STUDYING REPLICATION BIASES IN DEVELOPMENTAL PSYCHOLOGY

ECOLE NORMALE SUPÉRIEURE
M2 Cogmaster – Master's thesis:
Patricia Mirabile
Advisors:
Brent Strickland & Paul Égré

Date :
31st August 2016
(September session)

Language:
English
Word count:
13 273 words

TABLE OF CONTENTS

Declaration of originality

The present work applies in a novel way the distinction between *proximate* and *ultimate causes* to the topic of causes of non-replicability. Its main originality rests in the use of meta-analytic methods to investigate empirically and quantitatively the claims made by the developed theory of causes of non-replicability. It makes use of the distinction between two types of causes of non-replicability at the experimental level (*sources of error* and *sources of bias*) to identify possible methodological issues in experiments in the field of developmental psychology. It presents a set of 10 new meta-analyses in the domain of developmental psychology (using strict criteria of similarity) and a novel meta-analytical work that combines the result of those 10 different literature searches into one unique data set, in order to investigate the impact of the identified methodological factors on the effect sizes and pooled standard deviations of those experiments.

Declaration of contribution

This work was made possible by the contributions of Pr. Brent Strickland, Pr. Paul Égré, Dr. Christina Bergmann, Pr. Alex Cristia and Aurélien Allard. Pr. Brent Strickland and Pr. Paul Égré contributed to the definition of the research project, and Pr. Brent Strickland played a key role in the development of the theory of causes of non-replicability. The literature review and the coding of the R script were the result of my own work, but Pr. Alex Cristia and Dr. Christina Bergmann provided crucial advice on the methods of literature review, meta-analysis and on the development of the R script. Pr. Brent Strickland, Pr. Paul Égré and Dr. Christina Bergmann contributed to the choice of the analytical methods and gave very insightful input on the analyses and the interpretation of the results. Aurélian Allard shared some feedback following a short presentation of my work at the Institut Jean Nicod. Pr. Brent Strickland and Pr. Paul Égré gave me much feedback on my thesis, including theoretical notes, style and presentation notes.

Thanks

I am very grateful for all the help of the listed contributors, who guided and encouraged me all throughout this master's thesis. I also wish to thank all the authors who kindly responded to my queries for supplementary information. On a more personal level, I am sincerely thankful to all the people who surrounded me during the year with their material and emotional support, and to myself for finding the motivation to always learn new things and for bringing this project to completion.

Additional Material

All the R-code, as well as the data collection spreadsheets are available online and publicly on the following Git Hub Repository : https://github.com/PLMir/replication-biases

I. INTRODUCTION

The replication of experimental results has been described as one of the "hallmarks" (Drummond, 2007) or "cornerstones" (Simons, 2014) of science. Replicating a study entails the reproduction of its experimental paradigm with the aim of testing the veracity of a study's empirical claims (in the case of strict replications), and of its theoretical claims (in the case of conceptual replications). Recently, especially since the publication in August 2015 of the "Republication Project: Psychology" by the Open Science Framework project (OSF, 2015), which presented a rate of 61% of failure when attempting to replicate a 100 studies published in three major academic journals, the replicability rate of psychology experiments has been called into question. Weaknesses both at the individual and at the institutional level have been identified as possible causes of this situation: in particular, rationality biases and socio-economical characteristics of the academic world might have led to what is often called a "replication crisis" in psychology and amongst the solutions that have been suggested, the most frequent one is a call for more replications (for instance, the editorial of the 2010 issue of the *Journal of Pediatric Psychology* is titled "A Call for Replications of Research in Pediatric Psychology and Guidance for Authors"). However, if efforts are going to be made to generalize the practice of replicating studies and experiments, then it seems necessary to inquire whether causes of non-replicability might also be identified at the level of the experiments themselves. Indeed, there might be features of experimental paradigms which render them more susceptible to bias or make them more error-prone. The main focus of this essay will be to develop a novel theory of causes of non-replicability at the experimental level, and of the effects those causes have on replication attempts, and to present the results of a systematic review in developmental psychology which attempted to test the empirical predictions outlined by the aforementioned theory.

The first half has a more theoretical nature: it lays out how the practice of replication fits into a normative model of science where scientific fields strive to achieve high-levels of credibility and reliability by submitting scientific facts to processes that promote the self-correction of errors and falsehoods. It then delves into the practical limitations of this model once the cognitive and socio-economical realities of the

scientific world are taken into account, and advances a theory about two different types of causes of non-replicability that might affect experiments: causes that are sources of bias, and causes that are sources of error, and about how their influence might be empirically recognized. The second half proposes an application of that theory to the domain of developmental psychology. Three methodological factors were identified either as possible sources of bias or as possible sources of error, and a meta-analysis was conducted on a set of 185 studies, corresponding to ten different experimental paradigms, to investigate whether those methodological factors were at the source of variations in the effect sizes or in the precision of the studies.

The normal functioning of science

The scientific method can be described as a set of practices that allow the production of facts that may reasonably be believed to be true. The truth of a scientific fact rests both on its correspondence with the states of the world, and on the procedures that were used to obtain it: if a fact is true but was not established by the use of scientific standard practices, then it is not a scientific fact. These procedures entail at the very least the observation of empirical evidence and the elaboration of theories and hypotheses, which serve as attempts to organize and explain the evidence, and which are then put to the test by empirical experiments. In addition, the scientists who are enacting this practical process are also expected to be guided by a certain list of values, or even virtues, such as rigor, precision, objectivity, honesty and a deep-seated desire for finding and communicating the truth (Kühn, 1977).

This general conception of science, which is taken to refer both to the whole set of scientific facts and methods, and to the group of scientists and institutions that enable, structure and regulate the production of those facts, has come under important theoretical criticism since the 20th century. Karl Popper called into question the assumption that the scientific method made the discovery of true propositions possible (Popper, 1959). In fact, because scientific theories take the form of universal, or general truths, they must follow the logical rules of induction, and as such, they can never be confirmed by empirical evidence but only be refuted when a contradictory event is encountered. For instance, to confirm the proposition that “all crows are black”, I would need to verify that all existing crows, past and future, are indeed black, whereas one simple instance of a white crow is enough to disprove it. Popper’s position entails that science cannot solely

focus on the production of new facts, but rather it must always adopt a critical mindset towards already existing scientific facts. Those facts must be submitted to new empirical tests, i.e. to replication and to falsification attempts, and successfully passing those tests can admittedly increase confidence in them, but it is still not enough to fully verify them. This entails that the reliability of scientific facts does not rest on their truth (although they must not be obviously un-true), since they can never actually be proven to be true, but on the rigor of the methods used to produce them and on the intensity of the attempts made to disprove them. It also entails that the capacity to admit that a sometimes long favored hypothesis is in fact wrong must be added to the list of scientific abilities. When a scientific field follows standard scientific methods, when it submits its facts to rigorous testing and when it readily discards facts once they are discovered to be false, then it should enjoy a high level of credibility, and it can also be described as a self-correcting scientific field. This definition of scientific credibility differs from the one developed in "Why Science Is Not Necessarily Self-Correcting" (2012), where Iaonnidis defines scientific credibility as "the proportion of scientific facts that are correct": since it has been argued that the correctness of a fact cannot be assessed, determining such a proportion is not a possible endeavor. Instead, when the credibility of a scientific field is tied with the presence of a "self-correcting mindset" in the actors of that field, then the reliability of a fact can be evaluated without needing to refer to its correctness. In other words, a scientific fact will be reliable because it will describe reliably empirical events but also because it will be possible to assume that the scientific community submits facts to strenuous testing and to assume that if the fact was known to be wrong, or to have been produced by unreliable methods, then a record of its incorrectness would be available.

More recently, these conceptions of the scientific method have come under the fire of another type of criticism. Indeed, it has been argued that while such frameworks might be normatively accurate (they state how science should work), they are descriptively inaccurate. In particular, a number of biases, both at the individual and at the institutional level, have been shown to negatively affect the production of scientific facts. A bias is a structure (either mental or institutional) which distorts a process in such a way that its results are skewed, irrational, inefficient or, sometimes, ethically wrong. Another common and important characteristic of a bias is that its existence goes mostly undetected and that it often comes with the illusion of objectivity. For instance, a gender bias might affect the recruiting process of a firm so that instead of

comparing the abilities of applicants regardless of their gender, men's applications will be unduly favored and women's applications will be unduly undervalued. However, the recruiters might not be aware of their own gender bias and they will then believe that they are simply selecting the best applicant, a belief which can often be disproven when gender-blind recruiting procedures (such as anonymized applications) are used instead.

A rationality bias is a specific type of bias, which renders a piece of reasoning irrational: for instance, when assessing the truth of a position, it would be rational to take into account both arguments in favor and against that position. However, research has shown (Wason, 1968; Nickerson, 1998) that human beings tend to exhibit a confirmation bias, such that they will grant more weight to arguments that confirm their own favored position than to arguments that challenge it. More generally, the term of confirmation bias has been used to refer to a group of mental structures that tend to render human reasoning about evidence and about opinions irrational because they will favor, without realizing it, the information that allows them to uphold the opinion or the thesis that they have come to believe. This might affect activities such as searching for evidence, selecting what is deemed as relevant evidence, producing empirical data, interpreting that evidence, reasoning, producing arguments and understanding and responding to counter-arguments.

Two biases are particularly relevant to the scientific domain: the *experimenter bias* and the *publication bias*. The experimenter bias (Rosenthal, 1968; Strickland & Suben, 2012) is a type of confirmation bias that affects more specifically researchers and the production of experimental data. It can lead scientists to unconsciously influence the participants or the data collection process so that the outcomes of the experiment will match their favored hypothesis. The publication bias is a bias that appears at the level of the institutions that are in charge of the publication and diffusion of scientific results, and it leads them to favor certain types of experiments and certain types of results. In particular, there is an industry of academic publications (it should be remembered that the main academic publishers are, first and foremost, profit-driven companies) which encourages the publication of novel experiments with positive outcomes, or at least of experiments with significant results, because they are considered to be more interesting or valuable, and might as such improve or maintain the perceived standing of an academic journal. As a consequence, papers presenting replication attempts and papers presenting non-significative results will be more difficult to publish (Rosenthal, 1979), specially in top-tier journals, which in turn tends to drive scientists away from this

important aspect of scientific work. Indeed, there are real social and economical stakes involved in the publication of scientific articles for the individual scientist: impacts on future career opportunities, on reputation within a field, on access to grants and funding for future research are all concrete considerations which need to be taken into account when choosing a research project. Failing to publish any novel or significant results when time and resources were invested in a research project, or publishing too many null-results might be costly for a researcher's status within his or her field, so much so that the phrase "publish or perish" has been coined. The main consequence of the publication bias is that scientific results available to the public tend to present skewed evidence: replication attempts, in particular failed replication attempts, and experiments which produced null results are not published and the accessible evidence is mostly confirmatory. Moreover, publication bias and experimenter bias might both influence simultaneously a researcher so that he or she might unconsciously take into account the potential to publish a paper when evaluating possible theories or hypotheses.

The context of the replication crisis

Those worries about the role of biases in the scientific domain must be taken seriously because they can have major consequences on the reliability of scientific facts, and on the credibility of science. Standard scientific methods might not be sufficient to ensure the objectivity of experimental results and the expectation that replication and falsification attempts are being pursued and published diligently might not be met. This criticism has come to a head in the field of psychology with the identification of what has been called a "replication crisis". This replication crisis has two important aspects: the first one is the discovery that a high ratio of published results in the field are not replicable and the second one is the realization that many scientists might have been using questionable research practices in their experimental projects (John, 2012).

A research practice is questionable when, without being fraudulent, it allows researchers to (perhaps unconsciously) exploit weaknesses in experimental protocols and in statistical methods in order to obtain certain results. For instance, a researcher might use optional stopping, i.e. collecting data until the statistical tests produce significant results, or selectively including or excluding outliers, or run unplanned analyses on the data until a significant result is found, all of which reduces the statistical power of the

analysis, and thus increases the risk of finding false positives (i.e. erroneously producing positive test outcomes). Other questionable research practices also include (John, 2012) selective reporting of dependent measures, "rounding off" of p values, and post-hoc story-telling (presenting an unexpected finding as having been predicted from the start). Questionable research practices can in other terms be described as a set of practices that facilitate the expression of confirmation bias, and more specifically of experimenter bias.

Coming back to the first aspect of the "replication crisis" in psychology, the field of psychology was put to the test by the Open Science Framework, a collective of researchers who lead a collaborative project, known as the "Reproducibility Project: Psychology" (RP:P), whose aim was to assess the reproducibility of published results in psychology. Replication attempts on a sample of 100 studies, following a pre-registered protocol aiming to produce high-quality replications (OSF, 2015: "Estimating the reproducibility of psychological science"), were conducted by a group of 270 contributors. The results were published in August 2015: "39% of effects were subjectively rated to have replicated the original result" (OSF, 2015). In other words, replication attempts failed approximately six times out of ten, casting important doubts on the correctness of the original studies, and on the correctness of published results in psychology as a whole. It should be noted that in themselves, replication failures are a normal part of the scientific process, and the discovery that published results do not replicate is not enough to conclude to the existence of a crisis. However, it is the rate of replication failure that was worrying, because it suggested both that false positives in the field were more common than expected (normal statistical analyses are expected to produce a rate of only 5% of false positives) and that there was a problem at the level of the replication process: replication of published studies might not be systematic enough, or at least the results of those replications might not be published.

