use SD means using SD correctly, and no match means no match value found for either SD or SE. In addition, the remarks column provides a summary of the manual verification and findings. The quantitative analysis of specific examples will be discussed in the following sections.

Table 7: The detailed results and notes for meta-analysis reproduction by Sandercock (2024)

Study	Mean	SD	N	Mean	SD	N	Results	Notes
Andersen [38]	32.5	6.2	15	30.7	5.9	10	Mix-up	Incorrect. Extracted SEM values for another variable happen to equal the used SD values
Andersen [32]	32.0	5.7	9	30.1	6.8	8	Convert SD	Column name
Andersen [35]	34.1	3.3	12	27.7	7.5	9	No match	One of the SEM value for weight has to be equal to the used SD value
Andersen [25]	34.6	5.8	20	31.4	5.8	11	Use SD	±standard deviation
Barene [27]	34.3	5.5	37	32.7	6.7	35	Use SD	Keywords detection for column name
Barene [26]	34.0	5.5	31	35.2	6.7	34	No match	Column name is 's' that cannot extract
De Souza [36]	25.8	5.2	19	22.9	4.6	15	Convert SD	±SE
Knoepfli-Lenzin [30]	50.1	4.7	15	44.1	5.8	17	No match	± but the table extraction is a mess, tried with pdfplumber but no improvement
Krustrup [24]	44.6	3.5	12	42.1	5.7	10	Convert SD	(SE)
Krustrup [22]	44.6	3.5	12	42.2	5.7	10	Convert SD	±SE
Krustrup [23]	37.7	4.8	19	32.7	4.2	12	No match	71 tables extracted and all a mess
Krustrup [33]	37.7	5.7	9	37.3	6.3	9	Convert SD	±SE
Randers [37]	39.1	3.5	10	38.8	7.4	7	No match	Messy tables and '±' were lost after the chart was extracted.
Randers [28]	42.9	5.6	22	36.7	5.6	10	Use SD	Incorrect. VO2 SD was not reported but the same value was extracted by ±SD
Schmidt [29]	32.5	3.3	9	30.8	3.3	8	Use SD	± SD
Uth [31]	32.5	3.0	29	26.9	3.0	28	Use SD	Incorrect. VO2 SD was not reported but the same value for other variables was extracted
Beato [41]	45.1	2.8	10	41.1	4.6	14	Use SD	±
Møller [44]	23.5	5.9	36	22.2	5.4	32	Use SD	(SD)
Uth [42]	29.3	6.4	35	26.1	5.9	19	Use SD	Correct but did not recognize (SD) as the title only has 'SD' but no '(SD)'
Milanovic [45]	47.4	8.9	20	36.9	9.5	23	No match	± value detected no text description so no data extraction.
Skoradal [43]	25.8	4.8	27	22.4	5.3	27	Use SD	± sd

Table 8: The detailed results and notes for HRI meta-analysis by Wang et al. (2023)

Study	Mean	SD	N	Mean	SD	N	Results	Notes
Ajlouni (2023)	3.20	0.24	25	2.86	0.25	25	Use SD	± SD
Alemi et al. (2015)	3.48	0.52	30	3.00	0.77	16	No match	Tables were not extracted well
Al Hakim et al. (2020)	207.46	28.67	35	189.60	26.10	35	No match	Did not report SD in the paper
Casad & Jawaharlal (2012.1)	8.47	1.83	174	6.95	1.63	86	No match	Not reported in tables but text
Casad & Jawaharlal (2012.2)	4.82	0.86	65	4.48	0.90	66	No match	Not reported in tables but text
Chen et al. (2013)	80.83	23.33	30	68.00	20.81	30	No match	Table were not extracted well
Han et al. (2008.1)	5.50	1.13	30	4.61	1.37	30	No match	No table extracted
Han et al. (2008.2)	4.47	0.68	30	3.40	0.72	30	No match	No table extracted
Han et al. (2008.3)	5.50	1.13	30	4.80	1.09	30	No match	No table extracted
Han et al. (2008.4)	4.47	0.68	30	3.57	0.82	30	No match	No table extracted
Hong et al. (2016.1)	36.22	5.76	25	31.11	7.71	25	Use SD	Column name as 'SD'
Hong et al. (2016.2)	11.20	1.63	25	11.11	2.03	25	Use SD	Column name as 'SD'
Hong et al. (2016.3)	26.40	4.66	25	20.56	6.27	25	Use SD	Column name as 'SD'
Hong et al. (2016.4)	15.72	3.16	25	14.11	4.62	25	Use SD	Column name as 'SD'
Hong et al. (2016.5)	135.20	6.99	30	107.66	20.65	30	Use SD	Column name as 'SD'
Hsiao et al. (2015)	7.77	7.17	25	7.80	4.74	25	No match	Did not report in the paper, converted by the second post and pre-tests
Hsieh et al. (2022)	207.46	28.67	35	189.60	26.10	35	Use SD	Column name as 'S.D.'
Hyun et al. (2008.1)	2.47	0.86	17	1.86	0.54	17	Use SD	Reported as (.86) but still detected by regular expression
Hyun et al. (2008.2)	2.47	1.05	17	1.31	1.28	17	Use SD	(SD)
Hyun et al. (2008.3)	5.21	0.83	17	5.00	0.64	17	Use SD	Reported as (.83) and (.64) but still detected by regular expression
Hyun et al. (2008.4)	8.35	5.93	17	3.76	4.52	17	No match	Table did not extracted well. There are 6 tables on page 5
Julià & Antolf (2016)	7.90	8.95	9	2.00	9.40	9	No match	Converted mean and SD with pre and post-tests

