ABSTRACT

LONDON, JENNIFER E. The Effect of Time Period, Field, and Coding Context on Rigor, Interrater Agreement, and Interrater Reliability in Meta-Analysis. (Under the direction of Dr. Mark A. Wilson).

Meta-analyses are used in all fields of research, and they are typically assumed to be the same regardless of field or time. This paper explores changes in the meta-analytic process over time as well as across the fields of Computer Science, Education, Medicine, and Psychology on a representative sample of meta-analyses. Facets of the process of meta-analysis were also examined as well as how they impact meta-analytic Rigor, interrater agreement (IRA) and interrater reliability (IRR). Data visualization, ANOVAs, multiple regression, and path analysis were used to connect variables and make sense of relationships. Surprisingly, IRA and Rigor were not related, and IRR had a strong negative correlation with Rigor. Furthermore, the reporting rates of IRA and IRR were very low across all fields.

In terms of procedure, meta-analyses in the field of Medicine were the most well-developed and offer several ideas for improving meta-analyses in the social sciences. Suggestions for improving IRA and IRR process and reporting are included. In examining many meta-analyses from different fields, some surprising discoveries were made. Most notably, studies labeled as meta-analyses in Computer Science frequently failed to include the analyses that typify meta-analysis; instead, they would be better described as vote-counting studies or qualitative reviews.
© Copyright 2016 by Jennifer Elise London
All Rights Reserved
The Effect of Time Period, Field, and Coding Context on Rigor, Interrater Agreement, and Interrater Reliability in Meta-Analysis

by
Jennifer Elise London

A dissertation submitted to the Graduate Faculty of North Carolina State University in partial fulfillment of the requirements for the degree of Doctor of Philosophy

Psychology

Raleigh, North Carolina

2016

APPROVED BY:

Dr. Mark A. Wilson
Committee Chair

Dr. Samuel B. Pond, III

Dr. Adam W. Meade

Dr. Eric N. Wiebe
BIOGRAPHY

Jennifer Elise London was born in 1983 in Huntington Beach, CA. She attended Sonoma State University before transferring to North Carolina State University to finish her Bachelor of Arts degree in Psychology. She was accepted into the Industrial and Organizational Psychology Doctoral program at North Carolina State University in 2007 and received her Masters of Science degree in 2011. Jennifer has had the good fortune to work in several applied settings, including the Friday Institute for Educational Innovation, the City of San Francisco, and currently Stanford University School of Medicine. She enjoys working with data and statistics to inform organizational decisions.
ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Mark Wilson, for his invaluable contributions to my education. I would also like to thank my committee for their feedback and contributions to making this study more rigorous.

Finally, I would like to thank Dr. Mary Luong for being the second coder, and for not divorcing me in the process.
# TABLE OF CONTENTS

LIST OF TABLES .......................................................................................................................... vii
LIST OF FIGURES ........................................................................................................................ viii

Introducti................................................................. 2

Conceptual Model and Variables ................................................................. 5

Study Elements ......................................................................................... 6

Time period ......................................................................................... 6
Field of study .......................................................................................... 7
Number of coders (NCoders) .................................................................. 9
Total number of variables (TotalVs) ...................................................... 10
Variable type .......................................................................................... 10
  Low inference ...................................................................................... 11
  High inference ..................................................................................... 11
  Meta variables ..................................................................................... 12
Rigor ........................................................................................................ 12
IRA and IRR ........................................................................................... 14

Method ...................................................................................................... 17
Sampling Method ................................................................................... 17
Definition of Fields ................................................................................ 17
  Computer Science .............................................................................. 17
  Education ........................................................................................... 18
  Medicine .............................................................................................. 18
  Psychology .......................................................................................... 18
Field Overlap ........................................................................................... 18
Inclusion Criteria .................................................................................... 19
Coding of Dependent Variables ............................................................ 19
  Coders and IRA ................................................................................ 19
  Consolidating IRA and IRR values. ................................................... 20
Rigor ........................................................................................................ 20
Coding of Independent Variables ........................................................ 20
Time period.................................................................................................................. 20
Field .................................................................................................................................. 20
Number of coders. ............................................................................................................. 20
Total number of variables............................................................................................... 21
Percent high inference variables.................................................................................... 21
Percent low inference variables...................................................................................... 21
Number of meta variables............................................................................................... 22
Journal impact factor. ..................................................................................................... 22
Pages .................................................................................................................................. 22
Number of authors. .......................................................................................................... 22
Number of failsafes. ......................................................................................................... 23
Results.................................................................................................................................. 23
How Many Meta-Analyses are Available Each Year? (RQ1) ................................................. 23
Are Meta-Analytic Rigor, IRA, and IRR Interrelated? (RQ2).................................................. 25
Accounting for Time Period, How Do Meta-Analytic Rigor, IRA, and IRR Each Vary by Field of Study? (RQ3) .................................................................................................................. 30
Accounting for Time Period, How Does Field of Study Affect Number of Coders, Total Number of Variables, and Variable Type? (RQ4).......................................................... 30
Individually, Which of the Proximal Variables Affect IRA and/or IRR? (RQ5) ...................... 34
Additional Findings............................................................................................................ 34
False meta-analyses. ....................................................................................................... 34
Meta variables in Medicine. ............................................................................................. 35
Use of failsafes. .................................................................................................................. 36
Proportion of coders to authors. ...................................................................................... 37
Note. Background shade indicates number of authors, line indicates number of ............ 38
coders.................................................................................................................................. 38
Figure 5. Comparison of Number of Authors to Coders...................................................... 38
Visualization of IRA, rigor, and time.............................................................................. 38
Inclusion flow charts. ...................................................................................................... 39
Journal impact factors..................................................................................................... 39
Discussion......................................................................................................................... 39
LIST OF TABLES

Table 1. *Frequencies of IRA and IRR Values Available in Sample* .................................................. 25
Table 2. *Correlation Matrix of Variables with Variable N* ................................................................. 28
Table 3. *Variable Means and Standard Deviations by Field* ............................................................... 29
Table 4. *Estimates from full unconstrained model* .............................................................................. 33
Table 5. *Comparison of model fits with varying path constraints* ...................................................... 33
Table 6. *Multiple Regression Coefficients Predicting IRA* ................................................................. 34
Table 7. *Descriptives for Failsafes by Field* ......................................................................................... 36
LIST OF FIGURES

Figure 1. Conceptual Model ................................................................. 5
Figure 2. Meta-analysis Search Result Volume by Year .......................... 24
Figure 3. Path Model ........................................................................ 31
Figure 4. False Meta-Analyses Frequency by Time Period ...................... 35
Figure 5. Comparison of Number of Authors to Coders .......................... 38
Figure 6. Kappa, Percent Agreement, and ICC by Rigor and Time .......... 41
Figure 7. Example Inclusion Flow Chart .............................................. 42
Figure 8. Scatterplot of Rigor and JCR ............................................... 43
The Effect of Time Period, Field, and Coding Context on Rigor, Interrater Agreement, and Interrater Reliability in Meta-Analysis

Meta-analyses are an important part of modern research but they are susceptible to bias and methodological weakness. They are also exceptionally difficult to evaluate as a reader. This difficulty is due to several factors. First, readers must rely on the judgement of the meta-analysts. Meta-analytic results are largely dependent on the choices made by the meta-analysts. There is a plethora of difficulties in making such choices (Hunter & Schmidt, 2004; Wanous, Sullivan & Malinak, 1989). For example, meta-analytic coding is an intensive process that requires deep knowledge of the subject matter, repeated consensus meetings, and hundreds or thousands of hours of work. Second, readers rarely find the meta-analysts’ choices and processes documented in the published articles. Thus, it is usually impossible to directly evaluate bias and coding quality in a meta-analysis for anyone other than the authors despite the use of proxies for coding quality such as interrater agreement (IRA) and interrater reliability (IRR), which are quantifications of the degree of coding consistency between coders. Third, most readers are ignorant of how meta-analysis methodologies have evolved over time and in different fields of research. A fourth challenge in evaluating meta-analyses is that readers discuss results as though they were monolithic and equivalent regardless of when they were published, but they have likely evolved over time and might have a life expectancy. Similarly, readers often cite meta-analyses indiscriminately from different fields, but they do not know whether some fields hold a less rigorous standard. They do not know how IRA, IRR, rigor, time, field, and other features of meta-analysis relate to one another, nor do they have a high-level perspective on how those variables have evolved. Given these
difficulties in evaluating meta-analyses this paper has four objectives: to examine whether meta-analytic rigor and elements of the coding process predict IRA and IRR, to describe the level of meta-analytic rigor across different fields of study and time, to thoroughly describe the current state of meta-analysis and how it has evolved over time and field of study, and to make recommendations as to how we might use this information to better understand meta-analyses.

**Introduction**

It is not always apparent to readers of meta-analyses that data collection for meta-analysis is extremely difficult, it is subjective, and data are wholly dependent on the handful of people doing the task. Data collection is highly labor intensive; it requires patience, cooperation, and an intolerance for ambiguity. Meta-analyses can take years to complete, and the remarkable length of time can be attributed to the data collection and coding process (Cochrane Oversight Committee, 2012). Much of this time is spent revisiting agreed-upon definitions and debating how data should be classified, or if it is closely enough aligned to the construct of interest to be included at all. Unlike qualitative research, the coder must generate data from sources that were not originally intended to become data for meta-analysis. Each study is of varying degrees of relevance, and every data point within each study is framed differently depending on the original focus. The cognitive demand on coders is significant, and it is likely that fatigue and difficulty have an effect on error rate and data validity.