Both individual problems, such as the questionable research practices identified earlier, and institutional problems, such as the publication bias, the lack of incentives to conduct replication attempts and weaknesses in the statistical training of psychologists, have been identified as possible causes for the replicability crisis in psychology. Efforts have also been made to incentivize replication attempts, for instance *Perspectives on Psychological Science* has opened a specific section for replication reports and

repositories for null results have been created[1]. New methodological practices, such as pre-registering analyses and making data publicly available, are also becoming more wide-spread. However, this new focus on the necessity of conducting replications also calls for a more careful consideration of what a replication is and of what determines its quality.

II. A THEORETICAL FRAMEWORK TO THINK ABOUT REPLICATION

A replication is the reproduction of the experimental paradigm described in a study, in order to assess whether the outcomes of the original study can be verified or falsified. A distinction can be introduced between two types of replications : *direct replications* and *conceptual replications*. Direct replications strive to be as similar as possible to the original paradigm, they will use similar experimental manipulations, sample from a similar population (for instance using the same age group), and will generally attempt to control for any variation with the original study. Conceptual replications, on the other hand, will introduce variations into the experimental paradigm, generally in order to test whether the effect can be generalized to other populations or to other situations in that way that a given theory predicts. For instance, a conceptual replication of the priming effect might use of population of children instead of adults, or it might attempt to replicate a visual priming effect with an auditory priming effect. In the current work, we will focus on the implications of conducting direct replications but with one important limitation: direct replications generally require intent. The attempt to directly replicate an experiment should be planned as such in order to increase the similarity with the original, for instance the authors of the original paper might be contacted to clarify ambiguities in the protocol. Here, however, this notion of intent was not taken into account when identifying replications and only the similarity between the original study and its replication was considered. Strictly speaking, some of those replications might be better defined as conceptual replications since some of the variations they introduced could have been meant to investigate about the strength of the theory rather than the reliability of the original experimental results. This corresponds to an important feature of the distinction between direct and conceptual replications: it is inscribed on a continuum rather than being a strict dichotomy.

1 http://www.psychfiledrawer.org

The notion of replication raises two important difficulties. First, direct replications can never be perfect replications: different research team, different laboratories and different populations will necessarily introduce variations in the experimental paradigms, which might drive the effect of a manipulation in unexpected, and unknown ways. Secondly, a replication attempt might fall prey to the same shortcomings as the original study, for instance by using biased experimental methods that increase artificially the effect size, or that produce a higher rate of false positives. In other words, there might be specific causes at the level of the experimental paradigms that lead to the non-replicability of scientific results. It is the identification and the investigation of those specific causes of non replicability that is the focus of the rest of this essay.

Introducing a theory of causes of non-replicability

In evolutionary biology, a distinction (which can be traced back as far as Plato's dialog *Phaedo*) is often established between *proximate causes* and *ultimate causes* (E. Mayr, 1961). A proximate cause has a direct effect on an event whereas an ultimate cause has an indirect effect, which is generally mediated by a proximate cause. For example, the proximate cause of a death might be that a bullet penetrated the victim's heart, interrupting its normal functioning, whereas the ultimate cause would be the murderous intent of the person firing the gun. In this case, it seems obvious that the intent to kill (ultimate cause) can have no direct effect since a weapon (proximate cause) is needed to assassinate somebody: a proximate cause must be present in order for the ultimate cause to have an effect. We will apply this distinction to the topic of causes of non-replication[2].

As explained earlier, one difficulty of direct replications is that they can never be truly identical to the replicated study : the imitation of an original always introduces some level of variation and reproducing the experiment in another laboratory, with a different research team, sometimes in another country can change slightly the conditions of the experiment. In the case of psychology, this is complicated by the fact that psychological experiments can never be run in perfectly controlled environments,[3] since factors such as, for instance, the weather conditions, the political climate or the personal history of the participants, might influence subtly their states of mind and their reactions in the study. Another aspect is that when replicating

2 The use of this distinction was strongly inspired by B. Strickland, and the manuscript of a research proposal he submitted to the CNRS, titled "Research Proposal 53: Improving expert performance: An interdisciplinary approach to the search for truth".

3 A fact which was already discussed by J.S. Mill, 1843, in the *Logic of the Moral Sciences,* Chapter VII, §2.

an experiment, the original study will constitute a base of beliefs, for instance about which outcomes to expect, beliefs which might then create confirmation bias in replication attempts. Expecting a study to replicate, or even being disinclined to fail to replicate some major result in a field, because it would threaten one's world view or because it might harm some professional relationships with other researchers[4] could for instance contribute to the effects of the confirmation bias. In other words, variability is inherent to the practice of replication. Here we propose to identify some proximate sources of this variability, and we will distinguish these from ultimate causes that (often unconsciously) capitalize on this variability in order to match the outcomes of an experiment with the expectations of the researcher (and of the scientific community). In the case of causes of non-replicability, proximate causes in themselves are sources of error — they introduce more variability in the results — and ultimate causes are sources of bias, thus creating deviation from the truth in specific directions.

Sources of bias and *sources of error* are two of the main types of causes that might affect the replicability of results. In an experiment, a methodological factor is a source of error when it increases the amount of variability of the study, and in particular when it makes measurements less precise. An example of this is when a manually-activated stopwatch is used to time an event (like a race) instead of a system that is activated by lasers, which automatically detect when the event starts and ends. A manually-activated stopwatch will be less precise because it depends on highly variable factors such as the concentration level of the experimenters and their interpretation of what counts exactly as the start and the end of the measured event. In some cases, a source of error becomes a source of bias because the variability it introduces can be harnessed to produce data that confirms the expected outcome of the experiment. A classical example of a source of bias is the failure to use a double-blind procedure, i.e. a procedure where experimenters interacting with the subjects are left unaware of the experimental condition in which those subjects have been placed. For instance, in a medical trial, experimenters will not know whether a patient is being given the actual drug or simply a placebo because this might introduce variability in how the experimenters interact which the patients, and thus bias the results

Contrarily to the way the distinction is used in evolutionary biology, where a connected proximate

4 Bargh's reaction to replication failures of his priming experiments in "Priming Effects replicate just fine, thanks", a blog post originally titled "Nothing in their heads", is a striking example of this risk.

and ultimate cause produce the same effect (in the example given earlier, they both result in the death of the victim), in this case proximate causes should produce different effects when they are under the influence of an ultimate cause, *i.e* proximate and ultimate causes will not have the same effects. Sources of error increase the variability, or in other words, they increase the amount of noise in the data, which is a random effect. Sources of bias, on the other hand, have a directional effect: they should drive the results of the experiment towards what the experimenter expects or desires them to be. This will have the following consequences on a scientific field: if sources of error are prevalent, then the discovery of true effects might be more difficult because of the higher variability in experimental results. If sources of bias are prevalent (in addition to the proximate causes that are necessary to allow them to become expressed), however, then the amount of false positives might increase, since most results will seem to confirm the most popular theories and the major findings of the field.

On the basis of this distinction between sources of bias and sources of error, we can formulate the following predictions:

- if a methodological factor is merely a source of error, then it will increase the amount of variability in an experiment. This variability can for instance be measured by using the pooled standard deviation of the study, which is an indicator of the precision of the measurements ;

– if a methodological factor is a source of bias, then it will increase the effect size of the study (so long as proximate mechanisms are available). The effect size is an indicator of the efficacy of an experimental manipulation, and statistically significant, high, effect sizes are often interpreted as signifying that the hypothesis has been confirmed and that the replication is a success.

III. APPLYING THE THEORY TO THE DOMAIN OF DEVELOPMENTAL PSYCHOLOGY

General methodology

To test those hypotheses, the protocol for a systematic literature review was developed. The goal was to investigate whether the presence, or the absence, of a given methodological factor, identified either as a source of error or as a source of bias, would influence the results of an experiment, and more generally whether it would influence its likelihood of replicating by increasing either variability or effect sizes. In order to do this, we would select a set of original experiments within a specific field and then perform a systematic review, following PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analysis) guidelines, and collect data about direct replications of those original experiments. We would then use meta-analytic statistical techniques in order to evaluate whether effect sizes were higher when a source of bias was present, and simple statistical techniques to evaluate whether pooled standard deviations were higher when a source of error was present.

We selected the field of developmental psychology to conduct this empirical application of our theory of causes of non-replication. Two reasons motivated this decision. First, developmental psychology is a subfield of psychology that is very definite: it has a specific subject population (infants and children) which requires specific types of experimental techniques, with specialized researchers research teams and laboratories, and its own specialized academic journals. This would ensure that the literature review could be conducted on experiments with common standard practices and often similar methods.

A second motivation was the publication of an article, titled “The Baby Factory: Difficult Research Objects, Disciplinary Standards and the Production of Statistical Significance”, (David Peterson, 2016), which presented the results of an ethnographic study. This study was conducted in three laboratories that perform experiments in developmental psychology. The main thesis of this article is that the natural variability of experimental objects, which scientists generally strive to control in their experiments, is “untamable” in developmental psychology because of its focus on infants. Indeed, the manipulations used to

control for variability in inanimate objects, or even in animals, cannot be generalized to infants, since they are protected by ethical and legal considerations, and infants are obviously unable to understand and follow instructions in the same way as adults. This "untamable variability" renders the production of rigorous empirical data, and of statistically significant results, very difficult, a difficulty which is exacerbated by the important costs in time and resources necessary to recruit infant participants and to run the experiments. Peterson claims that as a result the field of developmental psychology has developed a set of four "strategies" that help the "production of studies with statistical significance", in other words those strategies can be described as questionable research practices that enable researchers to obtain positive results but also drastically increase the risk of those results being false positives. To illustrate those strategies, Peterson draws from the observations he made during his visits, either as an observer or as a laboratory assistant, to three different laboratories. Let us focus on the first strategy identified by Peterson: protocol flexibility. This refers to a set of behaviors that bend the pre-defined protocol of the experiment in order to diminish loss of data. For instance, the data from a participant might be included in the final sample even if the experimenter committed a mistake when presenting the stimuli. Protocol flexibility is described as diminishing the quality of the experimental data and as having a strong potential for biasing it. From the example of protocol violations that Peterson relates, it seems that protocol flexibility can concern three main aspects of an experiment: the decision to include or to exclude the data of a participant, the production of a controlled experimental setup and of experimental stimuli, and the interpretation of the infants' reactions during the coding process.

Peterson's study is a qualitative investigation of how questionable research practices might influence the production of unreliable scientific results in the field of developmental psychology. Because of its ethnographic nature, it focuses on a limited number of observations and is liable to cherry-picking: with this method it is not possible to objectively assess how common the described violations are in the everyday practice of those laboratories, and of laboratories in general. Besides, the thesis that those strategies enable the production of significant results, although compelling, is not directly proven. In particular, the article does not offer any data to back the claim that protocol flexibility is correlated with the generation of confirmatory results or of statistically significant outcomes. In some respects, the study presented (below) in this essay can also be understood as an attempt to conduct a quantitative investigation of the claim that

protocol flexibility is one of the strategies used by developmental psychologists to obtain a higher rate of statistically significant results.

Here identified three methodological factors that might affect the replicability of developmental psychology experiments: the way the stimuli were presented to the infants, the way the infants' reactions were coded and the ratio of infants that were included in the final sample, compared to the total number of infants that participated in the experiment. In most developmental experiments, stimuli are either presented manually (for instance, the infant will be shown a puppet show) or automatically (the infant will be show a video). Manual presentation seemed to allow for more flexibility in the protocol, and we predicted that this mode of presentation would be a source of bias and error because the direct interaction between the experimenter and the participant might give rise to expectancy effects. Essentially, the experimenter may move or present the stimuli in a way that makes it more likely that they will obtain an expected result. We predicted that manual presentation of stimuli would produce higher effect sizes, compared to cases where stimuli had been produced automatically.

During the coding process, two methods are generally used in developmental experiments: the coding will either be done live (also known as "online coding") or it will be done on a video recording of the infants' reactions. Online coding is often done by watching the infants' faces through a small hole in the display and by pressing a button to record the infants' looking-times. When coding is done on a recording, it will often involve several coders, and tests will be run to check for inter-coder agreement. A third frequent option consists of a hybrid between online coding and coding on a recording: the data will be coded online, but a certain portion of the trials (generally between 25 and 75%) will be re-coded using a recording in order to check for inter-coder agreement. Because the conditions of online coding seem liable to decrease the precision of the measurement, we identified it as a source of error, but not as source of bias. We predicted that studies with online coding would produce more noise in the data than studies that were not coded online, and that as a consequence they would tend to have higher pooled standard deviations. However these should not produce bias because coders are generally blind to experimental condition.