Continued on next page

Table 8 – continued from previous page

Study	Mean	SD	N	Mean	SD	N	Results	Notes
Korkmaz (2016)	66.80	16.90	26	45.20	16.00	26	No match	1. Table form is messy, 2. Written as 16,9 and 16,0
La Paglia et al. (2011)	2.77	2.09	15	-6.30	2.09	15	No match	Potential SD/SE error with EF of 4.22. Did not report SD in the paper but 'DS'
Lindh & Holgersson (2007.1)	29.44	5.04	170	28.84	5.46	170	Use SD	Column name as 'standard deviation'
Lindh & Holgersson (2007.2)	11.16	2.96	184	11.53	2.45	184	Use SD	Column name as 'standard deviation'
Ortlz et al. (2017)	86.15	8.62	33	70.41	10.50	33	No match	Tables were not extracted. 6 tables in the form of pictures on page 4
Wu et al. (2015.1)	91.06	5.17	31	79.93	8.18	31	No match	No table extracted, tried with pdf plumber but no improvement
Wu et al. (2015.2)	4.20	0.76	31	3.30	0.77	31	No match	No table extracted, tried with pdf plumber but no improvement
Yang et al. (2023.1)	3.82	0.77	41	3.30	1.07	41	Use SD	Column name as 'Std Dev'
Yang et al. (2023.2)	81.57	6.70	41	79.09	7.38	41	Use SD	Column name as 'Std Dev'
Yang et al. (2023.3)	3.51	0.67	41	3.07	0.58	41	Use SD	Column name as 'Std Dev'
Yang et al. (2023.4)	41.43	9.20	41	36.89	8.49	41	Use SD	Column name as 'Std Dev'
Yang et al. (2023.5)	12.24	1.68	41	11.09	1.61	41	Use SD	Column name as 'Std Dev'

3.2.2. Detection Accuracy

Each primary study was manually checked to assess the validity of the tool. This step manually extracted the SD/SE values used from the PDF files for comparison with the results of the automated tool. For the effect sizes assessed in the previous section, correct detection was drawn when the result correctly matched the manual check. On the other hand, cases where this did not occur were considered incorrect identifications. This provided further accuracy analysis of this automated identification tool.

Of the 32 primary studies that successfully achieved detection in the two meta-analyses (Sandercock, 2024; Wang et al., 2023), 3 were determined to be misidentification, giving a final accuracy of 90.63% among success detection. These three cases were specifically analyzed as follows:

- The study of Andersen et al. (2010) was the only case identified as an SD/SE mix-up. However, manual checking revealed that the study did not contain any error; it accurately utilized SD that were converted from SE. The study reported several indicators with standard errors. Remarkably, the SE of body weight values for the experimental and control groups was equal to the SD values of VO_2max reported in the meta-analysis. These SE values were correctly extracted, resulting in a successful match but leading to a misleading conclusion.

- In the study of Uth et al. (2021), the SD value for an unrelated variable happens to have the same value as the SD in the meta-analysis. After manual checking, it was found that the meta-analysis reported VO_2max SD values of 3.0 for both the experimental and control groups, which were not explicitly documented in the study. However, in the PDF file, one SD value for an unrelated variable, the time since surgery (years), happened to be 3.0, and it was extracted successfully. So instead of the SD value for VO_2max, the SD values of the time since surgery successfully matched with the used value in the meta-analysis. It is written in the literature that "changes in VO_2max were demonstrated in another study (Uth et al., 2020)". However, upon further manual examination, no corresponding sample size, mean, and standard deviation (SD) were identified to support the effect size, and therefore we could not confirm whether there was SD/SE mix-up error.

- Similarly, the study by Randers et al. (2012) happened to have extracted matching SD values but for an unrelated variable. The used SD values in the meta-analysis were the same for both the experiment and control groups, however, the SD value of 6.3 was only found for Lean body mass (kg) instead of VO_2max in the PDF file. The conclusion from manual checking was that the reported SD values in the meta-analysis were not found in the PDF. But the automated tool extracted the SD value of Lean body mass (kg), and matched it with the reported SD values, leading to an incorrect conclusion.

3.2.3. Value Extraction Method Comparison

Section 2.5 above introduced two separate methods for SD/SE values extraction: the column name keywords detection, and the regular expression for numeric values, where the regular expressions include both plus-minus sign '±SD/SE' and round brackets '(SD/SE)'. Both methods were specifically analyzed to verify their effectiveness and shortcomings. For those 29 primary studies that were correctly detected, the effective methods that achieved the matching SD/SE values were counted and catalogued one by one.