Every piece of data collected for a meta-analysis is the subject of a judgment on the part of the coder. Coding requires a doctoral level of understanding (Oremus, Oremus, Hall,
& McKinnon, 2012) of the topic and related topics in order to distinguish complex ideas from one another and to make judgments on propriety of categorization. It is not unusual for experts to come to different but valid conclusions. Coding decisions are dependent on the coder’s experience, and the data do and should reflect that to some extent. The subjective nature of data collection is not inherently problematic because we often trust the judgment of experts in research but it requires us to question whether the data collected is valid.

Despite the absolute criticality of coders to the validity of results, the identity of the people doing the coding is almost always absent from published meta-analyses. Even when it seems likely that the authors were the coders, it is usually not clear which authors were involved and to what extent. In several cases (e.g., Bellini, Peters, Benner, & Hopf, 2007; Corrigan, Morris, Michaels, Rafacz, & Rüsch 2012; Eisend, 2009) full or partial coding was done by unnamed graduate students or undergraduates with no description of their expertise. In other cases, coding was done by someone who had a vested interest in a particular outcome. This is particularly troubling because meta-analysis has been used to justify decisions that had great human and financial consequences.

There is plenty of cause to suspect that the reliability and validity of the coding process has improved over the years. There are now template codebook spreadsheets available to help researchers organize data more efficiently and avoid time-consuming data reclassification mistakes (The Cochrane Collaboration, 2015), and guidelines are widely available on when and how to conduct consensus meetings (Cooper, Hedges, & Valentine, 2009). Journals are becoming more selective and are currently refining their requirements for meta-analysis in the interest of increasing validity among other goals (Weller, 2001). In the
field of Medicine, all contributors to a meta-analysis are required to be listed as either authors or contributors, and there is a mandatory statement of disclosure about each person’s task and about any potential conflict of interest (Higgins, 2011a). Most impressively, the Cochrane Collaboration was formed with the goal of creating completely transparent healthcare meta-analyses with a standardized format (Higgins, 2011b). However, we have no way of knowing how meta-analysis has actually evolved over time and very little perspective on how we use it in relation to other fields. Finally, we do not know whether any lessons learned have translated into better quality meta-analyses.

The question of whether meta-analysis has become more reliable and valid can be broken down into two other questions: “Do coders have minimal error in coding and agree on operationalizations?” and “Is the meta-analysis thorough and replicable?” Because of the difficulty of meta-analysis and the propensity for error, having a high level of accordance amongst coders is evidence that errors have been caught and corrected before analysis. In terms of subjectivity, agreement suggests that experts were able to reconcile their opinions on the data, and that the data are likely meaningful. Furthermore, it provides assurance that if there was a disparity in coder experience or even if all coders were relatively unskilled, there was still consistency of opinion. This information is quantified in meta-analysis by measures of IRA and IRR.

The second question of whether the meta-analysis was thorough and replicable is essential to reliability and validity. Without enough information as to procedure, it is impossible to make a judgment about reliability or validity. Similarly, if the work cannot be
replicated then it is neither reliable nor valid. To measure the degree of rigor in a meta-analysis one can use a validated rigor benchmark.

**Conceptual Model and Variables**

Because there was very little preliminary literature on meta-analysis from a meta perspective, this study was necessarily exploratory. The Conceptual Model shown in Figure 1 was used to guide research and analysis. In the following sections the variables and their operationalizations for this study are discussed.

*Figure 1. Conceptual Model*
Study Elements

**Time period.** The term “meta-analysis” was coined in a discussion paper by the psychologist Gene Glass (1976) for the purpose of drawing conclusions from numerous but muddled findings in Psychology. He noted that in the hard sciences, ten studies might be enough to resolve the nature of a construct, but in the social sciences one might not see the same pattern of results even twice in ten studies. Early meta-analytic procedure was simple; a bare-bones meta-analysis is an average of all studies with each study being weighted by the size of its sample. What followed was a period of expository articles, books, and proceedings focused on strategies to address the inherent weaknesses of bias and attenuation due to unreliability (e.g., Chalmers, 1991; Abrami, Cohen & d’Apollonia, 1988; Hedges, 1982; Hedges, 1986; Levine & Hunter, 1983). As meta-analysis became more popular, critics concerned with this lack of consensus began to focus on the human processes that introduced error, namely coding strategies and issues related to reliability between raters. Although there is no guiding research on this topic, some speculations can be made. First, as researchers and publishers became more well-informed about technique and meta-analyses with multiple coders became the norm, it is possible that there were corresponding upward trends in IRA and IRR values. One might also expect the relative use of IRA to IRR to change over time as meta-analysts began to use more IRR measures where appropriate. Finally, changing publishing requirements in each field regarding minimum IRA or IRR values might be reflected in the analysis of time.

Figure 1 proposes that Time Period works through the coding task-related elements of meta-analysis number of coders (NCoders), total number of variables (TotalVs), and
proportion of high inference variables (HINF) to low inference variables (LINF), where inference refers to the degree of discretion required of coders. Documenting changes in these variables could explain a great deal about how meta-analyses have evolved. The time periods (three year groupings) used in this study was conceived for practical and theoretical considerations. Practically, missing data produces gaps that can obscure patterns in data visualization, and deciding on time periods up front eliminates a source of potential misunderstanding of the figures. Theoretically, three year groupings were chosen from an estimation of the lag time from starting a meta-analysis to publication. The large time demands associated with meta-analysis prevent immediate adoption of best practices. This lag could obfuscate statistical detection of change over time, so time periods were used to help account for some of this lag. A Cochrane Collaborative report stated that the typical meta-analysis took 23 months from start to publication in the Cochrane Online Library (Cochrane Oversight Committee, 2012). We also know that the average length of time from submission to publication in a journal or conference is about 24 months (Kling & Swygart-Hobaugh, 2002). Thus, increment of three years seems like a reasonable period of time to group meta-analyses.

Field of study. In exploring the relationship between field of study and the other variables illustrated in Figure 1, it should be noted that meta-analysis, like any other research methodology, was not adopted simultaneously across all fields of research, nor was it likely to have developed identically. Interdisciplinarity and paradigm development of each field may have some influence on how meta-analysis spread. Interdisciplinarity refers to the amount of research collaboration between researchers across multiple fields. Such cross-
fertilization of ideas has been described as critical to the advancement of science (Morillo, Bordons & Gómez, 2003). Fields that are less interdisciplinary are more isolated from research that originates externally and are less likely to integrate those findings. Paradigm development refers to how a field has developed shared theoretical structures and methodological approaches about which there is a high level of consensus (Pfeffer, 1993). Fields that are low in paradigm development are more fragmented, with more definitions and operationalizations for the same term, and they are less unified on methodology. In addition to Computer Science, Education, Medicine, and Psychology provide a representative window through which to view integration of meta-analysis under varying conditions of interdisciplinarity and paradigm development. Although it is impossible to guess at all the ways that meta-analyses in these fields differ, it is reasonable to suspect that they are in fact different.

IRA and IRR values are suspected to vary by field, perhaps due to how concrete a field is or differing coding guidelines. Oremus et al. (2012) conducted a study on the peer review process to discover how highly educated reviewers differed in how they reviewed studies. They found that when coders are stringently observed and were provided with explicit guidelines, they were less likely to achieve high IRA or IRR values, and low IRR values were more accurate to the situation. Such explicit guidelines are characteristic of fields with high paradigm development, where there exists the expectation that there is little room for further interpretation on a construct or technique by the time it is well-developed enough to be included in a meta-analysis. It is unclear what is the role of publishing outlets for each field in terms of minimum requirements and standards for IRA and IRR values. A
low IRA or IRR value might be discouraged amongst publishers, depending on the level of tolerance and understanding of the field, and they may act as gatekeepers to keep IRA and IRR at a certain level. For these reasons, this study explored whether variability in IRA and IRR was due to field.

Similarly, meta-analyses in some fields may differ systematically from others in their coding procedures. Exploring the relationship between field and the proportion of high inference variables to low inference variables was done to provide an insight into how each field tends to structure their meta-analyses. Meta-analyses in Medicine may not have any high inference variables to contend with at all, while Education may have a much larger proportion of high to low variables. Further, clinical trials collect a great deal of demographic information on participants such that they are accounted for; these low inference variables may be prevalent. So far, there is no research that describes the makeup of meta-analyses by field, and this study attempted to fill that gap in knowledge.

**Number of coders (NCoders).** Achieving the common frame of reference necessary for acceptable IRA or IRR is the most complicated step in conducting a meta-analysis, and while adding more coders improves the perception of possible bias, it also means integrating more perspectives into the common frame of reference. Inexperienced coders also present a challenge, because they must be trained to have a highly advanced level of understanding of the variables of interest (Oremus et al., 2012). Ultimately, everyone involved in coding a meta-analysis must come to think of each variable the same way and share the same conceptual approach. Communication between coders must take place every time a coder encounters information that does not fit or challenges the existing operationalization and the
resulting decision rule recorded (Orwin & Cordray, 1985). Some codes are easier to agree on, but for those variables that require effort, more coders may attenuate IRA and IRR values.