Finally, the ratio of infants included in the final sample seemed like a possible indication of how flexible the protocol was when it came to deciding whether an infant's results should be excluded from the

final data. Studies where loss of data was very low could of course be interpreted as having been extremely lucky when recruiting participants, but it seemed also possible that flexibility in the protocol might have been used to prevent the exclusion of some of the tested infants. Because the collection of data in psychology is a costly process, it can be assumed that there is an effort to exclude as little data as possible, regardless of how that data might influence the final results. As such, we identified the ratio of inclusion as only a source of error and not of bias. We predicted that studies with a high ratio of inclusion would produce more noise in the data than studies that had a relatively lower ratio of inclusion, and that this would result in higher pooled standard deviations.

Literature search and data collection

In order to test those specific predictions, we selected eleven original "seed" experiments that we identified as presenting major results in the field of developmental psychology. The selection process took into account recommendations from specialists as well as the following criteria: the experiments needed to be run on healthy infants at the pre-verbal stage (which we defined as infants younger than 24 months) and they needed to collect behavioral measures (as opposed to collecting cerebral activity). We surmised that infants having reached the verbal stage might not present the same level of "untamable variability" as the one described by Peterson and we rejected measures of cerebral activity because they required specific technical knowledge that would make them difficult to compare with studies measuring behaviors such as looking time, head turning or grabbing of a puppet. A systematic literature review was conducted for each of those eleven seed experiments. For each seed experiment, we identified the original study and used the systematic review to find replications of that original study that were as similar as possible to it Because it would have limited the size of our sample too much, we included studies that were not strictly speaking direct replications, since they had not necessarily been planned as such and sometimes introduced variations in the experimental paradigm (at the very least at the level of the methodological factors we considered). The criteria used to select those replication attempts were nonetheless defined with the goal of minimizing that variability as much as possible and in particular studies that tested for additional factors (such as for instance longitudinal studies that tested for the evolution of an ability) were generally excluded from the data set. Out of those eleven seed experiments, one seed experiment had to be excluded because it was impossible to

compute effect sizes for most of the experiments (although means were reported, standard deviations or standard errors were not available) and because the studies were too old to make it possible to contact the authors: this was Fantz's 1961 experiment which compared looking times of infants who were shown pictures of human faces and pictures of scrambled faces (Fantz, 1961).

Following PRISMA guidelines, spreadsheets were built during the systematic reviews to record every step of the data collection process. The first step consisted in a systematic literature review. Once a seed experiment was identified, the website Web of Science,[5] was used to create a list of all the articles that cited that original study. This list was then submitted to several selection processes, and the criteria used to include or exclude a study were recorded as well. A first selection was operated on the base of the title of the article. If it was necessary to read the abstract, this was recorded, as well as the screening decision. Full text articles were then retrieved and their inclusion was decided based on their eligibility (whether they respected the constraints established to determine what qualified as a direction replication) and whether they reported data that could be use for the planned analyses.

The second step was to collected the data for the analyses. We used the spreadsheets provided by the MetaLab project, a collaborative project whose goal is to build meta-analysis on the domain of infant language acquisition[6] in order to ensure that we were collecting all the relevant data. We also coded whether each experiment had a manual or an automatic presentation of stimuli, how coding was conducted and the necessary information to compute the inclusion ratio. Effect sizes and pooled standard deviations were not computed until the full process of data collection was finished.

Here is a list and a description of the ten seed experiments that were included in the final analyses :

- **False Belief** seed experiment: "Do 15-Months-Old infants understand False Beliefs?" (Onishi, Baillargeon, 2005) and its replication attempts. This experiment is an adaptation of the Sally-Ann experiment to pre-verbal infants. It tests whether infants are able to predict how an actor will act on the basis of her true or false beliefs about a toy. The dependent measure is the infants looking time to the scene and uses a violation of expectation paradigm.

- **Number Discrimination** seed experiment: "Large number discrimination in 6-months old infants" (Xu, Spelke, 2000) and its replication attempts. This experiment tests whether infants are able to distinguish

5 https://apps.webofknowledge.com/
6 http://metalab.stanford.edu/

displays where different numbers of objets are represented visually (auditory or dynamic comparisons were excluded). Only experiments with a comparison ratio higher than 1:2 and with numbers higher than 3 were included. The dependent measure is the infants looking time to the different displays.

- **Word Segmentation** seed experiment with word first presentation : "Infants' detection of the sound patterns of words in fluent speech" (Jusczyk, Aslin, 1995) and its replication attempts. This experiment tests infants' abilities to recognize words they have been habituated when presented with longer passages of the same language containing those words. Only studies using natural languages were included. The time the infant turned its head towards the source of the auditory stimulus was the dependent measure, and a familiarization paradigm was used.

- **Word Segmentation** seed experiment with speech first presentation : "English-learning infants' representations of word forms with iambic stress" (Johnson, 2005) and its replication attempts. This experiment was identical to the precedent one except that the order of presentation was reversed: infants where habituated to passages of fluent speech and they were then tested on lists of words that had been used in those passages.

- **Head Turning to Scrambled Faces** seed experiment: "Visual following and pattern discrimination of face-like stimuli by newborn infants" (Goren, 1975) and its replication attempts. Infants are shown pictures of faces, either normal or scrambled, on a mobile display. The degree of head turn used by the infants to follow the mobile display is measured. Comparisons are made within subject (infants are shown successively normal and scrambled faces) and stimuli are presented one after the other (not concurrently).

- **Inferential Statistics** seed experiment with sample first presentation: "Intuitive statistics by 8-month-old infants" (Xu, Garcia, 2008) and its replication attempts. Infants are first shown a sample of colored balls and are then shown the population from which those balls were extracted. Their ability to detect incompatibilities between samples and populations is tested using a violation of expectations paradigm which measures the infants' looking times. Only studies with visual stimuli were included.

- **Inferential Statistics** seed experiment with population first presentation: also "Intuitive statistics by 8-month-old infants" (Xu, Garcia, 2008) and its replication attempts. This experiment is identical to the precedent expect that the population was shown during the familiarization and the sample was shown during the test.

- **Visual Statistics** seed experiment: “Visual statistical learning in infancy: evidence for a domain general learning mechanism.” (Kirkham, 2002) and its replication attempts. This experiment habituated infants to sequences of visual shapes where the pairings of shapes had a certain likelihood ratio and then compared looking times when infants were presented either with familiar or with novel pairs of shapes.

- **Intentional Action** seed experiment: “Taking the intentional stance at 12 months of age.” (Gergely, 1995) and its replication attempts. This experiment investigated whether infants where able to attribute intentional actions to agents. Infants reactions when an agent performed an irrational action instead of the expected rational action (for instance taking the longer path to reach a goal although an obstacle has been removed) were compared using measurements of looking times.

- **Grabbing of Pro-Social Actors** seed experiment: “Social evaluation by preverbal infants.” (Hamlin, Wynn, 2008) and its replication attempts. Infants were introduced to actors who presented either pro-social or not-prosocial behavior. Their preference for pro-social actors was measured depending on the puppet (representing the actor) the infants grabbed (it was assumed that they would grab the puppet of the actor they preferred).

We include a PRISMA flowchart that represents a summary of the ten literature searches. As such, it might contain duplicates for the citations that appeared in different literature searches. The list of the included studies is also available in the second section of the bibliography at the end of this essay.

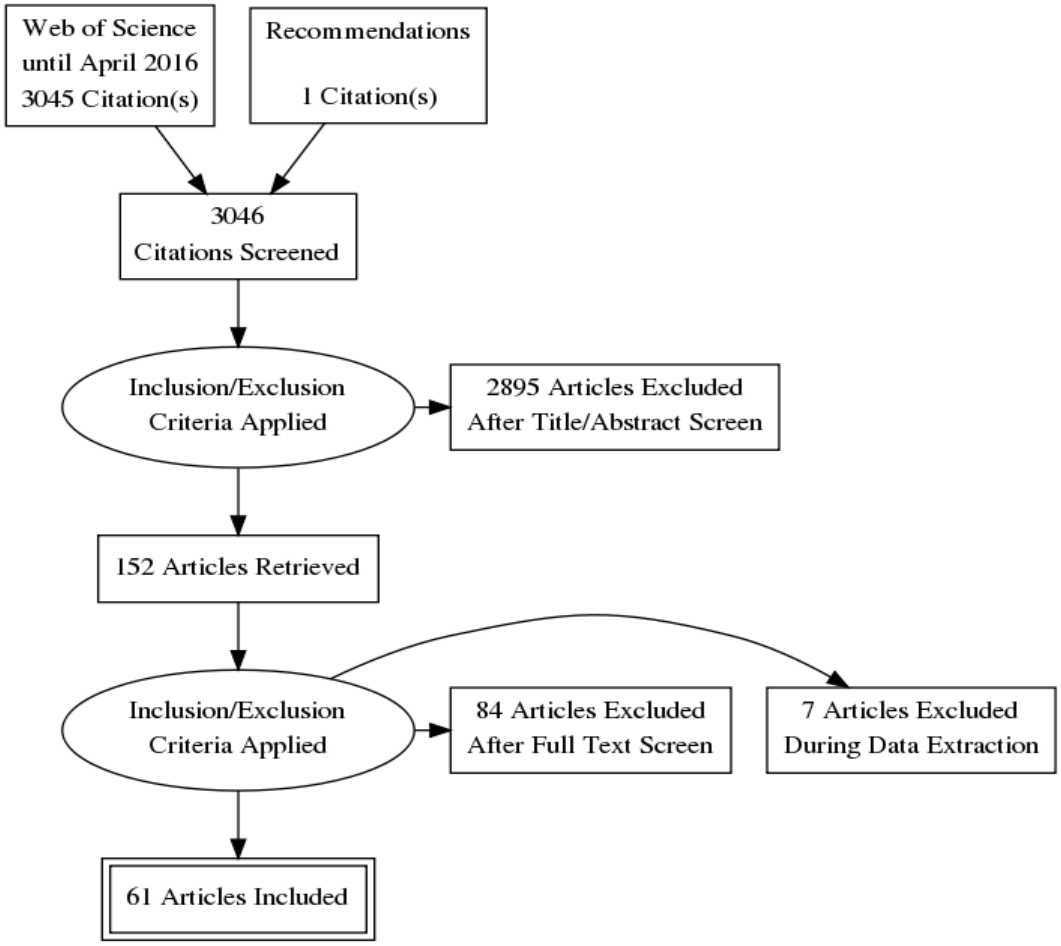

Figure 1:PRISMA Flowchart summary of the literature searches of the 10 seed experiments

Problems encountered

The main limitation of the literature search is that it was limited to published results. Replication attempts, and in particular failed replication attempts that weren't published, couldn't be included since their existence was not recorded in academic databases and their results could therefore not be accessed. The consequence is that the collected data will necessarily reflect any publication bias inherent to academic journals, who often discourage the publication of replication attempts and of non-significant results.

Two problems were encountered when collecting the data. The first difficulty was that the data were sometimes reported visually, in the form of charts, so that the values had to be estimated. A web-application called WebPlotDigitizer[7] was used in order to ensure more precision during the reading of the charts. A second problem was that experimental results (generally standard deviations) or information about the number of participants were sometimes not fully available in the published article. In those cases, an attempt was made to contact the authors of the study. When this attempt was unsuccessful, or that the contacted author was unable to provide the data, a statistical method called "imputation of standard deviations"[8] was used to estimate the missing standard deviations based on the standard deviations of the other studies of that same seed experiment.

As a result of this systematic literature review, the data from 185 experiments (distributed across 10 seed experiments) was collected. A script coded in R was developed in order to analyze these data.

IV. DATA ANALYSES AND RESULTS

Preliminary Analyses

The effect size and the pooled standard deviation of each study was first computed. The sample size was often fairly small: half the studies had less than 24 participants (going as low as 7 participants), with a mean sample size of 30 participants. Because of that the effect sizes were computed using Hedge's g, which allows one to correct for the bias that Cohen's d tends to introduce when used with small samples. A meta-analysis was then performed on each seed experiment separately, which allowed us to compute a summary

7 arohatgi.info/WebPlotDigitizer/app/?

8 The function "impute_SD" from the metagear package (cran.r-project.org/web/packages/metagear/index.html) was used. This function applies Bracken's (1992), "Statistical methods for the analysis of effects of treatment in overviews of randomized trials", it imputes SD by using the coefficient of variation from all complete cases.

effect size for each of them. Because the studies in each seed experiment did not follow the assumptions required for a fixed-effect model, a random-effects model was used. A fixed-effect model requires that all studies be extremely similar, ideally performed by the same team on very similar populations, so that a single true effect size can be assumed to underlie all studies. The random-effects model takes into account the variability induced by different experimental settings (different labs or populations from different countries, for instance) when evaluating the precision assigned to each study.