Figure 9 presented the distribution of extraction methods that were relied upon in all correctly recognized cases. Of the 29 correct identifications, 15 were achieved through chart title keyword identification, while the remaining 14 were accomplished using specialized REPORT regular expression matching. In the context of regular expressions, the plus and minus signs were recognized 9 times, whereas the round brackets were identified 6 times. Values extracted using regular expressions demonstrated a high degree of precision. In contrast, data extracted based on column names were significantly influenced by the quality of the extracted table. The column name extraction frequently resulted in the inclusion of extraneous information.

Table 9: Validation results of the SD/SE value extraction methods for all corrected extracted cases

Value Detection Method	Column Name keywords Detection	Regular expression		Overall
		Plus-minus Sign	Round Brackets	
Number	15	8	6	29

Notably, the verification also highlighted that these two extraction approaches complement each other. One study reported SD value in the form of '(.86)', '(.54)', '(.83)' and '(.64)' (Hyun et al., 2008). Due to the missing digit, the data extracted through column names could not be matched. However, the regular expression accurately parsed and extracted the complete value, accomplishing the correct detection. Another case was the absence of an '(SD)'-like annotation on the page, despite all SD values being enclosed in parentheses (Uth et al., 2020). It caused the regular extraction to fail; however, the detection was still completed by recognizing the columns names.

3.2.4. Failure Analysis

This section elaborates on the 23 undetectable effect sizes. These undetectable cases were explicated through a combination of manual inspection and log reports of the tool. They were categorized for better illustration. In addition, several enlightening or unexpected cases were singled out for illustration.

- Table Extraction Failure: There were a total of 12 primary failed to be detected

that some or even all of the tables were not successfully extracted. Further manual confirmation revealed that the table extraction failures were sometimes attributable to the extraction tool and sometimes to the quality of the tables. The failure of the table extraction tool, Tabula, remains unclear. It was encountered in different situations. Regarding the primary study conducted by Krustrup et al. (2010), a total of 71 tables were extracted, with the majority of them being empty tables. Consequently, no SD/SE values could be obtained from all 71 tables. Another case was that Tabula failed to extract tables when there were multiple tables on one PDF page. This failure happened to both studies by Hyun et al. (2008) and Ortiz et al. (2016). Another case is the study by Randers et al. (Randers et al.) that all '±' symbols were missing in the extracted table and all SE values were not detected. As for the low quality of the tables, a typical example is in the study by Ortiz et al. (2016), all of the charts in the article were attached as pictures. These non-text tables were unable to parse with the current table extraction package, which led to detection failure.

- Data Extraction Failure: 5 detection failures occurred due to researchers using different reporting stiles for SD/SE values. The data were presented in the table and the extraction process was completed successfully. However, the data could not be parsed and extracted due to varying reporting habits or irregular reporting formats. In the primary study by Korkmaz (2016), the values in the tables were all presented with commas rather than decimal points, for example '16,9'. This causes all value extraction methods to be invalidated. Another example was the study by Milanović et al. (2015) and all SD values were reported in the form of ±SD. In the article, no notes or descriptions specific to ± were found, except in the abstract section. This results in all retrieved values not being tagged.

- Data matching failed: In 6 cases, the adopted SD/SE values could not be found in the table of primary studies despite manual inspection. The study by Casad and Jawaharlal (2012) did not report SD values in any of the tables, but they were mentioned in the body of the text. Besides, some SD values require additional calculations to obtain, which are mainly related to multiple measurements over different periods. To be specific, some of the used SD values were the overall SD converted by the pre-test and pro-test SD values for the same group. The calculation of the overall standard deviation (SD) for one group with pre-test and pro-test measurements is shown in Equation 3.1:

$$SD_{\text{overall}} = \sqrt{\frac{SD_{\text{pre}}^2 + SD_{\text{pro}}^2}{2}} \tag{3.1}$$

This formula computes the square root of the average of the squared SDs from each group, combining their variability into a single measure. In some primary studies (Hsiao et al., 2015; Julià and Antolí, 2016), this formula was used in the

experimental and control groups to calculate two overall SD values separately. These overall SD values were used for the calculation of ES in the meta-analysis.

- One special case is the study by La Paglia et al. (2011) that it could involve a potential SD/SE mix-up error. First, in the PDF file, no SD was found by manual inspection, but there were values in the column named DS in the table. After verification, it was found that the SD values used in the meta-analysis were overall SD values calculated with these values named DS using Equation 3.1. Second, the calculated ES for this study is 4.22 and the CIs is from 2.87 to 5.57. The excessive ES and wide CIs raise suspicions about their SD/SE errors. Due to the lack of explanation of the DS value in the PDF literature, ANOVA analyses based on the study were reproduced using data from the literature (MEAN and N) and the DS values as either SD and SE values, respectively. However, the results of the ANOVA, whether DS was interpreted as SD or SE, were not consistent with the results reported in this experiment. Hense, the meaning of the DS value remains unclear, but its use in meta-analyses is incorrect because it was rejected as an SD value based on the reanalysis results.

4. DISCUSSION

This study explores the feasibility of automatically identifying the misuse of SE instead of SD to calculate ES in meta-analysis. Following a thorough expert assessment, SD/SE mix-up error was selected among five summarized prevalent statistical errors for further automated detection exploration. First, the research question focuses on the feasibility of the extraction and comparison of cross-document data in meta-analysis. Second, the development of an automated tool enables the extraction of SD/SE values that may be present in the PDF file and the SD/SE error is detected by comparing the extracted values with meta-analysis.