**Total number of variables (TotalVs).** Coding for meta-analysis is inarguably fatiguing in terms of time and mental expenditure, and the total number of variables was selected as a representation of the overall workload. As such, it was expected to influence IRA and IRR values. The number of variables that require coding per study is the most direct indication of fatigue because it tells us how much work needed to be completed, and it is indicative of the scope and spread of the meta-analysis (Bornmann, Mutz, & Daniel, 2010). Although some meta-analyses are highly focused in terms of variables explored, many others include a large number of variables to maximize the chance of finding a significant effect. It is impossible to tell from reading a meta-analysis how many variables were coded originally but did not make it into the published work, so it is not a perfect predictor of fatigue or scope. However, it stood to reason that, on average, a study reporting a greater total number of variables might have lower IRA and IRR values.

**Variable type.** The variables in this category were meant to provide a quantitative estimate of the difficulty of coding each item. Cooper et al. (2009, pp. 99-102) described the source of this difficulty as being the amount of inference required of coders, and that “low inference” as opposed to “high inference” variables should require less effort and time to code. Coding for low inference variables was argued to be more accurate overall and to be preferred over high inference coding where possible. At worst, it has been suggested that, for very high inference variables, coders become just another level of participant in the research
and should be treated as such by researchers (Greenberg & Folger, 1988, pp. 177-194). Meta
variable type was not coded a priori, but emerged as a variable that provided valuable insight.

**Low inference.** Low inference variables are those that have a single
operationalization according to the meta-analysis and thus are fairly simple to find in primary
studies and transcribe for meta-analysis. For example, intelligence has many accepted names,
definitions, and assessments. However, if the operationalization calls for the result of the
Wechsler intelligence scale for children (Wechsler, 2003) because it is specifically of
interest, then it does not require the coder to evaluate whether the test measures intelligence
in a way that is aligned closely enough with his own conception of the construct, nor is
evaluation of the sample or setting necessary. Most demographics fall under the low
inference heading, as well as contextual information such as study setting (lab vs. field) and
most binary variables because they typically represent presence or absence (e.g., presence or
absence of rash). It was expected that low inference variables were less controversial than
high inference variables, and associated with higher IRA and IRR values.

**High inference.** High inference variables are those that have two or more acceptable
operationalizations according to the meta-analysis. For example, adopting a broad definition
of intelligence would require coders to evaluate every source carefully. Not all high inference
variables have multiple operationalizations, however. Occasionally complex judgments must
be refined into a binary code. An example of binary operationalization that is still high
inference would be where a coder had to determine whether a student was given support
during a task where support was broadly defined. It was expected that high inference
variables are associated with lower IRA and IRR values.
**Meta variables.** Meta variables are those variables that are not found in the primary literature text at all; instead they are observations or evaluations of the primary studies. Meta variables were generally assessments of quality and were used to weight or otherwise assess data at the primary study level. Meta variables can be either high or low inference, but they were unique in that they are somewhat confounded with rigor because assessment of quality is one of the points that make up the rigor score, and the presence of meta variables tend to fulfill that point.

**Rigor.** Rigor in meta-analysis is a difficult construct to measure because the amount of description and disclosure of procedure is typically poor. Primary study quality measures ask the reader to assess how well the study accomplished a methodological goal from commonly reported information, but with meta-analysis it is just as likely that the information is not present at all. Regardless of when it was published, meta-analytic rigor within Psychology has been inconsistent (London, 2011). Rigor in this study was measured by counting the number of points on Oxman, Cook, and Guyatt's (1994) six-question guide on how to determine whether a meta-analysis is valid, from the perspective of the reader. This guide was published in the medical field, is cited relatively frequently, and is still referenced often today. It has been adapted by the authors to fit the common language of multiple fields because the areas are broadly relevant to meta-analysis, but still specific enough to operationalize. The areas addressed by Oxman, Cook and Guyatt (1994) and the operationalized points that comprise each area are shown in Appendix A. The broad questions recommended by the authors are:
1. Did the overview address a focused clinical question?
2. Were the criteria used to select articles for inclusion appropriate?
3. Is it unlikely that important, relevant studies were missed?
4. Was the validity of the included studies appraised?
5. Were assessments of studies reproducible?
6. Were the results similar from study to study?

The inclusion criteria feature prominently because they most directly address bias and disclosure. There is evidence that less disclosure of meta-analytic features such as inclusion criteria and variable operationalization is associated with larger results (London, 2011; Oxman, et al., 1994), and that this phenomenon is more prevalent with meta-analysis than other methodologies (Ravnskov, 1992). London (2011) found that certain aspects of rigor such as validity criteria for inclusion in the sample of studies (<50%) were reported to a much lesser extent than other aspects. In addition, Markowitz and Hancock (2015) were able to successfully predict retractions with an obfuscation index they created to describe disclosure avoidance language in medical articles. There are obvious incentives to bias research results, especially in a publish-or-perish academic environment, or where profits are at stake. Gøtzsche (1987) reviewed citations in meta-analyses that examined a new non-steroidal anti-inflammatory drug for rheumatoid arthritis, and out of a pool of 77 possible primary studies that were roughly equal in positive and negative outcome, 44 (60%) of the meta-analyses cited a higher proportion of positive outcomes. The present study describes how much variability there is in rigor for a representative sample of meta-analyses.
**IRA and IRR.** At its heart, meta-analysis consists of a series of complex professional judgments about other researchers’ work. It is a mistake to think that the scraping process consists of simply transcribing a piece of information from a primary study to the data spreadsheet for the meta-analysis and that transcription error is the main threat to validity (Card, 2011). In reality, the process requires a great deal of human decision-making and accommodation, and is compounded by the number of humans working on the task. IRA and IRR represents the mathematical approximation of the validity of the coders’ judgments.

The reason for choosing either IRA or IRR is often a matter of convention rather than best practice, but their underlying statistical assumptions are different enough that equating the two is inappropriate. IRA is very common in Psychology and is often operationalized as simply the percent of total items that coders scored identically. Other common types of IRA are Cohen’s kappa, intraclass correlations, or rwg (LeBreton & Senter, 2008). IRA differs from IRR in that it measures the “extent to which the different [coders] tend to make exactly the same judgments about the rated subject” (Tinsley & Weiss, 1975, p. 359). Because meta-analyses often require binary coding, IRA can be more appealing than IRR (Tinsley & Weiss, p. 359). IRR measures the “degree to which the ratings of different [coders] are proportional when expressed as deviations from their means” (Tinsley & Weiss, p. 360). The effect of this is that if one coder scored an item in a consistently different way than another coder, the IRR would still be high. IRR is more appropriate when there is an interval judgment to be made, for example on quality or performance where a coder can be said to be more lenient or strict than another. The most common measure of IRR is the intraclass correlation, a calculation similar to a Pearson correlation except that data are centered and scaled using a pooled mean.
and standard deviation (LeBreton & Senter, 2008). Some methods are more conservative than others and a discussion of relative strengths and weakness is beyond the scope of this paper (see Brown & Hauenstein, 2005), but all purport to measure conformity between raters.

One might expect that perfect interrater reliability or agreement should be the goal of meta-analysts, but this is not practically true. Coding studies can be extraordinarily difficult, requiring hundreds of hours of work and discussion. The threat of fatigue cannot be overstated; in addition to errors, the temptation to concede or compromise as a convenience is strong. Complete reliability amongst multiple coders can be almost impossible to achieve.

For consumers of research who are simply looking for an effect size, IRA and IRR may not be outwardly concerning. However, this information provides a window into the black box that is the coding process. In order to describe the overall state of IRR regardless of field and time, the variability of IRR was examined to provide an overview of IRR’s consistency or lack thereof.

On the surface, higher rigor might seem to predict greater IRA and IRR. Either the variables were not complex enough to warrant disagreement or coders were aligned in their perspectives, either of which speaks to greater validity of the estimated effect size (i.e., outcome of a meta-analysis; Oxman et al., 1994). However, we should explore this assumption because less meta-analytic rigor has been found to predict higher estimated meta-analytic effect sizes in published Psychology meta-analyses (London, 2011). In other words, the less information disclosed about its methodology, the stronger the estimated effect size result of the meta-analysis. This evidence is supported by another study that suggests lower interrater reliability, rather than higher reliability, is a product of more rigorous coding.
(Oremus et al., 2012). Thus, this study attempted to quantify a bidirectional relationship between IRA and IRR.

In summary, this study describes the state of meta-analysis over time and across the fields of Computer Science, Education, Medicine and Psychology while exploring the intermediate influence of number of coders, total number of variables, and percent low inference, percent high inference, and number of meta variables as described in Figure 1. The following questions were addressed, starting with a focus on the DVs:

   Research Question 1: How many meta-analyses are available each year?

   Research Question 2: Are meta-analytic Rigor, IRA, and IRR interrelated?

   Research Question 3: Accounting for Time Period, how do meta-analytic Rigor, IRA, and IRR each vary by Field of Study?

   Research Question 4: Accounting for Time Period, how does Field of Study affect Number of Coders, Total Number of Variables, and Variable Type?