To interpret effect sizes (and summary effect sizes), the following thresholds, formulated by Cohen (Cohen, 1998) were referred to, although the use of such benchmarking practices should be subject to caution, since an effect size should normally be interpreted in the context of the corresponding experiment (G. Glass, 1981): 0.20 and higher corresponds to a small effect size, 0.50 and higher to a medium effect size, 0.80 and higher to a large effect size and 1.30 and higher to a very large effect size. Out of the ten considered experiments, three experiments didn't reach a significant summary effect size as they had a p-value higher than 0.05 (speech first Word Segmentation, Inferential Statistics with a population first presentation and Visual Statistics); two had a small summary effect size (word first Word Segmentation and Inferential Statistics with a sample first presentation), one had a large effect size (Interpreting Intentional Action) and the remaining four had medium effect sizes.

Seed experiment	Number of studies	Summary Effect Size	P value
False Belief	11	**0.68**	< 0.0001
Number Discrimination	8	**0.58**	< 0.0001
w. Word Segmentation	47	**0.21**	0.001
sp. Word Segmentation	36	0.06	0.418
Head Turn to Faces	13	**0.7**	< 0.0001
s. Inf Statistics	3	**0.42**	0.002
pop. Inf Statistics	10	0.15	0.230
Visual Statistics	20	0.22	0.380
Intentional Action	7	**0.87**	0.004
Grabbing Social Actor	30	**0.74**	< 0.0001

Figure 2: Summary Effect Sizes and p-values

As a preliminary analysis, we first attempted to assess how the original studies compared to their replications. To this end, we compared the summary effect size of the original study with the summary effect size of the replication attempts. Although this method does not allow, strictly speaking, to compute a replicability rate – which is a more complex endeavor, since on the one hand, variations in true effect sizes might explain variations in estimated effect sizes and, on the other hand, the statistical power of both the original study and the replication attempt need to be taken into account to assess the success of a replication (Simonshon, 2015) –, it was chosen because it allows for a clear comparison between original studies and replication attempts. In 7 seed experiments out of 10, the difference between the original study and its replication attempts was not significant. In the 3 remaining experiments, there was a significant difference. In the False Belief experiment, the effect size sank from being very large (1.47, $p = 0.0006$) to being medium (0.56, $p = 0.0006$); in the Word Segmentation experiment with speech first presentation, it sank from being medium (0.57, $p = 0.0061$) to being insignificant (0.04, $p = 0.56$) and in the Grading of Social Actors experiment, it sank from being large (0.93, $p < 0.0001$) to being medium (0.73, $p < 0.0001$). Those results can be understood as showing that one seed experiment did not replicate (the Word Segmentation experiment with speech first presentation) whereas the other three still replicated the effect but not its magnitude, which might also be attributed to a declining effect size trend across time.

Seed expt	**Difference in effect size between original and replications**	**P value of difference**
Fl. Belief	**0.912**	**0.046**
Numb Disc.	- 0.290	0.575
w. Word Seg.	0.387	0.092
sp. Word Seg.	**0.527**	**0.016**
H. Turn to Faces	0.283	0.174
s. Inf Statistics	0.236	0.436
pop. Inf Statistics	0.442	0.106
Vis Statistics	0.394	0.236
Intent. Act.	-0.180	0.663
Grabbing Social Actor	**0.207**	**< 0.0001**

Figure 3: Replicability of original studies

In the following pages, we include a set of figures (Figure 4 to 13: forest plots) to represent effect sizes and confidence intervals of the studies included in our data set of 185 studies. In a forest plot, each study is represented by a box and bounded by a confidence interval (Borenstein, Hedges, et all, 2009). The area of the box is proportional to the inverse of that study's variance. The summary effect size is represented by a diamond, and the width of the diamond represents its confidence interval. The reference line (dotted vertical line) is set at 0.20: generally, effect sizes under 0.20 are not considered to be significant.