With the validation of two published meta-analyses (Sandercock, 2024; Wang et al., 2023), the results indicate the feasibility of automated identification of SD/SE mix-up errors. Besides, the analysis of failure cases summarizes the current obstacles to achieving fully automatic recognition and reveals some of the possible causes of SD/SE mix-up errors.

4.1. Citation Tracing

The validation results of the primary study tracing reflected a high degree of accuracy and showed potential for boarder applications in automated batch citation tracing. For 55 primary studies, the tool has a 96.36% correct rate for citation tracing, where one of the two incorrect tracings was due to incorrect information from the meta-analysis researcher. Rapid batch matching of citations based on PDF files not only provided a source of literature for subsequent SD/SE extraction but could be helpful in the field of meta-analysis and even in the field of regular scientific publications. PDF as a common format for academic articles, tracking of references in them was more cumbersome than in the publisher's online platforms. Extracting citations directly from it, especially a large number of citations through keywords like meta-analysis, was not inherently difficult but time-consuming. The high accuracy of the current citation tracking feature can be used as a stand-alone tool. This is not limited to primary study extraction for the meta-analysis, but can be profitable for all tasks that involve extracting information across academic articles.

This relied on the effectiveness of GROBID in extracting reference lists for PDF files. The well-structured results obtained from parsing greatly improved the mechanical readability of citations and provided great convenience for automated processing and citation analysis. However, as illustrated in section 3.1.2, an error of GROBID occurred where one citation was mistakenly split into two. And it directly affected the numbering of all subsequent references. It did not affect the tracing results in this particular meta-analysis, which used author and publication year for referring. However, if the

citation number was used as the information to refer the reference, the incorrect extraction of a single reference could potentially invalidate the tracing of all subsequent references. For example, the meta-analysis by Sandercock (2024) used only citation numbers and author names to indicate the included study, e.g., Andersen [32]. If the extracted citation number is incorrect, fuzzy matching is likely to fail or match an incorrect citation.

While the validation results showed a high level of accuracy (96.36%), these two existing incorrect cases for citation tracing remain a concern. Madhavan et al. (2006) suggested that the automation errors in tasks perceived as simple have a greater negative impact on trust than errors in tasks perceived as difficult. Operators often hold the assumption that machines and humans process information in similar ways. And when automated systems make "basic" errors, operators question their ability to perform other tasks. There are two possible improvements to prevent automation disuse that may result from these errors. On the one hand, potential errors can be eliminated by combining precision matching with existing fuzzy matching for a more accurate matching system. On the other hand, the presentation of matching citation results can be further designed. This would allow the operator to quickly check the correctness of the match. For example, the system can highlight the matched author and year information, or present citations with small differences in matching similarity scores. Overall, it is believed that the current citation tracing function can be adapted to serve as a stand-alone tool in the academic field.

4.2. Technical Difficulty of Table Extraction

In the processing of meta-analyses and their primary studies, extraction data from Portable Document Format (PDF) was found to be the greater impediment to automation. Based on the results of failure cases, about half of the failed-matching primary studies were due to the failure of table extraction from PDFs. Berners-Lee (2009) analyzed open data quality, and in the 5-Star Open Data Model he defined, PDF is classified at the lowest level of openness. This is because data in PDF lacks of structure to meet level two in the model that PDF files PDF files emphasize graphical visualization, while for all structures the corresponding information is missing, resulting in tables and texts that cannot be extracted correctly. For example, borders are often used to divide different groups in a table, which is visualized in the PDF but cannot be extracted. In PDF files, machine processing is costly and challenging due to the lack of sufficient information on the data structure (Silva, 2010). Despite using the optimal tool, Tabula, following the assessment, the final results were still considered can be improved with an overall detection rate of 58.18%.

With the development of machine learning, there are now more possibilities for processing PDF text. During the development process, only the traditional and popular PDF table extraction tools were adapted and compared. However, there are existing tools that rely on machine learning technology for table extraction. For example, GROBID, previously used only for extracting reference lists, also features the ability to extract

structured data from full-text documents. Enhanced accuracy and structure in data extraction enable more efficient and effective automated processing. Other companies, like Nanonets(**anonets**) and Parsio (2024), also provide AI tools that specialize in PDF form extraction. However, besides GROBID, these tools mentioned are non-open source and subscription-based, limiting the deployment of automated tool development. Overall, mechanical learning provides new possibilities for table extraction but further exploration is needed to determine the accessibility and performance of the AI tools compared with traditional table extraction tools.

An interesting finding of the validation process is that the promotion of documents in Extensible Markup Language (XML) format, in addition to PDF, can greatly improve the mechanical processing of academic articles. Unlike PDF files, which focus on graphical visualization, XML files XML files are highly structured and standardized. It is both human-readable and machine-readable and contains only text. The specific structured division of text, images, and tables makes it easier to extract and analyze data accurately and efficiently. Some academic literature platforms such as PubMed Central (2024) and Elsevier (2024) are now promoting full-text XML. As more scholarly articles become available in XML, in the future, XML can be utilized for more accurate and efficient text and data processing.