   Research Question 5: Individually, which of the proximal variables (Number of Coders, Total Number of Variables, and Variable Type) affect IRA and/or IRR?

In addition to these initially proposed research questions, it was expected that insights would be discovered as the research is conducted. The findings are discussed in the results section, and implications and connections are explained in the discussion section.
Method

Sampling Method

In total, 153 meta-analyses were sampled from four domains: Computer Science, Education, Medicine, and Psychology from the Web of Science Core Collection. An attempt was made to find one for each year since 1977, the first year a meta-analysis was published. In addition, Inspec, an engineering database, was used to find additional meta-analyses for Computer Science. Meta-analyses were randomly selected by downloading a large list of citations for each year using the following search criteria: TI= “meta-analysis”. A random number generator was used to select articles for inspection from that list. In the instance where no IRA or IRR containing meta-analyses was found for a particular year, an additional meta-analysis from within the three-year time period was included if possible. This was not always possible because of sparse populations of meta-analyses containing this information, and in those cases a meta-analysis that did not contain IRA or IRR was coded for all other variables to flesh out the other analyses.

Definition of Fields

Computer Science. The Computer Science search was conducted with two methods. The first was to search the Web of Science Core Collection using the Subject Area refinement for Computer Science; this narrowed findings to those that fell in that subject area. The title search “meta-analysis” was then used. Inspec was also searched with the following subject keywords: “comput*, robot*, cyber*, programming, IEEE, software,” as well as “meta-analyses” in title.
**Education.** The Web of Science Core Collection was searched with the Subject Area refinement for Education and the in-title keyword “meta-analysis.” No further databases were searched, as ERIC and most other well-known Education databases are included in Web of Science. Additional subject keywords “educ*, learn*, teach*, school*, university,” and “student*” were also used without the subject area refinement when the initial search was exhausted.

**Medicine.** Medline was used to search for meta-analyses relating to Medicine using the in-title keyword “meta-analysis.” Medicine is an incredibly broad field, so it was necessary to be specific on the areas that would best inform this research. Medicine for this study is limited to those fields that deal with clinical trials and medical research using human subjects. Behavioral therapies and programmatic evaluations were numerous, but were not included as they were quite different in form from high stakes research that set Medicine apart from the other fields.

**Psychology.** The Web of Science Core Collection was used to search for articles relating to Psychology. The Psychology subject area refinement was used. The constraints described in the previous fields were also applied to Psychology (e.g., Educational Psychology).

**Field Overlap**

In some instances, meta-analyses overlapped the content area of two fields and were difficult to categorize, and these areas were eliminated. There was frequent overlap in three areas in particular. The Psychiatry literature was an overlap Medicine and Psychology, and they showed up in journals of both kinds as well as in Psychiatry-specific literature. Articles
also showed characteristics of both, such that there were clinical trials as well as behavioral research. Unsurprisingly, Education overlapped with all of the other fields because there is always a need to learn. Hybrid research areas such as Computer Science Education and Educational Psychology were accepted only if they were clearly one or the other, and this was determined by content, author background, and journal of publication. However, they were often a true hybrid, for example human-computer interaction studies combined with student acceptance and outcome research in the same meta-analysis. These meta-analyses were eliminated from the study.

**Inclusion Criteria**

Inclusion criteria are minimal due to the focus on how meta-analyses are conducted rather than any specific content area. The only criteria for inclusion was whether the journal is English-language or not and whether the study uses an agreement or reliability index. It was expected that meta-analyses would vary in topic, quality, type (odds-ratio vs. correlational), and effects model (random vs. fixed), amongst other characteristics. By keeping the meta-analytic sampling methodology mostly random in terms of characteristics, a representative sample of the population of meta-analyses was expected. In the cases where there were insufficient meta-analyses meeting the criteria of having an IRA or IRR measure, a meta-analysis lacking this criterion from the same time period was accepted in order to inform the other exploratory goals of this study.

**Coding of Dependent Variables**

**Coders and IRA.** Two coders, the author and a Psychology Ph.D., coded 100% of the variables. Weekly meetings were held during data collection, for a total of 26 meetings.
Consensus was realigned at the beginning of each meeting. IRA at the end of the coding phase was 88%, and final disagreements at the end of 26 weeks were resolved by the author.

**Consolidating IRA and IRR values.** Agreement percentages and kappa values were converted to z-scores. Where both values were reported, they were averaged to obtain a single IRA value. There were no other types of IRA (e.g., rwg) reported. For IRR, only intraclass correlations were reported, so there was no need to consolidate.

**Rigor.** Count of disclosure points, from 0-10 as per Oxman et al.’s (1994) document on how to assess the validity of a meta-analysis as a reader, was used to calculate rigor.

**Coding of Independent Variables**

**Time period.** To measure time period, year of publication of each meta-analysis was coded and then categorized into one of twelve three-year increments starting in 1977, the first year a meta-analysis was published.

**Field.** Field is categorical, representing each of the four domains of interest: Computer Science, Education, Medicine, and Psychology. Because each journal was specifically drawn from one of these areas, field was coded at the inclusion stage of this study.

**Number of coders.** Number of coders was transcribed directly from the meta-analyses where possible. In cases where the number of coders was not noted but where there was some evidence of plurality, inferring a value was attempted. Meta-analysts indicate that “we” were responsible for the coding, the number of coders was counted as the number of authors. If the discussion only indicates that “coders” were involved, then the study assumed that there were two coders.
**Total number of variables.** Total number of variables was a count of all variables, per meta-analysis, that were apparent in each of the meta-analyses. This included values that were listed in tables, or from demographic variables and study characteristics. In the event that the meta-analyst discussed variables that were coded but not subsequently included in the final analyses, those variables were counted because they were representative of the coding effort. On occasion, a variable would be comprised of other variables, such as Total Number of Variables in this study. In this case, the variables comprising it would be counted, but the calculated variable would not because it was not the product of coding. Excluded variables included primary research title, authors, years, publication status, and sample size. These variables were not included because they are necessary for every meta-analysis, and only some meta-analyses report their measurement. With the exception of sample size, those variables are also available in the citation and do not require meaningful coding effort.

**Percent high inference variables.** Variables identified as high inference were counted and then calculated as a percentage of total number of variables. High inference variables are defined as those with multiple operationalizations or binary variables that require judgment. On the rare occasion it was unclear whether a variable should be high or low inference based on these criteria, the variable was counted as high inference.

**Percent low inference variables.** Variables identified as low inference were counted and then calculated as a percentage of total number of variables. Low inference variables are defined as those that have a single operationalization. These variables could be binary (1 or 0) or a value or string (42, male) so long as they did not require judgment.
**Number of meta variables.** Variables that meta-analysts created for the purpose of evaluating primary research were counted. Usually, it was a quality judgment or scale about the fitness of the study for inclusion in the meta-analysis. Meta variables were also either high or low inference, and were included in the calculation of high and low inference.

**Journal impact factor.** The attempt was made to find a Journal Citation Reports® (JCR®; 2015) impact factor score for every journal. In those cases where an impact score existed, the score for the year of publication of the meta-analysis was used. In most cases the impact factor scores were only available starting in 1997, so any publication date earlier than 1977 was assigned the score for 1997. Two journals could not be found in the JCR database. None of the scores for proceedings from conferences were available from JCR. A few impact factor scores from other sources were found but they were not used due to a lack of source consistency, rendering them meaningless for comparison.

**Pages.** Pages was the number of pages starting at the main body of the meta-analysis and ending on the last page of text before references. The purpose of this variable was to determine whether publication restrictions on length was associated with rigor. Tables and appendices were included, but references were not included because the intention was to measure the main body of text, and references would be confounded with number of studies. Furthermore, formatting for references is done differently by field; the formatting in Medicine is highly abbreviated and therefore takes less space.

**Number of authors.** Number of authors was simply a count of the number of authors listed on the meta-analysis. Contributors, where listed, were not included.
**Number of failsafes.** The number of failsafe variables consisted of a count of the following methods used in the meta-analysis to detect publication bias: the funnel plot, failsafe N, Beggs method for detecting variance of the funnel plot (Begg & Masumdar, 1994), and the Eggers test of asymmetry of the funnel plot (Egger, Smith, Schneider & Minder, 1997). A single meta-analysis could have a value of four for this variable if it contained all of these methods.

**Results**

**How Many Meta-Analyses are Available Each Year? (RQ1)**

The Web of Science “Analyze Citations” feature was used to get counts of meta-analyses by year to get a picture of overall availability of meta-analyses seen in Figure 2. See Appendix B for the exact search terms used for the counts. Computer Science meta-analyses were scarce, especially in early years. This chart does not account for those meta-analyses present in the proceedings literature, where many of the Computer Science meta-analyses from this study were drawn. Furthermore, many of these meta-analyses from that field were false, an issue discussed later in this paper. Education had an uneven distribution due to lower volume, with very little consistency from one year to the next until about 2007, when there is a noticeable and consistent increase in publications. The distribution of Medicine was exponential. Due to its scale it was more informative to show two charts, one showing the first half of the timeframe and one showing the second. Use of meta-analyses is nonexistent until 1985, but quickly increases afterwards. This pattern appears to be specific to operationalization of Medicine used for this study, as meta-analyses on program and behavioral evaluations were conducted before 1985. The gap may be due to a prolonged
period of discussion about the appropriateness of meta-analyses for high-stakes stakes clinical trials and other research examining how pharmaceuticals and procedures affect human patients. DerSimonian & Laird (1986) seem to have broken the drought with an article describing how meta-analysis could and should be used in such research, and their research was cited more than 17,000 times. Psychology had a fairly uniform distribution with slight increase until with a marked increase around 2006.