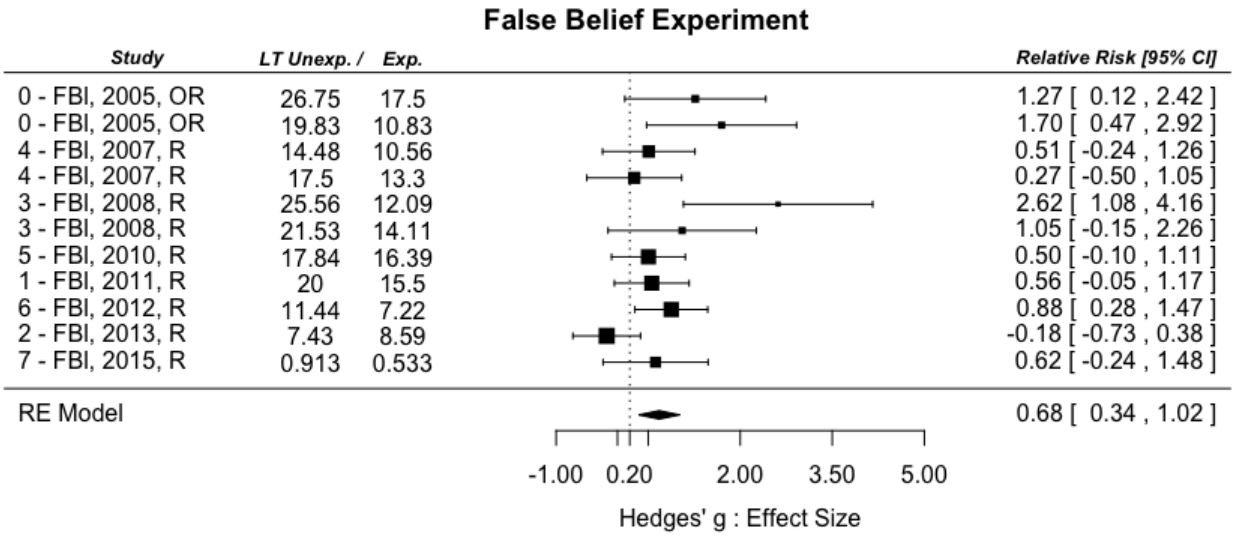

Figure 4: Forest Plot: False Belief Experiment

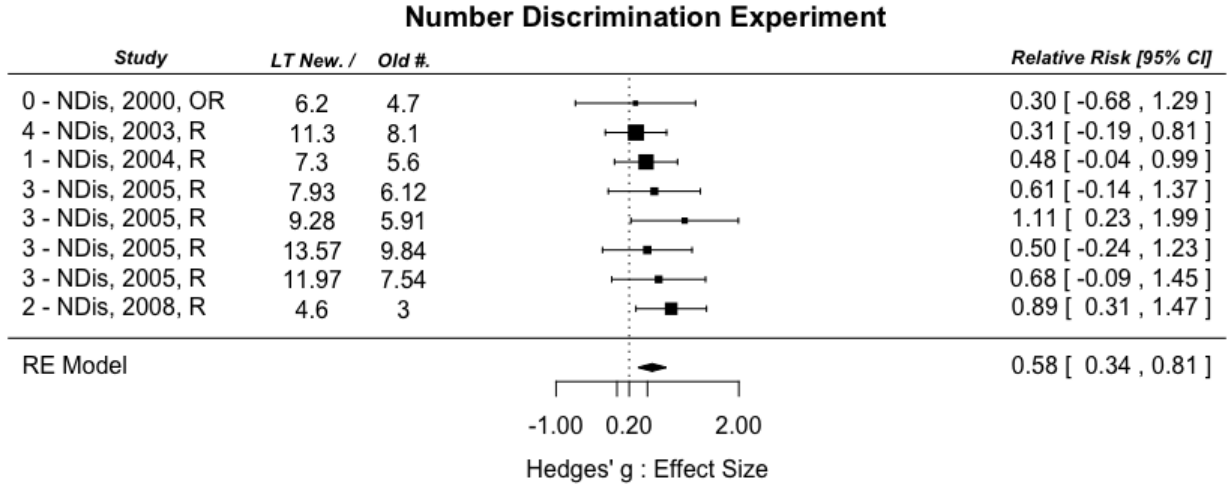

Figure 5: Forest Plot: Number Discrimination Experiment

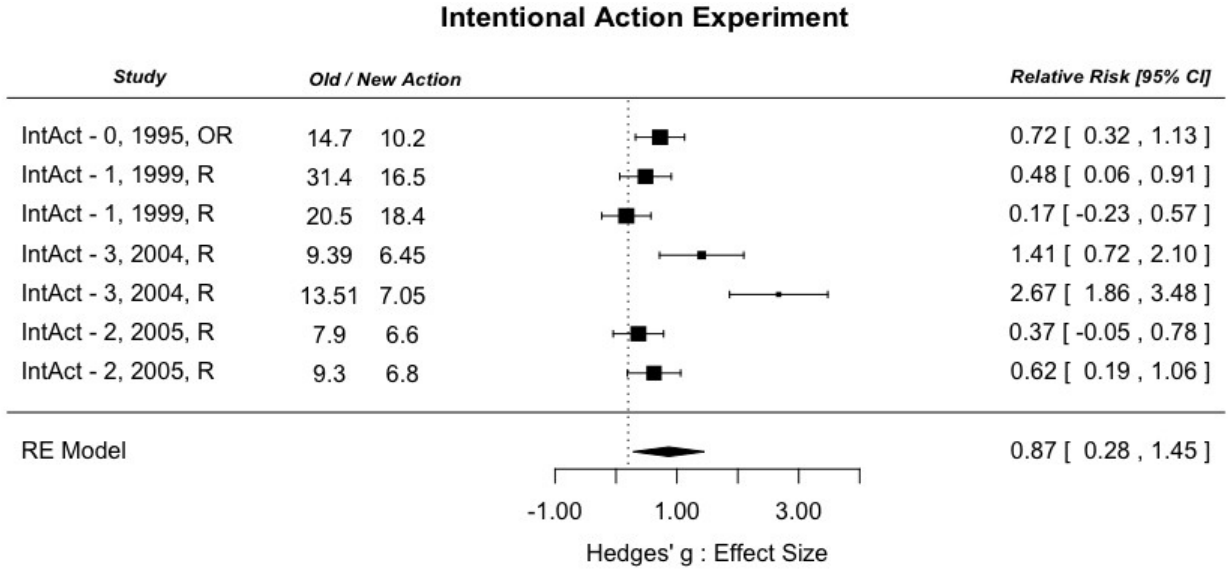

Figure 6: Forest Plot: Intentional Action Experiment

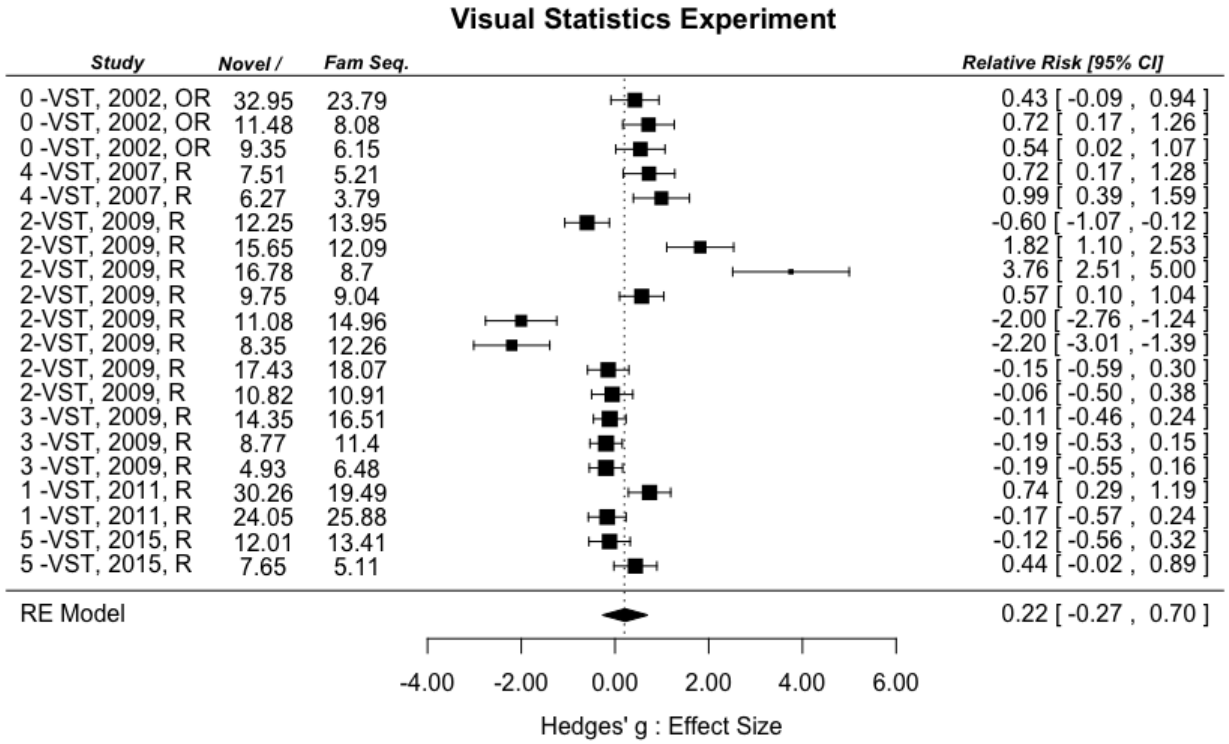

Figure 7: Forest Plot: Visual Statistics Experiment

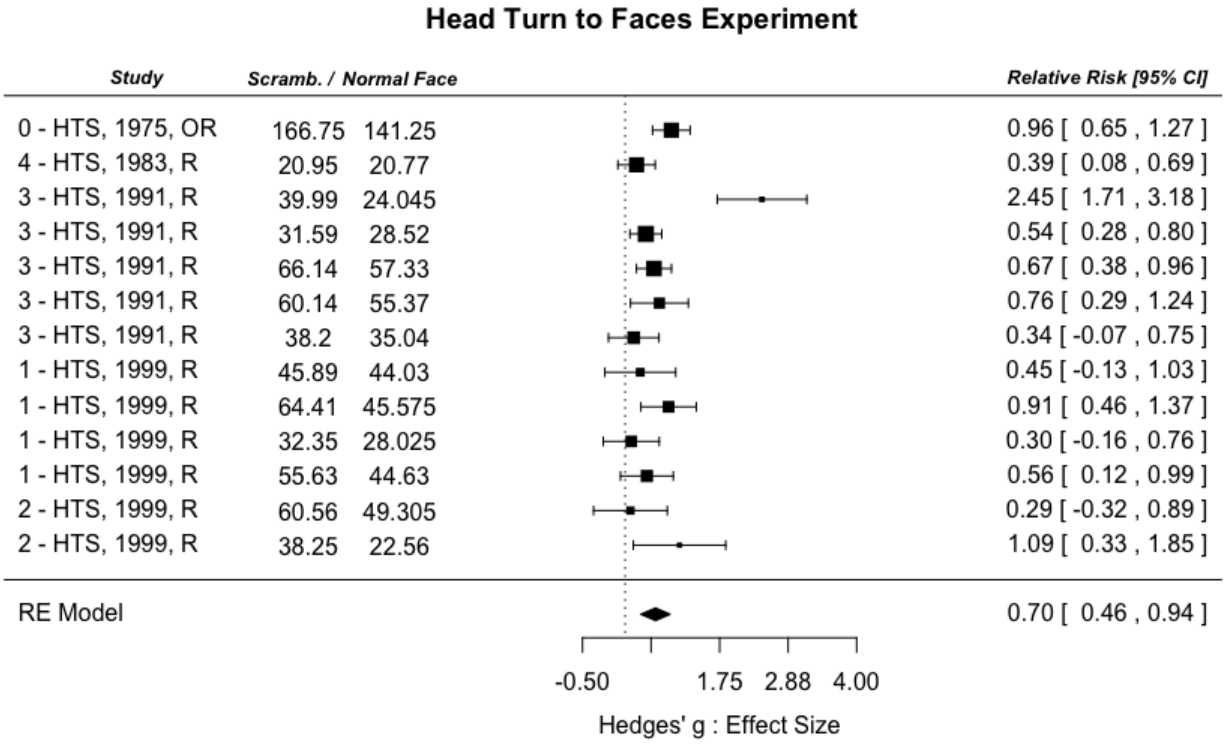

Figure 8: Forest Plot: Head Turn to Faces Experiment

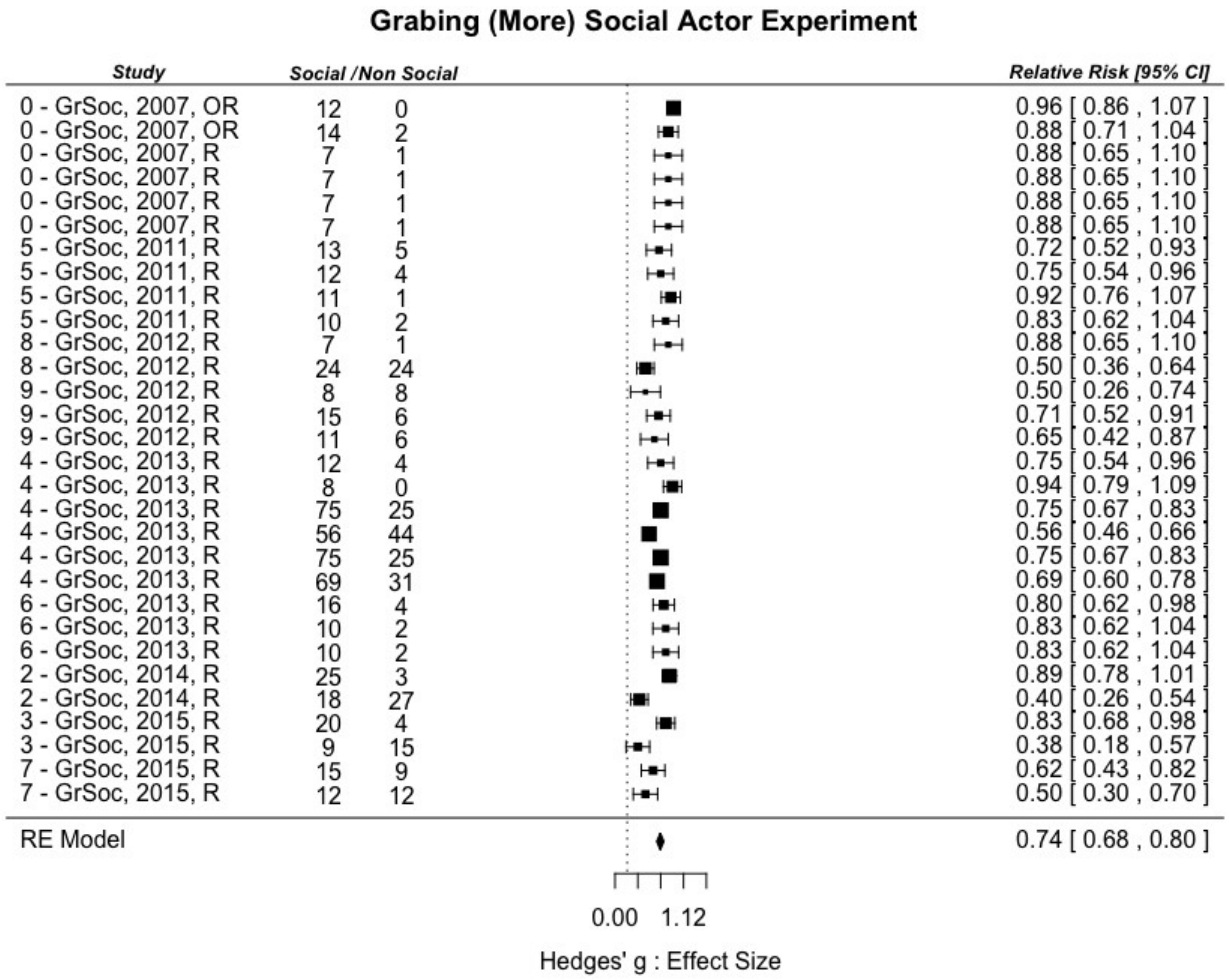

Figure 9: Forest Plot: Grabing Social Actor Experiment

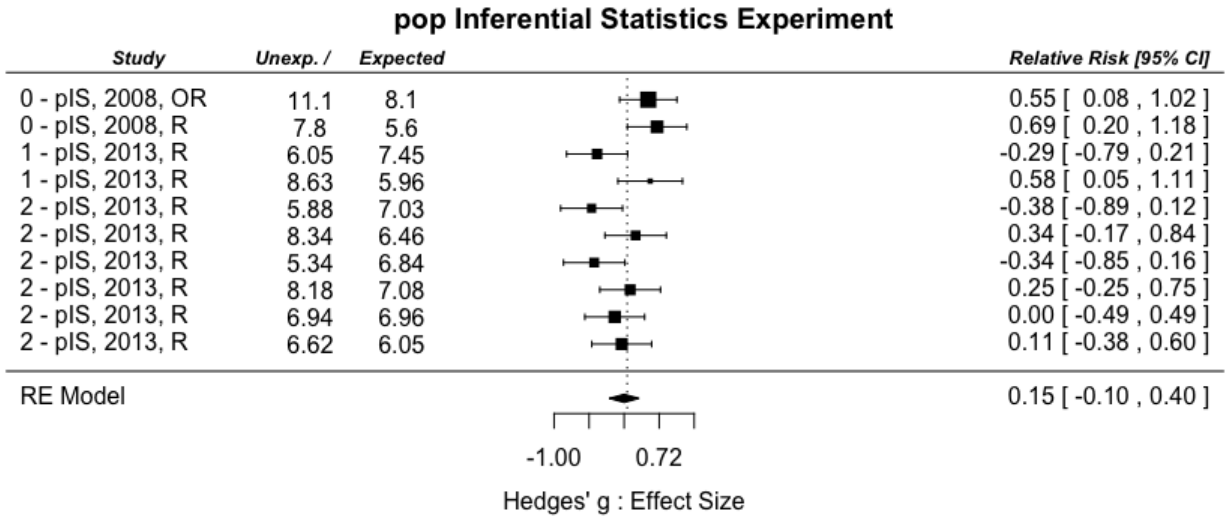

Figure 10: Forest Plot: population first Inferential Statistics Experiment

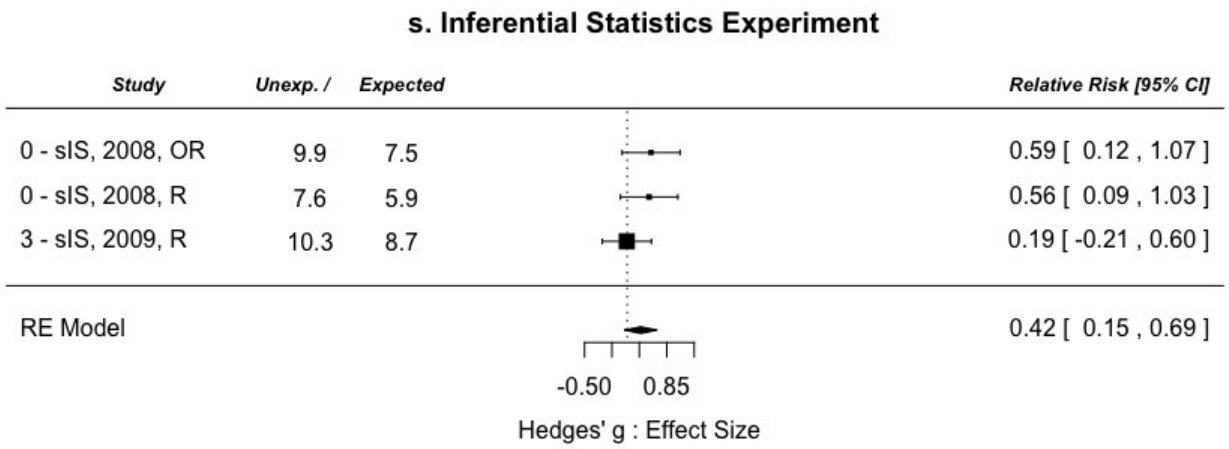

Figure 11: Forest Plot: sample first Inferential Statistics Experiment

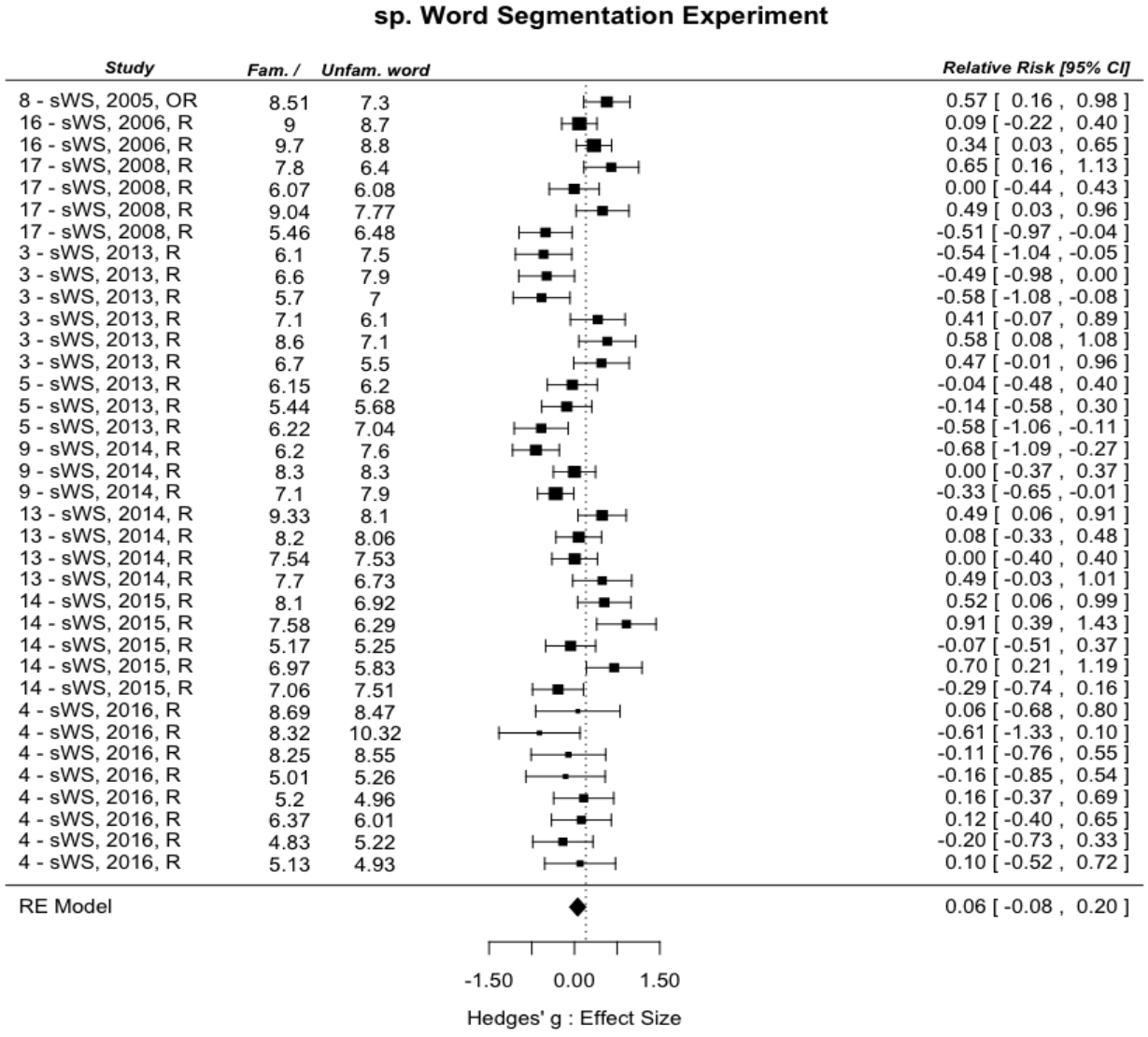

Figure 12: Forest Plot: speech first Word Segmentation Experiment

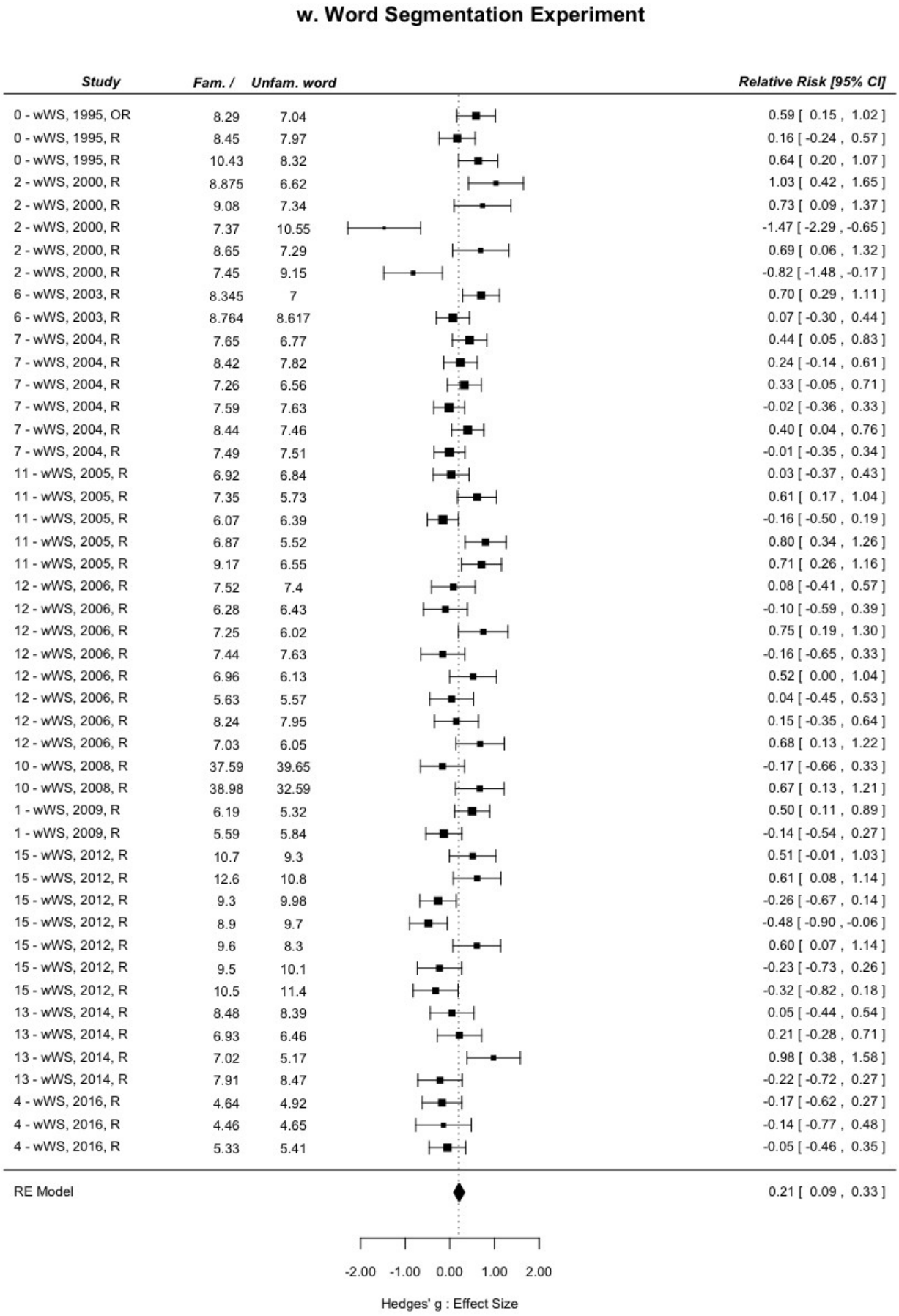

Figure 13: Forest Plot: word first Word Segmentation Experiment

A second preliminary analysis we performed was to examine whether the data set followed a trend of a declining effect size throughout the years. It has indeed been claimed that effect sizes tend to shrink in later studies (Lehrer, 2007), especially in fields such as psychology. To this end, we performed a meta-regression, which is a form of meta-analysis which compares the evolution of effect sizes with the variations of a continuous variable. More precisely, we used the year of publication of the studies as a continuous variable and the seed experiment type as a moderator. The analysis found an average decrease of -0.02 points in effect size per year, which was statistically significant ($p = 0.009$). This can be understood in the following way: if an effect size of 0.8 is large and an effect size of 0.5 is medium, then a given effect size would in average decrease from large to medium in approximately 15 years.

Main Analyses

Let us recall that we formulated two general hypotheses about the effect of methodological factors on the results of an experiment : firstly, if a methodological factor is a source of bias and error, then it will tend to increase the effect size of the experiment; secondly, if a methodological factor is a source of error (absent bias), then it will tend to increase the variability of the collected data. In this study, we decided to investigate three methodological factors: the way the stimulus was presented, the way the infants' reactions were coded and the ratio of participants tested but excluded from the final sample. The prediction was that the first one (stimulus presentation) was a source of bias, and that the last two (type of coding and inclusion ratio) were sources of error. For instance, we distinguished studies where the stimulus was presented manually from studies where it was presented automatically, and predicted that studies with a manually presented stimulus would tend to have higher effect sizes than studies with an automatically presented stimulus. To examine those predictions, we tested for each methodological factor whether it was a source of bias and whether it was a source of error.

Those analyses were run across the full set of studies we collected for the ten experiment seeds: in other words, we didn't test each one of the ten seed experiments separately but grouped them in a unique data set amongst which we distinguished, for instance, between studies with a manual presentation of the stimulus and studies with an automatic presentation of the stimulus. This method raised the following worry: it was possible that some experiments produced higher (or lower) effect sizes than the others because the

manipulation in itself was stronger, for instance, and this could then artificially inflate the overall results of a given condition (e.g. manually presented condition) if that experiment type contributed more studies to that condition than to the other. To control for this possibility, it seemed necessary to also run the analyses on a data set where each seed experiment would be represented in the exact same proportion in each comparison condition. As a result, the analyses would be ran on smaller data sets, but this loss in data would allow for more controlled comparisons because the variations in the data that were not due to the presence or absence of the methodological factors themselves would not drive the results of the analyses. Moreover, the simulation method used to create those matched data sets relies on a procedure which averages across random samples: the goal of this procedure is to ensure that the data sets that are generated are representative of the original data set.