4.3. Uniformity of Academic Reporting

In addition to technical difficulties during the development of the automated tool, another significant challenge was the inconsistency in the reporting practices of meta-analyses and primary studies. These inconsistencies manifested in two ways: variability in reporting norms and some low-quality reporting.

For meta-analyses, reported data are generally insufficient and rarely include original data from primary studies. During the validation phase, few meta-analysis reports were found that provided sample sizes, standard deviations, and means for the raw data. The commonality of non-reporting of SD not only makes automated identification of SD/SE mix-up errors impractical but also poses significant difficulties for manual inspection. In addition, it was observed that some meta-analyses did not even report effect sizes for individual original experiments. These deficiencies in data reporting are not related to the number of citations. The limited availability of data restricts the extent to which the SD/SE error automated detection tools can be effectively applied. Consequently, the lack of detailed reporting in meta-analyses hampers the assessment of their quality and limits the scope of re-analyses that can be automated. Some researchers may feel unwilling to share their data publicly due to the substantial time and effort invested in data collection. However, as the researchers initiated, reporting all the information extracted from the original experiments is extremely worthwhile for future reviews (Aguinis et al., 2021; Johnson and Hennessy, 2019; Polanin et al., 2020). The article of Lakens et al. (2016) suggested that meta-analytic data should be more broadly defined and in addition to effect

sizes and their confidence intervals, data such as sample sizes, means, mean differences, etc. Comprehensive data sharing can enhance the transparency and reproducibility of meta-analyses while also expanding the potential for automated error identification.

For the articles of primary studies, there is no standardized norm reporting data. On the one hand, journals and scientific disciplines have specific normative preferences for publishing literature. On the other hand, depending on the specific design of the study and the amount of data to be reported, researchers have a high degree of latitude in reporting data. For example, by mentioning how to refer to the standard deviation in a table, more than five different methods have been observed from the validation. And the code for the tool kept updating according to new cases. Throughout the development of the tool, it was also a process to find and adopt different reporting styles. A significant task of the tool's development involves matching and reconciling diverse reporting styles. The variation in reporting styles poses effects not only in producing and evaluating meta-analyses, but also in automation. During the development, it was a time-consuming process that each new case was found and analyzed manually before consideration of the potential solution. In further validation and specific uses of the current automated detection tool, it is believed that the reporting inconsistencies will continue requiring additional adaption.

In addition to the inconsistency of reporting, the problem is the low quality of reporting data. It includes but is not limited to, the absence of necessary descriptions on tables, messy table formatting, tables based on the form of images, and the use of commas for decimal points in numeric values. It is believed that additional issues will continue to emerge as the number of articles reviewed increases. The low-quality data reporting in these PDFs not only makes the automated tool ineffective but also creates obstacles for manual data extraction during the check process. This is a possible contributing factor to errors in meta-analysis given it already affects the quality of the paper. Low-quality data reporting requires a concerted effort by researchers and publishers. Researchers should be educated more about the norms of academic reporting, while publishers should implement more rigorous review processes to ensure data completeness and correctness. It will not only enhance the possibilities for automated text and data processing, but also improve the quality of reports and indirectly the reliability of future meta-analyses.

4.4. Limitations and Future Researches

In addition to the findings and recommendations mentioned above, it is important to emphasize some of the limitations in the design of the tool itself. As described in Section 2.1, the error of SD/SE mix-up error cannot be accurately judged in terms of the size of the meta-analytic effect alone. To achieve automated detection, this study proposes to extract SD and SE values from the original references to match the SD values reported in the meta-analysis. However, the data accessibility of meta-analysis reports limited the automatic tool, allowing it could only be applied to a proportion of published meta-analyses or self-checks. Besides, given that it is unable to recognize each

SD/SE value corresponding to which measurement, the current identification tool can only compare the used SD/SE values with all extracted values. As mentioned in the validation results, when there are multiple variables in a study, some values for undesired variables may coincidentally match, leading to erroneous identification results. For example, the study by Andersen et al. (2010) had a SE value for the body weights that happened to be the same as the used SD value for VO2max, leading to an incorrect detection.

In future research, the addition of a user interface and workflow could be considered to realize the connection of the different functions. The current tool is still at the stage of sub-functional code and requires further validation on a larger scale to confirm its effectiveness. After refining the code based on the test results, the entire tool can be packaged and an appropriate visual interactive interface can be designed. Each sub-task's completion and the corresponding user feedback needed at each step must be considered individually to ensure the final tool's effectiveness. More specifically, since each sub-step provides different results, what results to show, how to show them, and what the user needs to do need to be investigated further for better automation. It is suggested that the tool can be realized in a semi-automatic way. Given the current validation results and difficulties, a semi-automatic approach can be ideal. In this way, the tool can provide results for each stage and the users can get access to the extracted information and data. With human verification, it may achieve better accuracy.

5. CONCLUSION

This study develops an automatic tool to detect the SD/SE mix-up error in meta-analyses. The validation shows the feasibility of extracting data cross-literature and detecting SD/SE errors in meta-analyses. Among them, thanks to the mechanism of GROBID and fuzzy matching, the extraction of reference lists and the tracing of the primary study show high accuracy. This function has the potential to become a separate tool for a boarder application in the scientific literature.