Figure 2. Meta-analysis Search Result Volume by Year
Are Meta-Analytic Rigor, IRA, and IRR Interrelated? (RQ2)

An attempt was made to code one meta-analyses for each year since it was published until 2014, or 37 meta-analyses. In the case of Computer Science and Medicine, this was not successful. For Computer Science, every discoverable meta-analysis that qualified as such was coded. In the case of Medicine, the late start to the literature meant it was not possible to find meta-analyses in the early years, although they were overrepresented in later years. It was difficult to find meta-analyses that contained a measure of IRA or IRR. It was more common for meta-analysts to write something to the effect of “The results of coding were compared, and disagreements were resolved by consensus,” never giving a value for agreement before codes were consolidated. Table 1 shows how many meta-analyses containing measures of each type were collected by field. IRR was reported very infrequently overall, despite the fact that it was often an appropriate measure. IRA was reported more frequently, but overall it was still reported less than half of the time.

Table 1.

Frequencies of IRA and IRR Values Available in Sample

<table>
<thead>
<tr>
<th>Field</th>
<th>IRA</th>
<th>IRR</th>
<th>Total Coded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp Sci</td>
<td>7 (27%)</td>
<td>2 (8%)</td>
<td>26</td>
</tr>
<tr>
<td>Education</td>
<td>22 (55%)</td>
<td>1 (3%)</td>
<td>40</td>
</tr>
<tr>
<td>Medicine</td>
<td>13 (38%)</td>
<td>5 (15%)</td>
<td>34</td>
</tr>
<tr>
<td>Psychology</td>
<td>25 (50%)</td>
<td>5 (10%)</td>
<td>50</td>
</tr>
<tr>
<td>Total</td>
<td>66 (46%)</td>
<td>13 (9%)</td>
<td>145</td>
</tr>
</tbody>
</table>
While conducting data collection it was noted that reporting IRA and IRR seemed to fall out of favor for certain years and then once again regained popularity. The most significant pattern of IRA reporting was in Medicine, where authors stopped reporting IRA or IRR for the coding process around 2005 and began to report it for study selection instead. It was much more difficult to find studies containing IRA and IRR values in later years for this reason; unfortunately, this is not quantifiable because number of meta-analyses unsuccessfully searched was not coded in this study. Aside from a growing awareness of the importance of study selection in the field, the author could find no evidence as to why this change occurred. This slightly reduced the sample size for the Medicine field for the IRA and IRR variables.

IRA was significantly associated with a number of other variables as shown in Table 2 and means and standard deviations are shown in Table 3. Most interestingly, IRA was positively correlated with Time Period \( (r = .27, p = .03) \), meaning later meta-analyses had higher IRA values than earlier meta-analyses. There was not a pattern of higher or lower IRA associated with any of the fields except for Computer Science \( (r = .26, p = .04) \). There was no association between Journal Impact Factor and IRA or between Total Number of Variables and IRA.

IRR was strongly associated with several variables. IRR had a strong negative correlation with Medicine \( (r = -.76, p < .01) \) and Count of Meta Variables \( (r = -.64, p = .02) \). This indicates that Medicine tended to have low IRR values and that meta-analyses that had higher IRR values contained few or no Meta Variables. It was also negatively associated with
Number of Total Variables ($r = -.67, p = .01$), meaning meta-analyses with higher IRR tended to have fewer variables.

Rigor was correlated most strongly with the field of Medicine ($r = .57, p < .01$) and Number of Meta Variables ($r = .40, p < .01$). It was also correlated with Journal Impact Factor Score ($r = .35, p < .01$). It was not, however, correlated with Number of Pages ($r = -.04, p = .62$), suggesting that length of an article was not a factor in how much disclosure was present in the study.
Table 2.

**Correlation Matrix of Variables with Variable N**

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
<th>15</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>IRA</td>
<td>.032</td>
<td>-.167</td>
<td>-.128</td>
<td>-.149</td>
<td>.001</td>
<td>-.108</td>
<td>.244*</td>
<td>-.085</td>
<td>-.135</td>
<td>.034</td>
<td>.257*</td>
<td>-.065</td>
<td>.088</td>
<td></td>
</tr>
<tr>
<td>2.</td>
<td>IRR</td>
<td>7</td>
<td>1</td>
<td>-.627*</td>
<td>-.060</td>
<td>-.667*</td>
<td>-.197</td>
<td>.197</td>
<td>-.647*</td>
<td>-.109</td>
<td>.493</td>
<td>-.762**</td>
<td>.222</td>
<td>.199</td>
<td>-.302</td>
</tr>
<tr>
<td>3.</td>
<td>Rigor</td>
<td>66</td>
<td>13</td>
<td>1</td>
<td>.207*</td>
<td>.296**</td>
<td>-.133</td>
<td>.133</td>
<td>.431**</td>
<td>.140</td>
<td>-.111</td>
<td>.582**</td>
<td>-.195*</td>
<td>-.277**</td>
<td>.343**</td>
</tr>
<tr>
<td>4.</td>
<td>NCoders</td>
<td>66</td>
<td>13</td>
<td>145</td>
<td>1</td>
<td>.141</td>
<td>.182*</td>
<td>-.182*</td>
<td>.199*</td>
<td>.150</td>
<td>-.056</td>
<td>.176*</td>
<td>-.161</td>
<td>.069</td>
<td>.039</td>
</tr>
<tr>
<td>5.</td>
<td>TotalVs</td>
<td>66</td>
<td>13</td>
<td>148</td>
<td>144</td>
<td>1</td>
<td>.012</td>
<td>-.012</td>
<td>.464**</td>
<td>.173*</td>
<td>-.072</td>
<td>.269**</td>
<td>-.021</td>
<td>-.186*</td>
<td>.118</td>
</tr>
<tr>
<td>6.</td>
<td>HINF</td>
<td>66</td>
<td>13</td>
<td>148</td>
<td>144</td>
<td>149</td>
<td>1</td>
<td>-.100**</td>
<td>.183*</td>
<td>.170*</td>
<td>-.014</td>
<td>-.150</td>
<td>-.029</td>
<td>.220**</td>
<td>.090</td>
</tr>
<tr>
<td>7.</td>
<td>LINF</td>
<td>66</td>
<td>13</td>
<td>148</td>
<td>144</td>
<td>149</td>
<td>149</td>
<td>1</td>
<td>-.183*</td>
<td>-.170*</td>
<td>.014</td>
<td>.150</td>
<td>.029</td>
<td>-.220**</td>
<td>-.090</td>
</tr>
<tr>
<td>8.</td>
<td>Meta</td>
<td>66</td>
<td>13</td>
<td>148</td>
<td>144</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>1</td>
<td>.161</td>
<td>-.253**</td>
<td>.557**</td>
<td>-.136</td>
<td>-.146</td>
<td>.226*</td>
</tr>
<tr>
<td>9.</td>
<td>Time P.</td>
<td>67</td>
<td>13</td>
<td>150</td>
<td>146</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>1</td>
<td>-.253**</td>
<td>.068</td>
<td>-.106</td>
<td>.320**</td>
<td>-.023</td>
</tr>
<tr>
<td>10.</td>
<td>Psych</td>
<td>67</td>
<td>13</td>
<td>150</td>
<td>146</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>168</td>
<td>1</td>
<td>-.327**</td>
<td>-.362**</td>
<td>-.350**</td>
</tr>
<tr>
<td>11.</td>
<td>Med</td>
<td>67</td>
<td>13</td>
<td>150</td>
<td>146</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>168</td>
<td>1</td>
<td>-.275**</td>
<td>-.266**</td>
<td>.390**</td>
</tr>
<tr>
<td>12.</td>
<td>Ed</td>
<td>67</td>
<td>13</td>
<td>150</td>
<td>146</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>168</td>
<td>1</td>
<td>-.294**</td>
<td>-.286**</td>
<td>-.184**</td>
</tr>
<tr>
<td>13.</td>
<td>C.Sci</td>
<td>67</td>
<td>13</td>
<td>150</td>
<td>146</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>168</td>
<td>1</td>
<td>-.245**</td>
<td>.016</td>
<td>-.135</td>
</tr>
<tr>
<td>14.</td>
<td>JCR</td>
<td>60</td>
<td>11</td>
<td>128</td>
<td>124</td>
<td>127</td>
<td>127</td>
<td>127</td>
<td>127</td>
<td>127</td>
<td>140</td>
<td>140</td>
<td>140</td>
<td>140</td>
<td>140</td>
</tr>
<tr>
<td>15.</td>
<td>Pages</td>
<td>67</td>
<td>13</td>
<td>150</td>
<td>146</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>149</td>
<td>157</td>
<td>157</td>
<td>157</td>
<td>157</td>
<td>157</td>
</tr>
</tbody>
</table>

*Correlation is significant at the 0.05 level (2-tailed).