Since each methodological factor was tested using a distinction between two conditions, a specific data set needed to be built for each methodological factor. To create those three specific data sets, we used the following method: for each seed experiment, we counted how the studies were distributed between the two relevant conditions (*i.e.* manual vs automatic, live coding vs non-live coding, high inclusion ratio vs lower inclusion ratio). If a condition contained more studies than its counterpart, then a hypothetical set of studies was generated to represent it. For instance, in the False Belief experiment, there were 3 automatic experiments and 8 manual experiments, so a simulated list of 3 manual studies (whose results are representative of the 8 original experiments) was built so that there would be exactly as many automatic studies as manual studies in the final data set. To generate this simulated set of studies, a hundred samples were selected randomly from the relevant condition for that seed experiment. In the False Belief experiment, 100 samples of three studies were randomly selected amongst the 8 False Belief studies that used a manual presentation of stimuli. In other words, we built 100 lists of manual studies for the False Belief experiment, each containing three experiments, which had been sampled from the manual studies of that experiment. Finally, we used an averaging technique to create a simulated list of studies representing the condition: the effect size and the variance of the first experiment of every one of the 100 samples was averaged to generate the first simulated study's effect size and variance, the effect size and the variance of the second study of every one of the 100 samples was averaged to generate the second hypothetical study's effect size and variance, etc. For the False Belief experiment, three hypothetical manual studies were generated, their effect

sizes and variances were the average of the effect sizes and variances of (respectively) the first, second and third experiment of every one of the 100 samples that had been randomly selected from the initial set of 8 manual studies. In the case of the manual versus automatic distinction, the False Belief seed experiment had a majority of manual experiments, so a randomized hypothetical set of manual studies was generated, whereas for instance the Intentional Action seed experiment had a majority of automatic studies, so a randomized hypothetical set of automatic studies (which was the result of that same sampling and averaging method applied to the effect sizes and variances of automatic Intentional Action studies) was generated. At the end of this process, an equally-matched, randomized data set was constructed which contained for each study: the simulated, randomly generated, set of studies for one condition and the original set of studies for the other condition. For instance, the data set generated for the manual vs automatic methodological factor contained, for the False Belief experiment, the 3 simulated, randomly generated manual experiments and the 3 original automatic experiments (the same ones that were used in the full, non-matched data set). Once those three equally-matched, randomized data sets were generated (one data set for every one of the methodological factors), they were used to test for both the presence of bias (using a subgroup analysis) and for the presence of error (using a comparison of pooled standard deviations). In other words, those equally-matched, randomized data sets were only generated one time: the data used for the subgroup analysis and for the comparison of pooled standard deviations of a given methodological factor was exactly the same.

Seed Experiment	Total of studies	Manual	Automatic	Live	Not Live	High Ratio of Inclusion	Lower Ratio of Inclusion
False Belief	11	8	3	6	5	6	5
Number Dis.	8	6	2	3	5	4	4
w. Word Seg	47	0	47	38	9	13	29
s. Word Seg.	36	0	36	25	11	23	13
HT to Faces	13	10	3	2	11	3	9
s. Inf Stats	3	2	1	1	2	1	0
pop Inf Stats	10	4	6	8	2	5	3
Vis. Stats	20	0	20	16	4	16	4
Intent. Act.	7	1	6	5	2	0	7
Grab. Soci	30	22	8	5	25	17	13

Figure 14: Distribution of studies across conditions

Analysis 1 : Are the methodological factors sources of bias ?

To test whether a methodological factor was a source of bias, a subgroup analysis was performed. A subgroup analysis is a form of meta-analysis. As such, it builds a model that assigns a weight to each study depending on its precision and then determines a summary effect size (a weighted, averaged effect) for the experiment being studied. An experimental study can be described as an attempt to estimate the true effect size of an empirical manipulation. The true effect size of the population, however, remains unknown, and since it is estimated using only a sample from that population, a meta-analytic model takes into account the fact that estimated effect sizes will vary around the true effect size. For all the analyses, a random-effect model was chosen: this model assumes that the true effect sizes of a set of studies for the same experiment follow a normal distribution: this means that in addition to the variation of estimated effect sizes around the true effect size, the variation of true effect sizes themselves will need to be taken into account when determining the precision of the study and thus assigning it weights. Because the studies that were selected had some variation in their experimental methods (they were not performed by the same teams, in the same labs, etc), it was necessary to select such a model, since the assumptions of the alternative model – the fixed-effect model, which assumes that there is a unique true effect size – were not met.