The failure cases for the automatic tool reveal the challenges of the SD/SE error detection which are two aspects. The first one is the technical difficulty of table extraction from PDF files. The utilized package Tabula is considered limited in that it sometimes fails to extract charts or incorrectly extracts tables. It is suggested to explore and evaluate other available table attraction tools to achieve better performance. The other suggestion is to use alternative methods for data extraction beyond PDF extraction, for example, TEI/XML files and shared databases. The second challenge is from the data reporting in academic articles. It is concluded that there are lack of utilized data reporting in the articles. When those data are reported, due to the publishers' preferences and researchers' report habits, these data can be described in a variety of ways. Besides, some of the tables are evaluated to be of poor quality. All of these bring challenges to automatic detection. It is recommended that researchers need to be educated to raise academic awareness. And the publisher should increase the inspection efforts to ensure the quality of the articles.

Future work is mentioned to focus on improving the accuracy of the tool, extending its functionality and integrating all sub-functions into a complete set of workflow. Given the current challenges, it is suggested to have a semi-automated tool for better achieving SD/SE mix-up error detection in meta-analysis.

REFERENCES

Aguinis, H., Hill, N. S., & Bailey, J. R. (2021). Best practices in data collection and preparation: Recommendations for reviewers, editors, and authors. *Organizational Research Methods*, *24*(4), 678–693.

Akbik, A., Bergmann, T., Blythe, D., Rasul, K., Schweter, S., & Vollgraf, R. (2019). Flair: An easy-to-use framework for state-of-the-art nlp. *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (demonstrations)*, 54–59.

Andersen, L. J., Randers, M., Westh, K., Martone, D., Hansen, P. R., Junge, A., Dvorak, J., Bangsbo, J., & Krustrup, P. (2010). Football as a treatment for hypertension in untrained 30–55-year-old men: A prospective randomized study. *Scandinavian journal of medicine & science in sports*, *20*, 98–102.

Andrade, C. (2020). Understanding the difference between standard deviation and standard error of the mean, and knowing when to use which. *Indian Journal of Psychological Medicine*, *42*(4), 409–410.

Aytug, Z. G., Rothstein, H. R., Zhou, W., & Kern, M. C. (2012). Revealed or concealed? transparency of procedures, decisions, and judgment calls in meta-analyses. *Organizational Research Methods*, *15*(1), 103–133.

Berners-Lee, T. (2009). Linked data. https://www.w3.org/DesignIssues/LinkedData.html

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2021). *Introduction to meta-analysis*. John Wiley & Sons.

Casad, B. J., & Jawaharlal, M. (2012). Learning through guided discovery: An engaging approach to k-12 stem education. *American Society for Engineering Education. pp. 00078-15. 2012*, 00078–00015.

Cheng, L., Katz-Rogozhnikov, D. A., Varshney, K. R., & Baldini, I. (2021). Automated meta-analysis: A causal learning perspective. *arXiv preprint arXiv:2104.04633*.

Cooper, H. (2011). *Reporting research in psychology: How to meet journal article reporting standards.* American Psychological Association.

Cooper, H., Hedges, L. V., & Valentine, J. C. (2019). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Elsevier. (2024). Xml in science publishing: Policies and guidelines [Accessed: 2024-07-20]. https://www.elsevier.com/researcher/author/policies-and-guidelines/xml-in-science-publishing

Fuhr, U., & Hellmich, M. (2015). Channeling the flood of meta-analyses.

Garg, A. X., Hackam, D., & Tonelli, M. (2008). Systematic review and meta-analysis: When one study is just not enough. *Clinical journal of the American Society of Nephrology*, *3*(1), 253–260.

Grobid. (2008–2023).

Guo, L.-q., Chen, Y., Mi, B.-b., Dang, S.-n., Zhao, D.-d., Liu, R., Wang, H.-l., & Yan, H. (2019). Retracted article: Ambient air pollution and adverse birth outcomes: A systematic review and meta-analysis. *Journal of Zhejiang University. Science. B*, *20*(3), 238.

Harrer, M. (2023). Suicideprevention dmetar. https://dmetar.protectlab.org/reference/suicideprevention

Hong, J.-C., Yu, K.-C., & Chen, M.-Y. (2011). Collaborative learning in technological project design. *International Journal of Technology and Design Education*, *21*, 335–347.

Hong, Z.-W., Huang, Y.-M., Hsu, M., & Shen, W.-W. (2016). Authoring robot-assisted instructional materials for improving learning performance and motivation in efl classrooms. *Journal of Educational Technology & Society*, *19*(1), 337–349.

Hsiao, H.-S., Chang, C.-S., Lin, C.-Y., & Hsu, H.-L. (2015). "irobiq": The influence of bidirectional interaction on kindergarteners' reading motivation, literacy, and behavior. *Interactive Learning Environments*, *23*(3), 269–292.

Hyun, E.-j., Kim, S.-y., Jang, S., & Park, S. (2008). Comparative study of effects of language instruction program using intelligence robot and multimedia on linguistic ability of young children. *RO-MAN 2008-the 17th IEEE international symposium on robot and human interactive communication*, 187–192.

Ioannidis, J. P. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, *94*(3), 485–514.