**Correlation is significant at the 0.01 level (2-tailed).

Note. Correlations are in the right half of the matrix, sample size is on the left side of the matrix.
A cluster of high correlations around Medicine illustrated several ways in which it differed from other fields. Namely, Medicine implemented meta variables that usually measured quality often \((r = .56, p < .01)\), whereas other fields did not. As mentioned above, Medicine was also associated with higher rigor scores \((r = .57, p < .01)\), which is slightly contaminated by the presence of metavariables. Higher Journal Impact Factor scores were also associated with Medicine \((r = .11, p = .53)\), and interestingly, shorter papers were also associated with Medicine \((r = -.30, p < .01)\). This suggests that meta-analyses in Medicine were written in such a way to maximize disclosure in fewer words.

Table 3.

*Variable Means and Standard Deviations by Field*

<table>
<thead>
<tr>
<th>Variable</th>
<th>Comp Sci</th>
<th>Education</th>
<th>Medicine</th>
<th>Psychology</th>
</tr>
</thead>
<tbody>
<tr>
<td>IRA</td>
<td>95.33 (1.53)</td>
<td>85.71 (16.25)</td>
<td>89.28 (9.12)</td>
<td>84.38 (17.65)</td>
</tr>
<tr>
<td>IRR</td>
<td>92.00 (0.00)</td>
<td>96.25 (-)</td>
<td>72.00 (15.13)</td>
<td>94.20 (5.85)</td>
</tr>
<tr>
<td>Rigor</td>
<td>5.65 (2.43)</td>
<td>6.28 (1.68)</td>
<td>9.35 (1.32)</td>
<td>6.64 (1.75)</td>
</tr>
<tr>
<td>NCoder</td>
<td>2.52 (1.34)</td>
<td>2.08 (0.77)</td>
<td>2.73 (1.39)</td>
<td>2.28 (0.83)</td>
</tr>
<tr>
<td>TVar</td>
<td>11.00 (6.73)</td>
<td>15.83 (14.57)</td>
<td>22.53 (15.11)</td>
<td>14.96 (10.11)</td>
</tr>
<tr>
<td>HINF</td>
<td>51.90 (32.51)</td>
<td>35.44 (29.73)</td>
<td>28.49 (30.77)</td>
<td>36.32 (28.71)</td>
</tr>
<tr>
<td>LINF</td>
<td>48.10 (32.51)</td>
<td>64.56 (29.73)</td>
<td>71.51 (30.77)</td>
<td>63.68 (28.71)</td>
</tr>
<tr>
<td>Meta</td>
<td>0.28 (0.68)</td>
<td>0.85 (2.52)</td>
<td>7.97 (9.63)</td>
<td>0.10 (0.36)</td>
</tr>
<tr>
<td>Failsafes</td>
<td>0.36 (0.62)</td>
<td>0.17 (0.38)</td>
<td>0.40 (0.81)</td>
<td>0.06 (0.24)</td>
</tr>
<tr>
<td>JCR</td>
<td>0.81 (0.72)</td>
<td>1.25 (0.96)</td>
<td>5.34 (5.17)</td>
<td>3.31 (2.78)</td>
</tr>
<tr>
<td>Pages</td>
<td>16.21 (14.98)</td>
<td>20.05 (17.83)</td>
<td>7.94 (6.97)</td>
<td>17.19 (11.27)</td>
</tr>
</tbody>
</table>
Accounting for Time Period, How Do Meta-Analytic Rigor, IRA, and IRR Each Vary by Field of Study? (RQ3)

Next, one-way ANOVAs were performed on IRA, IRR, and Rigor variables using Field of Study as a factor to determine whether they varied by group. Rigor scores ranged from 2 to 11 overall, with Computer Science ($M = 5.65, SD = 2.432$), Education ($M = 6.28, SD = 1.679$), and Psychology ($M = 6.64, SD = 1.747$) falling below Medicine ($M = 9.40, SD = 1.276$). Medicine was also set apart in that the minimum score was 7, whereas the other fields each had at least one score of 2. A one-way ANOVA was conducted and a statistically significant difference was found in rigor between groups, $F(3, 142) = 25.691, p < .01$. Bonferroni post-hoc analysis revealed that rigor was statistically significantly higher in Medicine than the other three fields ($p < .01$). There were no statistically significant differences in Rigor between Computer Science and Education ($p = 1.00$), or between Computer Science and Psychology, ($p = .145$), or between Education and Psychology ($p = 1.00$). The one-way ANOVA for IRA and Field was nonsignificant, $F(3, 63) = 1.809, p < .155$, indicating that there were no differences in between Fields for IRA. Because the sample size for IRR was so low, especially for Computer Science, it was not possible to conduct an ANOVA for these data.

Accounting for Time Period, How Does Field of Study Affect Number of Coders, Total Number of Variables, and Variable Type? (RQ4).

Path analysis was performed to examine whether there was model invariance across the four fields (i.e. does field moderate the relationships between the dependent and independent variables?). As shown in Figure 3, paths were drawn from Time Period
(independent variable) to three dependent variables: Number of Coders, Total Variables and Percent Low Inference Variables (LINF). Several iterations of the path model were examined. The first model was a fully unconstrained model where estimates for each parameter in the model were obtained for each field group (Table 4). The next steps involved constraining the loadings for Field on various paths in the model so that they were equal among groups to determine if the model fit is a better or worse than the fully unconstrained model and the other subsequent models (Table 5).

![Path Model](image)

**Figure 3. Path Model**

Model fit is determined by the chi-square value which should be greater than 0.05, the RMSEA (should be lower than 0.08 and the closer to 0, the better the model fit) and CFI (closer to 1.0 indicates a better model fit). Another measure of model comparison is Akaike Information Criterion (AIC), whereby the lower the AIC the better the model fit. All
iterations of constraints were examined, first by using a fully constrained model where all of the paths were the same between groups, then constraining each individual path separately and finally constraining two paths at a time.

The estimates in Table 4 indicate that in the Computer Science group, there was a significant association between Time Period and Number of Coders (B=0.28, \( p < 0.05 \)), meaning that at the later time period there were more coders. In the Education category there was a significant association between Time Period total variables (B=0.81, \( p < 0.05 \)), meaning that at later time periods there were more variables.

The model fit of the unconstrained model, as measured by the chi-square (\( \chi^2 = 53.48(24), p > 0.05 \)) shows that this a poorly fitting model (other fit parameters are not available for this model). The AIC for this model was 2914.22. The model with the best fit is the one in which total variables was constrained. In this model, the AIC was 2908.39, the lowest among all of the models, and the RMSEA=0.00 and the CFI=1.00. It can be argued that the model which constrained LINF is also a good fitting model compared to the others, with an AIC only slightly higher than the model that constrained total variables (2910.61) with similar values for RMSEA (0.00) and CFI (1.00). In either case, the variable Number of Coders should be left unconstrained while simultaneously constraining total variables or LINF. Field moderates the relationship between Time Period and Number of Coders.
Table 4.

Estimates from full unconstrained model

<table>
<thead>
<tr>
<th></th>
<th>B</th>
<th>S.E.</th>
<th>β</th>
<th>Sig.</th>
<th>r-square</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Computer Science</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Coders</td>
<td>0.28</td>
<td>0.10</td>
<td>0.48</td>
<td>0.01</td>
<td>0.23</td>
</tr>
<tr>
<td>Total variables</td>
<td>0.94</td>
<td>0.52</td>
<td>0.34</td>
<td>0.07</td>
<td>0.12</td>
</tr>
<tr>
<td>LINF</td>
<td>-4.67</td>
<td>2.51</td>
<td>-0.35</td>
<td>0.06</td>
<td>0.12</td>
</tr>
<tr>
<td><strong>Education</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Coders</td>
<td>0.10</td>
<td>0.03</td>
<td>0.42</td>
<td>0.00</td>
<td>0.18</td>
</tr>
<tr>
<td>Total variables</td>
<td>0.81</td>
<td>0.65</td>
<td>0.19</td>
<td>0.22</td>
<td>0.04</td>
</tr>
<tr>
<td>LINF</td>
<td>-2.08</td>
<td>1.32</td>
<td>-0.24</td>
<td>0.12</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Medicine</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Coders</td>
<td>-0.05</td>
<td>0.10</td>
<td>-0.09</td>
<td>0.64</td>
<td>0.01</td>
</tr>
<tr>
<td>Total variables</td>
<td>0.79</td>
<td>1.09</td>
<td>0.14</td>
<td>0.47</td>
<td>0.02</td>
</tr>
<tr>
<td>LINF</td>
<td>-3.08</td>
<td>1.66</td>
<td>-0.33</td>
<td>0.06</td>
<td>0.11</td>
</tr>
<tr>
<td><strong>Psychology</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Number of Coders</td>
<td>-0.06</td>
<td>0.03</td>
<td>-0.24</td>
<td>0.08</td>
<td>0.06</td>
</tr>
<tr>
<td>Total variables</td>
<td>0.67</td>
<td>0.40</td>
<td>0.23</td>
<td>0.09</td>
<td>0.06</td>
</tr>
<tr>
<td>LINF</td>
<td>-0.93</td>
<td>1.15</td>
<td>-0.11</td>
<td>0.42</td>
<td>0.01</td>
</tr>
</tbody>
</table>

*B=unstandardized beta, β=standardized beta

Table 5.