In a subgroup analysis, a set of studies is divided into subgroups, whose estimated summary effect sizes are computed. Those summary effect sizes are compared with each other and the subgroup analysis tests whether the summary effect size of a given subgroup is significantly larger than the others. This allows one to examine whether a given parameter can be said to influence the effect size. In the subgroup analysis performed here, a fixed-effects model was used for the comparison because the heterogeneity within each subgroup was accounted for when fitting the random-effects models for each category. For easier reference, let us state again the thresholds used here for interpreting effect sizes : 0.20 for a small effect size, 0.50 for a medium effect size and 0.80 for a large effect size.

(a) The first subgroup analysis divided the set of studies into two categories: studies with a manual presentation of the stimulus and studies with an automatic presentation of the stimulus. There were 185 studies in total, out of which 132 were automatic and 53 were manual. The subgroup analysis gave the following results: manual studies had a summary effect size of 0.70 ($p < .0001$), and automatic studies had a

summary effect size of 0.24 ($p < 0.0001$), there was thus a significant difference of 0.46 ($p < 0.001$). In other words, it seems that a manual presentation tends to increase the effect size from being small to being medium, just as we had predicted.

That same subgroup analysis was then performed on the randomized, equally-matched data set built using the methods we described earlier. This smaller data set contained 44 studies in total (22 manual and 22 automatic): three out of the ten experiment seeds were not represented in this data set because they only contained automatic studies. Manual studies had a summary effect size of 0.68 ($p < .0001$) and automatic studies had a summary effect size of 0.58 ($p <.0001$), the difference of 0.10 was not significant ($p = 0.201$), so in this case a manual presentation didn't increase significantly the effect size. The important variation in the effect size of automatic studies between the full data set and the smaller equally-matched data set can seem surprising but it is explained by the fact that the three seed experiments which were excluded because they contained exclusively automatic studies had some of the lowest effect sizes of the full data set.

(b) The second subgroup analysis also divided the set of studies into two categories: studies with live coding of the infants' reactions and studies with not-live coding of the infants' reactions. This second category comprised studies with automatic coding, coding on a recording and hybrid coding (where coding was live but coder reliability was checked by re-coding 25 to 50% of the infants' reactions on a recording). Amongst the 185 studies, 76 were coded live and 109 were coded not live. The subgroup analysis produced the following results: live-coded studies had a summary effect size of 0.28 ($p <.0001$) and not-live-coded studies had a summary effect size of 0.52 ($p <.0001$), there was a significant difference of -0.24 ($p = 0.002$), i.e. live coding decreased the effect size from being medium to being small.

That same subgroup analysis was performed as well on a randomized, equally matched data set. This data set contained 88 studies in total (44 live and 44 not live), and all ten seed experiment were represented. Live-coded studies had a summary effect size of 0.45 ($p <.0001$) and not-live-coded studies had a summary effect size of 0.31 ($p <.0001$), with a non-significant increase of 0.14 ($p = 0.09$). In other words, it seems that live coding in this sample didn't increase significantly the effect size.

(c) The third subgroup analysis divided the set of studies into two categories : studies in the third and

fourth quartile, where 77.41% or more of the tested infants were included in the final sample, were defined as having a high inclusion rate, and studies in the first and second quartile (where less than 77.41 % of the tested infants were included in the final sample) were defined as having a comparatively lower inclusion rate. Studies where the inclusion ratio couldn't be computed due to lack of information, generally because the number of infants excluded from the final sample was not indicated in the paper, were excluded from the data. The remaining 175 studies contained 88 studies with a high inclusion ratio and 87 with a low(er) inclusion ratio. The subgroup analysis gave the following results: high inclusion ratio studies had a summary effect size of 0.34 ($p < .0001$) and lower inclusion ratio studies had a summary effect size of 0.40, with a non-significant difference of 0.061 ($p = 0.45$) In other words, it seems that a high inclusion ratio didn't increase significantly the effect size, keeping it at the level of a small effect size.

That same subgroup analysis was then performed on the randomized, equally matched data, using the same value of 77.41% as a threshold. This smaller data set contained 116 studies (58 high and 58 lower): two seed experiments out of the ten seed experiments are not represented in this data set because one of the categories didn't have any observations (the Inferential Statistics experiment didn't have any studies with a low inclusion ratio, whereas the Intentional Action experiment only had studies with a low inclusion ratio). High inclusion ratio studies had a summary effect size of 0.40 ($p < .0001$) and lower inclusion ratio studies had a summary effect size of 0.40 ($p < .0001$), with a non-significant difference of 0.01 ($p = 0.929$). In this case as well, it seems that a high inclusion ratio didn't increase significantly the effect size, keeping it at the level of a small effect size.

Analysis 2: Are the methodological factors sources of error ?

To examine whether a methodological factor was a source of error, an unpaired, independent t-test (Welch two sample t-test) was performed on the pooled standard deviations of the studies, divided along the categories we already described (type of stimulus presentation, type of coding, ratio of inclusion). The pooled standard deviation, also known as the variance of the effect size, expresses the amount of inter-participant variability within a given study: when an important source of error is present in an experimental setup, it seems likely that the data would have less precision and that, as a consequence, its variability would increase. The Welch t-test was chosen to compare the means of the two populations because it was not

possible to assume that the two samples had equal variances.

(a) For the type of stimulus presentation, the t-test compared the pooled standard deviations of manual studies with the pooled standard deviations of automatic studies. The means were estimated to 0.08 (SD = 0.119) for manual studies and 0.07 (SD = 0.046) for automatic studies. The analysis' results showed a non-significant difference of 0.016 between the means : t(58.30) = -0.94, with a p-value of 0.35. The type of stimulus presentation didn't produce a significant variation in the variance of the studies.

The same t-test was then performed on the randomized, equally matched sample (the exact same that was already used for the corresponding subgroup analysis described earlier). Manual studies had a mean pooled standard deviation of 0.10 (SD = 0.084) and automatic studies had a mean pooled standard deviation of 0.06 (SD = 0.054). The analysis showed a non-significant difference of 0.04 between the two means : t(35.642) = - 1.95, with p = 0.058. In this case also, manual studies didn't have a higher average pooled standard deviation than automatic studies.

(b) For the type of coding, the unpaired, independent t-test compared the pooled standard deviations of studies coded live with the pooled standard deviations of studies coded not live. The means of the pooled standard deviations were estimated at 0.08 (SD = 0.087) for live-coded studies and at 0.05 for not live coded studies (SD = 0.049), and the t-test showed a significant difference of 0.023 : t(175.67) = - 2.81, p = 0.006. Studies coded live tended to have significantly higher pooled standard deviations than studies coded not-live.

The t-test was also run on the same randomized sample used earlier for the corresponding subgroup analysis. Live-coded studies had a mean pooled standard deviation of 0.12 (SD = 0.093) and not-live coded had a mean pooled standard deviation of 0.07 (SD = 0.043). The difference of 0.05 was deemed statistically significant : t(60.363) = 3.24, p = 0.002. Again, studies coded live tended to have significantly higher pooled standard deviations than studies coded not-live.

(c) Finally, for the ratio of inclusion, the t-test was used to compare the pooled standard deviations of studies with a high inclusion ratio with the pooled standard deviations of studies with a lower inclusion

ratio. The same threshold of 77.41% (the median of the data) was used. The means were of 0.069 (SD = 0.069) for the high inclusion ratio studies and of 0.073 (SD = 0.08) for the lower inclusion ratio studies. The difference of -0.006 was not significant, $t(167.47) = -0.52$, $p = 0.6$: the ratio of inclusion didn't affect the pooled standard deviations in a significant way.

Again, a t-test was conducted on the same randomized, equally-matched sample as earlier. Studies with a high inclusion ratio had a mean pooled standard deviation of 0.10 (SD = 0.057) and those with a lower inclusion ratio had a mean pooled standard deviation of 0.09 (SD = 0.100). The two means had a difference of 0.01, which the t-test showed to be not statistically significant : $t(90.475) = 0.40$, with $p = 0.683$. In other words, in this sample, a high inclusion ratio did not tend to increase significantly the pooled standard deviation, compared to a lower inclusion ratio.

Interpretation of the results

First, by-and-large more than half of the seed experiments replicated well (meaning that there was not significant difference between the original studies' effect size and the replication attempts' effect size), and only one seed experiment did not replicate at all (meaning that the effect size of the replication attempts was not significant). This suggests that the dominant methods from developmental psychology can produce reliable results and scientific knowledge. From the meta-regression analysis, it can nevertheless be inferred that the data shows a decline in effect sizes across time. Since the declining effect size can be argued to be a general trend, at least in psychology, this result can be interpreted as a confirmation of the representativeness of our data set. In the light of this result, the ten experiments, which were chosen because they represented some of the main results in the field of developmental psychology, can be said to be a fairly representative sample of the research performed in that field. There are competing explanations for the drop in effect size: one is that original studies often have small samples with small powers which increases the risk of producing false positives (Ioannidis, 2005). Another possible explanation is that the publication bias is particularly strong for original studies: whereas replication attempts might sometimes profit from the legitimacy of the original experiment to get published even if their effect sizes are small or not-significant, the legitimacy of original studies might be help to higher standards of statistical significance. Finally, a third possible explanation is that researches might be conducting more and more fine-grained studies in order to better

understand the effects, which is likely to naturally reduce the effect size..

The goal of the subgroup analyses was to assess whether the factors were sources of bias, under the hypothesis that when a factor is a source of bias, it will tend to produce higher effect sizes. We predicted that manually presenting a stimulus would tend to be a source of bias and error. We did find a significant result when testing this hypothesis on the full data set, but the difference stopped being significant on the randomized data set. One possible limitation of this analysis is that the three experiments which only had studies with an automatic presentation of stimuli were also amongst the studies with the lowest summary effect sizes (two with a non-significant summary effect size and one with a small summary effect size), thus making it impossible for these seed studies with their manual counterparts. As those studies were entirely excluded from the randomized, equally matched data set, the summary effect size of automatic studies rose from 0.46 to 0.58. Obviously, since only ten different experiments were selected for the systematic review, and since there weren't any cases of experiments with exclusively manual studies, no conclusions can be drawn from the fact that those three automatic experimental paradigms tended to have lower effect sizes. All in all, the analyses presented in the current work are not sufficient to adopt confidently the hypothesis that the mode of presentation of stimuli is a source of bias, but they still suggest a possible relevance of that methodological factor, that would need to be further investigated. As for the two other methodological factors (coding of the infants' reactions and ratio of inclusion), we predicted that they wouldn't be sources of bias, and expected the results to show no significant increase when a study was coded live or when the ratio of inclusion was comparatively high. Indeed, we found that live-coding could be either associated with a decrease in effect size, in the full data set, or with a non-significant increase in the randomized, equally-matched data set. As for the high ratio of inclusion, it was associated with no significant variation in effect sizes, either for the full data set or for the randomized, equally-matched data set. Those results seem to confirm our predictions that those two methodological factors should not be considered to be sources of bias in experimental paradigms.

We also performed unpaired, independent t-tests on the pooled standard deviations to inquire whether the selected methodological factors might increase the amount of inter-study variability, or in other

words, if they might increase the pooled standard deviation of the studies. Our hypothesis was that a source of error would decrease the precision of the data in a study, and thus increase its pooled standard deviation. We predicted that the mode of stimulus presentation was not a source of error, whereas the type of coding and the inclusion ratio could be sources of error. No significant increase in the pooled standard deviation was found for the mode of presentation of the stimuli, which can be interpreted as meaning that it is not a relevant source of error. Live-coded studies, both in the full data set and in the randomized, equally-matched data set, presented a significant increase in their pooled standard deviations, whereas studies with a high inclusion ratio didn't present any such variation in either data sets. As we predicted, it seems that coding experimental results live (or "online") is a possible source of error, as it tends to increase the pooled standard deviation of the studies. A further possible analysis would be to examine how studies with hybrid coding (where a fraction of the data is recorded and coded again) fare when compared to studies coded exclusively live or exclusively on recording, and eventually to determine what is the minimal threshold of recording (25%, 50% or 75%) that allows to control satisfyingly for the increase in pooled standard deviation (and the decrease in precision). However, contrary to our prediction, high ratios of inclusion do not seem to affect significantly the pooled standard deviations in our data set, and as such they cannot be considered as sources of error in experimental paradigms.