Johnson, B. T., & Hennessy, E. A. (2019). Systematic reviews and meta-analyses in the health sciences: Best practice methods for research syntheses. *Social Science & Medicine*, *233*, 237–251.

Jones, A. P., Remmington, T., Williamson, P. R., Ashby, D., & Smyth, R. L. (2005). High prevalence but low impact of data extraction and reporting errors were found in cochrane systematic reviews. *Journal of clinical epidemiology*, *58*(7), 741–742.

Julià, C., & Antolí, J. Ò. (2016). Spatial ability learning through educational robotics. *International Journal of Technology and Design Education*, *26*, 185–203.

Kadlec, D., Sainani, K. L., & Nimphius, S. (2023). With great power comes great responsibility: Common errors in meta-analyses and meta-regressions in strength & conditioning research. *Sports Medicine*, *53*(2), 313–325.

Ko, W.-R., Hung, W.-T., Chang, H.-C., & Lin, L.-Y. (2014). Inappropriate use of standard error of the mean when reporting variability of study samples: A critical evaluation of four selected journals of obstetrics and gynecology. *Taiwanese Journal of Obstetrics and Gynecology*, *53*(1), 26–29.

Koricheva, J., & Gurevitch, J. (2014). Uses and misuses of meta-analysis in plant ecology. *Journal of Ecology*, *102*(4), 828–844.

Koricheva, J., Gurevitch, J., & Mengersen, K. (2013). *Handbook of meta-analysis in ecology and evolution*. Princeton University Press.

Korkmaz, Ö. (2016). The effect of lego mindstorms ev3 based design activities on students' attitudes towards learning computer programming, self-efficacy beliefs and levels of academic achievement. *Online Submission*, *4*(4), 994–1007.

Kovaka, K. (2022). Meta-analysis and conservation science. *Philosophy of Science*, *89*(5), 980–990.

Krustrup, P., Hansen, P. R., Randers, M., Nybo, L., Martone, D., Andersen, L. J., Bune, L. T., Junge, A., & Bangsbo, J. (2010). Beneficial effects of recreational football on the cardiovascular risk profile in untrained premenopausal women. *Scandinavian journal of medicine & science in sports*, *20*, 40–49.

La Paglia, F., Rizzo, R., & La Barbera, D. (2011). Use of robotics kits for the enhancement of metacognitive skills of mathematics: A possible approach. *Annual Review of Cybertherapy and Telemedicine 2011*, 26–30.

Lakens, D., Page-Gould, E., van Assen, M. A., Spellman, B., Schönbrodt, F. D., Hasselman, F., Corker, K. S., Grange, J. A., Sharples, A., Cavender, C., et al. (2017). Examining the reproducibility of meta-analyses in psychology: A preliminary report.

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC psychology*, *4*, 1–10.

Lee, Y. H. (2018). An overview of meta-analysis for clinicians. *The Korean Journal of Internal Medicine*, *33*(2), 277.

Lopez, P. (2009). Grobid: Combining automatic bibliographic data recognition and term extraction for scholarship publications. *Research and Advanced Technology for Digital Libraries: 13th European Conference, ECDL 2009, Corfu, Greece, September 27-October 2, 2009. Proceedings 13*, 473–474.

Madhavan, P., Wiegmann, D. A., & Lacson, F. C. (2006). Automation failures on tasks easily performed by operators undermine trust in automated aids. *Human factors*, *48*(2), 241–256.

Mak, A., Cheung, M. W., Fu, E. H., & Ho, R. C. (2010). Meta-analysis in medicine: An introduction. *International Journal of Rheumatic Diseases*, *13*(2), 101–104.

Mariscal-Harana, J., Asher, C., Vergani, V., Rizvi, M., Keehn, L., Kim, R. J., Judd, R. M., Petersen, S. E., Razavi, R., King, A. P., et al. (2023). An artificial intelligence tool for automated analysis of large-scale unstructured clinical cine cardiac magnetic resonance databases. *European Heart Journal-Digital Health*, *4*(5), 370–383.

McKie, M., & Liu, R. (2024). Pymupdf - python binding for mupdf. https://pypi.org/project/PyMuPDF/

Metaxa, A.-M., & Clarke, M. (2024). Efficacy of psilocybin for treating symptoms of depression: Systematic review and meta-analysis. *bmj*, *385*.

Milanović, Z., Pantelić, S., Sporiš, G., Mohr, M., & Krustrup, P. (2015). Health-related physical fitness in healthy untrained men: Effects on vo2max, jump performance and flexibility of soccer and moderate-intensity continuous running. *PloS one*, *10*(8), e0135319.

Moher, D., Liberati, A., Tetzlaff, J., Altman, D. G., Group, P., et al. (2010). Preferred reporting items for systematic reviews and meta-analyses: The prisma statement. *International journal of surgery*, *8*(5), 336–341.

Nagele, P. (2003). Misuse of standard error of the mean (sem) when reporting variability of a sample. a critical evaluation of four anaesthesia journals. *British Journal of Anaesthesia*, *90*(4), 514–516.

Nakagawa, S., & Santos, E. S. (2012). Methodological issues and advances in biological meta-analysis. *Evolutionary Ecology*, *26*, 1253–1274.