Comparison of model fits with varying path constraints

<table>
<thead>
<tr>
<th>Model</th>
<th>chi-square (df)</th>
<th>Sig.</th>
<th>AIC</th>
<th>RMSEA</th>
<th>CFI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unconstrained</td>
<td>53.48 (24)</td>
<td>0.00</td>
<td>2914.22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fully constrained</td>
<td>32.55 (11)</td>
<td>0.00</td>
<td>2924.77</td>
<td>0.23</td>
<td>0.27</td>
</tr>
<tr>
<td>constrain Number of Coders</td>
<td>16.64 (3)</td>
<td>0.00</td>
<td>2924.85</td>
<td>0.36</td>
<td>0.54</td>
</tr>
<tr>
<td>constrain Total variables</td>
<td>0.17 (3)</td>
<td>0.98</td>
<td>2908.39</td>
<td>0.00</td>
<td>1.00</td>
</tr>
<tr>
<td>constrain LINF</td>
<td>2.39 (3)</td>
<td>0.50</td>
<td>2910.61</td>
<td></td>
<td>1.00</td>
</tr>
<tr>
<td>constrain Number of Coders &amp; LINF</td>
<td>26.57 (7)</td>
<td>0.00</td>
<td>2926.79</td>
<td>0.28</td>
<td>0.34</td>
</tr>
<tr>
<td>constrain Total variables &amp; LINF</td>
<td>14.67 (7)</td>
<td>0.04</td>
<td>2914.89</td>
<td>0.18</td>
<td>0.74</td>
</tr>
<tr>
<td>constrain Number of Coders &amp; Total variables</td>
<td>23.71 (7)</td>
<td>0.00</td>
<td>2923.93</td>
<td>0.26</td>
<td>0.43</td>
</tr>
</tbody>
</table>
Individually, Which of the Proximal Variables Affect IRA and/or IRR? (RQ5).

Multiple regression analysis was used to test whether the Number of Coders, Total Variables, and Percent Low Inference Variables significantly predicted IRA (see Table 6). The results of the regression indicated the model was nonsignificant ($R^2 = .04$, $F(3,47) = 1.64$, $p = .19$). The same analysis with IRR was proposed, but was not possible due to the low sample size.

Table 6.

Multiple Regression Coefficients Predicting IRA

<table>
<thead>
<tr>
<th>Variable</th>
<th>B</th>
<th>SE</th>
<th>$\beta$</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NCoders</td>
<td>-.086</td>
<td>.127</td>
<td>-.097</td>
<td>-.677</td>
<td>.502</td>
</tr>
<tr>
<td>Total Variables</td>
<td>.003</td>
<td>.009</td>
<td>.057</td>
<td>.404</td>
<td>.688</td>
</tr>
<tr>
<td>LINF</td>
<td>-.010</td>
<td>.005</td>
<td>-.303*</td>
<td>-2.166</td>
<td>.035</td>
</tr>
</tbody>
</table>

*p<.05

Additional Findings

False meta-analyses. A pattern of studies titled as meta-analyses became apparent in the Computer Science literature starting in 1994. An example study of this type, “The Measurement of GSS Effectiveness: A Meta-analysis of the Literature and Recommendations for Future GSS Research” (Pervan, 1994), was a vote counting study. Vote counting typically featured in systematic reviews before meta-analysis replaced it as a method. It consisted of counting positive, neutral, and negative results and disregarding study sample size,
moderators, and a comprehensive study inclusion strategy. The majority of writing in these vote counting studies is discussion, similar to narrative reviews, and they did not follow the other conventions of meta-analysis such as reporting interrater agreement. Figure 3 shows the frequency of these publications by time period. For whatever reason, vote counting studies and similar alternatives became more prevalent and actual meta-analyses became scarce, which contributed to the low sample size in that field.

Figure 4. False Meta-Analyses Frequency by Time Period

**Meta variables in Medicine.** Reporting meta variables, or those variables that were a measure of quality of the primary research, was prevalent in the later Medicine literature, and barely seen in other fields. Meta-analysts used homemade checklists for quality using study
features they felt were important. They also used the Jadad Questionnaire (Jadad et al., 1996), which focuses on randomization, blinding, and disclosure of methods. They used the results of these measures to weight or eliminate primary studies. The means for the other fields were well below one, but the mean number of Meta Variables for Medicine is 7.65 ($SD=9.80$).

**Use of failsafes.** Failsafes methods for detecting publication bias, such as funnel plots and their associated diagnostic tests, were more prevalent in every field other than Psychology (see Table 7). Use of failsafes tended to increase in later time periods ($r = .22, p = .01$). However, this trend does not seem to apply to Psychology.

Table 7.

*Descriptives for Failsafes by Field*

<table>
<thead>
<tr>
<th>Field</th>
<th>Yearly Min</th>
<th>Yearly Max</th>
<th>Total Failsafes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp Sci</td>
<td>0</td>
<td>2</td>
<td>10</td>
</tr>
<tr>
<td>Education</td>
<td>0</td>
<td>1</td>
<td>7</td>
</tr>
<tr>
<td>Medicine</td>
<td>0</td>
<td>3</td>
<td>14</td>
</tr>
<tr>
<td>Psychology</td>
<td>0</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>
Proportion of coders to authors. One interesting finding concerned the proportion of people who coded the meta-analyses to those that had their name on the paper as authors. As can be seen in Figure 4, the numbers are for the most part equal in all fields but Medicine. This information should not be taken to imply that the authors are necessarily the ones doing the coding, as this information was rarely reported. In Medicine, it is obvious that there are many more authors than coders. The reason for this may be that meta-analysts in Medicine are often doctors, and may be highly specialized. Having more than one author may allow a broader research perspective. Whatever the case, it is clearly more of a convention to have more authors, but the same two or three coders common to the other fields.
Note. Background shade indicates number of authors, line indicates number of coders.

Figure 5. Comparison of Number of Authors to Coders

Visualization of IRA, rigor, and time. Figure 6 illustrates several of the relationships measured in this study. Cohen’s Kappa (IRA), Percent Agreement (IRA), and Intraclass Correlations (IRR) are broken out to reflect how they were found in the sample literature. Field is represented by color, and Rigor is represented by the size of each circle, with a bigger circle representing higher rigor. Kappa and IRR were reported less often, which is reflected by less crowding of data points compared to percent agreement. Interestingly, it appeared that Medicine reported visibly lower kappa and IRR values than other fields, despite having higher rigor. This is reflected in the correlation between IRR and Medicine ($r$
= -.76, \( p < .01 \), but not in the correlation between IRA and Medicine due to the fact that percent agreement and Cohen’s kappa values were combined.

**Inclusion flow charts.** A reporting feature emerged in the Medicine literature towards the later years and seemed to become convention. As part of a trend of emphasizing the importance of inclusion criteria in primary research, meta-analysts started using inclusion flowcharts, an example of which is shown in Figure 7. These flowcharts provide an excellent framework for thinking about inclusion, and they provide a great deal of insight into the inclusion process.

**Journal impact factors.** JCR impact factor score was associated with Rigor \( (r = .34, p < .01) \), such that a higher factor score tended to be linked to a higher Rigor score. The scatterplot shown in Figure 8, however, indicates that there are group differences in impact factor score. An ANOVA confirms that JCR differs by field \( F(3, 136) = 14.147, p < .01 \) The Bonferroni analysis showed that Medicine was significantly higher than Computer Science \( (p < .01) \), Education, \( (p < .01) \), and Psychology \( (p = .02) \). Computer Science and Education did not differ from one another, but Psychology was significantly higher than both \( (p = .01; p = .01) \).

**Discussion**

Probably the most surprising aspect of this study was the difficulty of finding IRA and IRR data. The frequency represented in this study is not representative of the population of meta-analyses, at least for the fields as they were operationalized. It often took 15-20 meta-analyses to find one with usable data despite emphasis on coding and IRA and IRR in textbooks. Yet, even in the field of Medicine, such information was omitted or were instead
given for study inclusion. Another issue with IRA was the lack of variance. With the exception of two outliers, a 25% in Psychology, and a 28% in Education, the vast majority of IRA percentages were high, between 89 and 100% in most cases. In addition, there was no relationship between IRA and Number of Total Variables ($r = -.13, p = .31$) or whether the variables coded were low or high inference ($r = -.00, p = .99$). This is surprising because agreement is very difficult to achieve, especially when there are more coders, many variables, and data are complex. This is probably one of the reasons for the lack of correlation between IRA and Rigor, because the two seem to be aligned closely in concept. Whereas IRA was initially expected to have a range of 30-40 percentage points, it was actually only seven for the majority of cases. It is possible that IRA and IRR are undervalued in importance, and due to the strict process requirements for true IRA measurement, they are reported more loosely than other meta-analytic features.
Figure 6. Kappa, Percent Agreement, and ICC by Rigor and Time
Figure 7. Example Inclusion Flow Chart
Rigor was associated with Number of Coders and Number of Variables as well as Meta Variables and Impact Factor Score. It was also significantly higher in Medicine than the other Fields. The measure created by Oxman et al. (1994) provided data that seems valid and has sufficient variance to be meaningful for analysis. However, certain items that comprise the scale have become less relevant over time, and probably reduce the reliability of the scale in later years. For example, contacting original authors of primary research for additional studies is a good strategy, but it was used less and less in later years. This seems more an effect of the ubiquity of the internet for searching and the improvements in databases and search features that reduce the need for personal communication.
Takeaways from Computer Science

Meta-analyses from the field of Computer Science should be evaluated with care. The presence of false meta-analyses is a concerning trend because they may be seen as more accurate and more valuable than they are. IRA was significantly higher in Computer Science than in the other fields but Rigor was lower. This calls into question the procedures that contributed to calculating IRA, and devalues the importance of IRA as a proxy for validity.

Takeaways from Education

Education seems the most similar to Psychology. Aside from difference in number of published meta-analyses, there seems to be few substantive differences between them. There are few differences in style or format. There may be sufficient interdisciplinary work between Education and Psychology to have homogenized meta-analysis methodologies.

Takeaways from Medicine

Medicine was quite different in fundamental ways from the other fields. Meta-analyses were shorter, but also had higher Rigor and more Quality measures. More people were generally involved in them and they developed some conventions that are as yet not seen in the other fields, for example inclusion flow charts and IRA values for inclusion instead of coding. Given the huge volume of meta-analyses in Medicine relative to other fields, it may be that processes in Medicine are more refined, and that there have been consequences and findings that have influenced their evolution.
Takeaways from Psychology

Meta-analyses in Psychology were generally rigorous and reported IRA or IRR most frequently. However, Psychology was slow to implement even those methods such as funnel plots that are known to provide additional support for valid results.

Recommendations for Future Meta-Analyses

First, there should a more established methodology for how to measure IRA. The construct it represents is essential to determining reliability and validity in the coding process, and there are no other metrics that provide that insight. Emphasis should be placed on the value of true disagreement, not on the benefit of being as close to 100% as possible. It should also be reported explicitly how ties are broken.

Second, it is worth considering the value of quality measures for Psychology meta-analyses. These measures often require a great deal of additional work; in the Medicine literature it was not uncommon to have far more quality variables than variables from within the study. There is plenty of literature in support of using weighting measures to ensure that poor studies do not overly bias results (e.g., Kahn, Daya & Jadad, 1996) and it is a recommendation of the Cochrane Collaboration (Higgins, 2011c). There have been some arguments in the Medicine literature regarding the suitability of weighting by quality. Jüni, Witschi, Bloch, and Egger (1999) note that results can be heavily influenced by using different quality scales, even though they purport to measure the same underlying construct. Similarly, Greenland and O’Rourke (2001) found no systematic differences in meta-analytic results based on quality of the study, and suggested that hierarchical meta-regression might be a better solution. At the very least, they conclude that there is no such thing as a universal
quality measure for all applications. All of these questions bear researching in order to improve meta-analyses in Psychology.

Llewellyn, Whittington, Stewart, Higgins, and Meader (2015) demonstrate one possibility for the future of quality measures. The Semi-Automated Quality Assessment Tool (SAQAT) method they recommend reduces required effort dramatically while creating a format that facilitates IRA. Because assessing the quality of primary studies is so labor intensive, any automation would only encourage such practices. Even if it were not ultimately used to weight primary studies, it might provide additional discussion points for study inclusion. Results might also be reported for the discretion of the reader, with the caveat that the tool was semi-automated and thus should be interpreted with some caution. The SAQAT is specific to the healthcare industry, but there may be opportunity for Psychology to capitalize on a good idea.

Third, inclusion flow charts are an insightful means of communicating information about inclusion decision-making and process. Not only are readers informed about the greater population of available research, they are informed about the major decision points at each layer of refinement. Inclusion is just as important in Psychology as it is in Medicine for establishing validity, and yet there is less information disclosed about how it was done. Not only do flow charts provide information, they provide a guideline for how meta-analysts should be thinking about meta-analysis.

Limitations

This was a very broad study, and depth was sacrificed for breadth. More meta-analyses in a shorter range of time would provide more nuanced information about meta-
analysis. It might be useful to do a meta-meta-analysis such as this one within a specific topic of interest in order to better see the effects of time without the contamination of different styles and disciplines. In addition, there was some overlap amongst fields. Although efforts were made to keep the fields relatively “pure,” the overlap was fairly subjective and done from the point of view of an I-O Psychologist.
References


Sample Meta-Analyses


memory: five years at HICSS. *Proceedings of the 36th Annual Hawaii International Conference on System Sciences*, (pp. 1-9). Washington, DC: IEEE.


Proceedings of the 48th Annual Hawaii International Conference on System Sciences, (pp. 4483-4492). Washington, DC, IEEE.


Yaghoobi, M., Mayrand, S., Martel, M., Roshan-Afshar, I., Bijarchi, R., & Barkun, A. (2013). Laparoscopic Heller's myotomy versus pneumatic dilation in the treatment of...


Appendix A

Operationalization of Oxman, Cook, and Guyatt’s (1994) Guide to Discerning Validity

Did the overview address a focused clinical question?
   1. Can the sample and at least one treatment and outcome be identified from the title
      and/or abstract?

Were the criteria used to select articles for inclusion appropriate?
   1. Are the sample criteria identified?
   2. Are the treatment criteria identified?
   3. Are the outcome criteria identified?
   4. Are the methodological standards (randomized, controlled, etc) identified?

Is it unlikely that important, relevant studies were missed?
   1. Are the databases searched identified?
   2. Is there evidence of communication with experts?

Was the validity of the included studies appraised?
   1. Are validity criteria supplied?

Were assessments of studies reproducible?
   1. Were judgment calls made by more than one person?

Were the results similar from study to study?
   1. Was a test of homogeneity reported?
Appendix B

Search Terms for Meta-Analysis Availability

<table>
<thead>
<tr>
<th>Field</th>
<th>Search Terms</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comp Sci</td>
<td>SU=Computer Science AND TI=&quot;META-ANALYSIS&quot; NOT TS=medical NOT TS=educ* NOT TS=student*</td>
</tr>
<tr>
<td></td>
<td>Analysis: [excluding]: WEB OF SCIENCE CATEGORIES:</td>
</tr>
<tr>
<td></td>
<td>(ENGINEERING BIOMEDICAL OR BIOLOGY OR BIOTECHNOLOGY APPLIED MICROBIOLOGY OR BIOCHEMICAL RESEARCH</td>
</tr>
<tr>
<td></td>
<td>METHODS OR EDUCATION SCIENTIFIC DISCIPLINES OR EDUCATION EDUCATIONAL RESEARCH OR BEHAVIORAL SCIENCES)</td>
</tr>
<tr>
<td>Education</td>
<td>SU=Educ* AND TI=&quot;meta-analysis&quot; NOT TS=&quot;computer science&quot; NOT TS=psych* NOT TS=medical</td>
</tr>
<tr>
<td></td>
<td>Analysis: [excluding]: WEB OF SCIENCE CATEGORIES:</td>
</tr>
<tr>
<td></td>
<td>(EDUCATION SPECIAL OR COMPUTER SCIENCE INTERDISCIPLINARY APPLICATIONS OR REHABILITATION OR PEDIATRICS OR</td>
</tr>
<tr>
<td></td>
<td>PSYCHOLOGY DEVELOPMENTAL OR PSYCHIATRY OR PSYCHOLOGY CLINICAL)</td>
</tr>
<tr>
<td>Medicine</td>
<td>(TI=&quot;meta-analysis&quot; NOT TS=psych* NOT TS=comput* NOT TS=edu* NOT LANGUAGE: (English))</td>
</tr>
<tr>
<td></td>
<td>Analysis: [excluding]: MeSH QUALIFIERS: (ORGANIZATION ADMINISTRATION OR ANALYSIS OR STANDARDS OR NURSING</td>
</tr>
<tr>
<td></td>
<td>OR LEGISLATION JURISPRUDENCE OR INSTRUMENTATION OR HISTORY OR STATISTICS NUMERICAL DATA OR VETERINARY</td>
</tr>
<tr>
<td></td>
<td>OR MANPOWER)</td>
</tr>
<tr>
<td>Psychology</td>
<td>SU=Psychology AND TI=&quot;meta-analysis&quot; NOT TS=psychiat* NOT TS=edu* NOT TS=medic*</td>
</tr>
<tr>
<td></td>
<td>Analysis: [excluding]: WEB OF SCIENCE CATEGORIES:</td>
</tr>
<tr>
<td></td>
<td>(GERIATRICS GERONTOLOGY OR PEDIATRICS OR ONCOLOGY OR PSYCHIATRY OR GENETICS HEREDITY OR PHYSIOLOGY OR</td>
</tr>
<tr>
<td></td>
<td>LAW OR PHARMACOLOGY PHARMACY OR EDUCATION SPECIAL OR NUTRITION DIETETICS OR GERONTOLOGY OR ERGONOMICS</td>
</tr>
<tr>
<td></td>
<td>OR REHABILITATION OR CLINICAL NEUROLOGY OR PSYCHOLOGY BIOLOGICAL OR SOCIAL SCIENCES BIOMEDICAL OR</td>
</tr>
<tr>
<td></td>
<td>ZOOLOGY)</td>
</tr>
</tbody>
</table>