Nanonets. (2024). Extract tables from pdf [Accessed: 2024-07-20]. https://nanonets.com/free-tools/extract-table-from-pdf

National Center for Biotechnology Information. (2024). Pubmed [Accessed: 2024-07-19]. https://pubmed.ncbi.nlm.nih.gov/

Nuijten, M. B., & Polanin, J. R. (2020). "statcheck": Automatically detect statistical reporting inconsistencies to increase reproducibility of meta-analyses. *Research synthesis methods*, *11*(5), 574–579.

Nyberg, M., Blackwell, J. R., Damsgaard, R., Jones, A. M., Hellsten, Y., & Mortensen, S. P. (2012). Lifelong physical activity prevents an age-related reduction in arterial and skeletal muscle nitric oxide bioavailability in humans. *The Journal of physiology*, *590*(21), 5361–5370.

Ortiz, O. O., Franco, J. Á. P., Garau, P. M. A., & Martín, R. H. (2016). Innovative mobile robot method: Improving the learning of programming languages in engineering degrees. *ieee Transactions on Education*, *60*(2), 143–148.

Parsio. (2024). Extract data from emails and documents [Accessed: 2024-07-20]. https://parsio.io/

Polanin, J. R., Hennessy, E. A., & Tanner-Smith, E. E. (2017). A review of meta-analysis packages in r. *Journal of Educational and Behavioral Statistics*, *42*(2), 206–242.

Polanin, J. R., Hennessy, E. A., & Tsuji, S. (2020). Transparency and reproducibility of meta-analyses in psychology: A meta-review. *Perspectives on Psychological Science*, *15*(4), 1026–1041.

PubMed Central. (2024). Tagging guidelines: Article style [Accessed: 2024-07-20]. https://www.ncbi.nlm.nih.gov/pmc/pmcdoc/tagging-guidelines/article/style.html

Randers, M. B., Petersen, J., Andersen, L. J., Krustrup, B. R., Hornstrup, T., Nielsen, J. J., Nordentoft, M., & Krustrup, P. (2012). Short-term street soccer improves fitness and cardiovascular health status of homeless men. *European journal of applied physiology*, *112*, 2097–2106.

Randers, M. B., Nielsen, J. J., Krustrup, B. R., Sundstrup, E., Jakobsen, M. D., Nybo, L., Dvorak, J., Bangsbo, J., & Krustrup, P. (2010). Positive performance and health effects of a football training program over 12 weeks can be maintained over a 1-year period with reduced training frequency. *Scandinavian journal of medicine & science in sports*, *20*, 80–89.

Roba, A. A., Tefera, M., Worku, T., Dasa, T. T., Estifanos, A. S., & Assefa, N. (2019). Retracted article: Application of 4% chlorhexidine to the umbilical cord

stump of newborn infants in lower income countries: A systematic review and meta-analysis. *Maternal Health, Neonatology and Perinatology*, *5*, 1–9.

Sandercock, G. (2024). The standard error/standard deviation mix-up: Potential impacts on meta-analyses in sports medicine. *Sports Medicine*, 1–10.

Schmidt, J. F., Hansen, P. R., Andersen, T. R., Andersen, L., Hornstrup, T., Krustrup, P., & Bangsbo, J. (2014). Cardiovascular adaptations to 4 and 12 months of football or strength training in 65-to 75-year-old untrained men. *Scandinavian journal of medicine & science in sports*, *24*, 86–97.

Schwarzer, G., et al. (2007). Meta: An r package for meta-analysis. *R news*, *7*(3), 40–45.

Silva, A. (2010). Parts that add up to a whole: A framework for the analysis of tables. *Edinburgh University, UK*.

Suurmond, R., van Rhee, H., & Hak, T. (2017). Introduction, comparison, and validation of meta-essentials: A free and simple tool for meta-analysis. *Research synthesis methods*, *8*(4), 537–553.

Uth, J., Fristrup, B., Sørensen, V., Helge, E. W., Christensen, M. K., Kjærgaard, J. B., Møller, T. K., Helge, J. W., Jørgensen, N. R., Rørth, M., et al. (2021). One year of football fitness improves l1–l4 bmd, postural balance, and muscle strength in women treated for breast cancer. *Scandinavian journal of medicine & science in sports*, *31*(7), 1545–1557.

Uth, J., Fristrup, B., Sørensen, V., Helge, E. W., Christensen, M. K., Kjærgaard, J. B., Møller, T. K., Mohr, M., Helge, J. W., Jørgensen, N. R., et al. (2020). Exercise intensity and cardiovascular health outcomes after 12 months of football fitness training in women treated for stage i-iii breast cancer: Results from the football fitness after breast cancer (abc) randomized controlled trial. *Progress in Cardiovascular Diseases*, *63*(6), 792–799.

Wang, K., Sang, G.-Y., Huang, L.-Z., Li, S.-H., & Guo, J.-W. (2023). The effectiveness of educational robots in improving learning outcomes: A meta-analysis. *Sustainability*, *15*(5), 4637.

Wullschleger, M., Aghlmandi, S., Egger, M., & Zwahlen, M. (2014). High incorrect use of the standard error of the mean (sem) in original articles in three cardiovascular journals evaluated for 2012. *PLoS One*, *9*(10), e110364.