V. CONCLUSION

We can now come back to the thesis defended by David Peterson in "The Baby Factory" and first note that despite what could have been inferred from that article, developmental psychology produces broadly replicated results One of the goals of the present empirical study was to investigate whether the claim that protocol flexibility is used in developmental psychology to favor the production of statistically significant results can be put to the test, using qualitative and systematic methods. Contrarily to what might have been expected, none of the methodological factors we identified turned out to increase effect sizes or statistical significance. In particular, the distinction between manual and automatic studies, that had seemed very convincing as a possible cause of bias at the onset of experiment, was not confirmed by the analyses. One of the methodological factors we identified (the type of coding), however, did produce a significant

variation in precision (pooled standard deviations) of studies. This confirms the predictions we had laid out, but it still does not allow to support the claim that that factors pertains to a system that will inevitably and inaccurately achieve significant results. Studies coded live did not have significantly higher effect sizes than studies coded not live, and a decrease in precision can not be directly connected to the generation of statistically significant results. In a sense, the present endeavor is also an example of the necessity to submit theoretical claims to rigorous empirical testing as opposed to merely focusing on cherry picked observations for qualitative analysis.

However, the present study does present an important limitation: some of the seed experiments contained only a low number of replication attempts, and this reduced the statistical power of the overall analyses. But on the other hand, the fact that finding replications of original experiments is difficult can also be interpreted as a sign that the practice of replication in developmental psychology is not as wide-spread as it could be and that the effects of the publication bias might discourage researchers from conducting replication attempts and from publishing them, in particular when those replication attempts failed. A few tentative recommendations for the field of developmental psychology can be drawn from the empirical analyses we conducted. From the present data set, it doesn't seem that the manual presentation of stimuli increases the risk of bias or of error in an experiment. It seems, however, that the use of online coding does increase the risk of error, even if does not increase the risk of bias or substantially decrease the likelihood of replicating a given finding. Finally, the ratio of inclusion was found to have to no measurable effect in the precision and the outcome of an experiment. In consequence, it seems that developmental psychologists should favor coding on recording when designing experiments if they wish to control for the risk of introducing more error in their experiments. More generally, it seems that the field of developmental psychology would greatly benefit if the practice of systematic replications, and the publication of the results of those replications, was more widespread.

Where the practice of replication is concerned, two final, and more general, recommendations can be drawn from this essay. First, when conducting direct replications, it is of course necessary to strive to reproduce as exactly as possible the original experimental setup, but this does not mean that direct replications should be blind imitations. On the contrary, it is critical to carefully consider whether some

methodological factors might be introducing sources of errors or sources of bias, and if possible to favor methods that are less likely to affect precision or to artificially increase effect sizes. Secondly, the development of a scientific culture of replication cannot happen only at the individual level. Academic institutions should also be responsible for the fostering of a climate where statistically significant results are not *sine qua none* conditions of the publication of an article and where replication attempts are also regarded as valuable scientific work. Indeed, those are necessary conditions if a field wishes to be truly self-correcting. At the onset of the thesis, we described a self-correcting field as one that enables and encourages the practice of replication and of falsification, and which is more focused on putting experimental results under scrutiny rather than preserving the beliefs of its scientific community. The two recommendations we suggest are thus aimed at improving the adequacy of effective scientific practice with normative frameworks of science. In particular, the practice of replication should take into account the criticism that has been made by Popper against the notion of a true scientific fact. Both the results of original studies and of their replications should be considered critically since neither is able to produce true definitive knowledge about the world, but this should not lead to a generalized skepticism towards scientific results. It should rather shift the focus from correctness to reliability and credibility, which rest precisely on the consistent use of replications and falsifications to ensure the detection and correction of falsehoods. To improve this self-correcting process, meta-analytical work should also be included more broadly in the "toolbox" of the critically-minded scientist because it allows to better understand the efficacy of a given experimental manipulation and also allows to think more critically about how replication attempts should be conducted (for instance when determining sample size or choosing experimental techniques). In particular, it would be interesting to replicate the study presented here with more experiments, with other methodological factors and on other fields in order to inquire whether the theory it advances (on the causes of non replicability) is itself generalizable, and because it could be used to investigate the reliability of certain experimental methods. However, such studies are only possible when the practice of replication is wide-spread enough that public reports of replication attempts are both numerous and publicly available.

In conclusion, scientific fields should strive to achieve and maintain high levels of reliability and credibility both from the scientific community and from the general public. To this end, they should adopt a self-correcting and critical mindset, aimed both at the established facts and at the standard experimental

methods of the domain. This is of critical importance because "replication crises", such as the one psychology is currently undergoing, can have a negative impact, not only on a given field, but on the perception of science as a whole. On the other hand, self-correcting practices, although they might seem more costly to put into place, since they require to spend time and resources on already discovered and published facts, can help develop a strong scientific ethos and have a lasting effect on epistemic practices, even outside of science itself.

VI. REFERENCES

MAIN REFERENCES

Bracken, M. B. (1992). Statistical methods for analysis of effects of treatment in overviews of randomized trials. In *Effective care of the newborn infant* (pp. 13-18). Oxford University Press, Oxford.

Cohen, J. (1988). Statistical analysis for the behavioral sciences. *Hillsdale: Lawrance Erlbaum*.

Cooper, H., Hedges, L. V., & Valentine, J. C. (Eds.). (2009). *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation.

Drotar, D. (2010). Editorial: A call for replications of research in pediatric psychology and guidance for authors. *Journal of Pediatric Psychology*, jsq049.

Drummond, C. (2009). Replicability is not reproducibility: nor is it good science. (Conference Paper)

Glass, G. V., Smith, M. L., & McGaw, B. (1981). *Meta-analysis in social research*. Sage Publications, Incorporated.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Med*, *2*(8), e124.

Ioannidis, J. P. (2012). Why science is not necessarily self-correcting.*Perspectives on Psychological Science*, *7*(6), 645-654.

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science*, 0956797611430953.

Kuhn, T. S. (1977). The essential tension: Selected studies in scientific tradition and change. *Chicago: The University of Chicago*.

Lehrer, J. (2010). The truth wears off. *The New Yorker*, *13*(52), 229.

Mayr, E. (1961). Cause and effect in biology. *Science*, *134*(3489), 1501-1506.

Mill, J. S. (1872). *The logic of the moral sciences*. Open Court Publishing.

Nickerson, R. S. (1998). Confirmation bias: A ubiquitous phenomenon in many guises. *Review of general psychology*, *2*(2), 175.

Peterson, D. (2016). The baby factory difficult research objects, disciplinary standards, and the production of statistical significance. *Socius: Sociological Research for a Dynamic World*, *2*, 2378023115625071.

Plato, *Phaedo*, Plato. *Plato in Twelve Volumes, Vol. 1* translated by Harold North Fowler; Introduction by W.R.M. Lamb. Cambridge, MA, Harvard University Press; London, William Heinemann Ltd. 1966. 1925.

Popper, K. R. (1959). The logic of scientific discovery. *London: Hutchinson*.

Rosenthal, R., & Jacobson, L. (1968). Pygmalion in the classroom. *The Urban Review*, *3*(1), 16-20.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological bulletin*, *86*(3), 638.

Simons, D. J. (2014). The value of direct replication. *Perspectives on Psychological Science*, *9*(1), 76-80.

Simonsohn, U. (2015). Small telescopes detectability and the evaluation of replication results. *Psychological Science*, 0956797614567341.

Strickland, B., & Suben, A. (2012). Experimenter philosophy: The problem of experimenter bias in experimental philosophy. *Review of Philosophy and Psychology*, *3*(3), 457-467.

Wason, P. C. (1968). Reasoning about a rule. *The Quarterly journal of experimental psychology*, *20*(3), 273-281.

REFERENCES OF THE STUDIES INCLUDED IN THE META-ANALYSIS

Bartels, S., Darcy, I., & Hoehle, B. (2009). Schwa Syllables Facilitate Word Segmentation for 9-month-old German-learning Infants. In J. Chandlee, M. Franchini, S. Lord, & G. M. Rheiner (Eds.), *Proceedings of the 33rd Annual Boston University Conference on Language Development*, Vols 1 and 2 (pp. 73–84).

Bortfeld, H., & Morgan, J. (2000). The influence of focusing stress on infants' recognition of words in fluent speech. In S. C. Howell, S. A. Fish, & T. KeithLucas (Eds.), *Proceedings of the 24th Annual Boston*

University Conference on Language Development, Vols 1 and 2 (pp. 151–163).

Bosch, L., Figueras, M., Teixido, M., & Ramon-Casas, M. (2013). Rapid gains in segmenting fluent speech when words match the rhythmic unit: evidence from infants acquiring syllable-timed languages. *Frontiers in Psychology*, 4, 106.

Brannon, E. M., Abbott, S., & Lutz, D. J. (2004). Number bias for the discrimination of large visual sets in infancy. *Cognition*, 93(2), B59–B68.

Bulf, H., Johnson, S. P., & Valenza, E. (2011). Visual statistical learning in the newborn infant. *Cognition*, 121(1), 127–132.

Cordes, S., & Brannon, E. M. (2008b). The difficulties of representing continuous extent in infancy: Using number is just easier. *Child Development*, 79(2), 476–489.

Csibra, G., Gergely, G., Biro, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of "pure reason" in infancy. *Cognition*, 72(3), 237–267.

Csibra, G., Gergely, G., Biro, S., Koos, O., & Brockbank, M. (1999). Goal attribution without agency cues: the perception of "pure reason" in infancy. *Cognition*, 72(3), 237–267.

Denison, S., Reed, C., & Xu, F. (2013). The Emergence of Probabilistic Reasoning in Very Young Infants: Evidence From 4.5- and 6-Month-Olds. *Developmental Psychology*, 49(2), 243–249.

Easterbrook, M. A., Kisilevsky, B. S., Hains, S. M. J., & Muir, D. W. (1999). Faceness or complexity: Evidence from newborn visual tracking of facelike stimuli. *Infant Behavior & Development*, 22(1), 17–35.

Easterbrook, M. A., Kisilevsky, B. S., Muir, D. W., & Laplante, D. P. (1999). Newborns discriminate schematic faces from scrambled faces. *Canadian Journal of Experimental Psychology-Revue Canadienne De Psychologie Experimentale*, 53(3), 231–241.

Fantz, R. The origin of form perception. *Scientific American*, 1961, 204, 66-72.

Floccia, C., Keren-Portnoy, T., DePaolis, R., Duffy, H., Delle Luche, C., Durrant, S., … Vihman, M. (2016). British English infants segment words only with exaggerated infant-directed speech stimuli. *Cognition*, 148, 1–9.

Gabalda, B. (2012). *Development of the sense of ownership: Social and moral evaluations.* PhD Manuscript.

Gergely, G., Nádasdy, Z., Csibra, G., & Bíró, S. (1995). Taking the intentional stance at 12 months of age.

Cognition, 56(2), 165–193.

Goren, C. C., Sarty, M., & Wu, P. Y. (1975). Visual following and pattern discrimination of face-like stimuli by newborn infants. *Pediatrics*, 56(4), 544–549.

Goyet, L., Nishibayashi, L.-L., & Nazzi, T. (2013). Early Syllabic Segmentation of Fluent Speech by Infants Acquiring French. *Plos One*, 8(11), e79646.

Hamlin, J. K. (2014). Context-dependent social evaluation in 4.5-month-old human infants: the role of domain-general versus domain-specific processes in the development of social evaluation. *Frontiers in Psychology*, 5, 6141.

Hamlin, J. K. (2015). The case for social evaluation in preverbal infants: gazing toward one's goal drives infants' preferences for Helpers over Hinderers in the hill paradigm. *Frontiers in Psychology*, 5, 1563.

Hamlin, J. K., & Wynn, K. (2011). Young infants prefer prosocial to antisocial others. *Cognitive Development*, 26(1), 30–39.

Hamlin, J. K., Mahajan, N., Liberman, Z., & Wynn, K. (2013). Not Like Me = Bad: Infants Prefer Those Who Harm Dissimilar Others. *Psychological Science*, 24(4), 589–594.

Hamlin, J. K., Wynn, K., & Bloom, P. (2007). Social evaluation by preverbal infants. *Nature*, 450(7169), 557–U13.

Hohle, B., & Weissenborn, J. (2003). German-learning infants' ability to detect unstressed closed-class elements in continuous speech. *Developmental Science*, 6(2), 122–127.

Houston, D. M., Santelmann, L. M., & Jusczyk, P. W. (2004). English-learning infants' segmentation of trisyllabic words from fluent speech. *Language and Cognitive Processes*, 19(1), 97–136.

Johnson, E. K. (2005). English-learning infants' representations of word forms with iambic stress. *Infancy*, 7(1), 99–109.

Johnson, E. K., Seidl, A., & Tyler, M. D. (2014). The Edge Factor in Early Word Segmentation: Utterance-Level Prosody Enables Word Form Extraction by 6-Month-Olds. *Plos One*, 9(1), e83546.

Johnson, m., Dziurawiec, s., Ellis, H., & Morton, J. (1991). Newborns preferential tracking of face-like stimuli and its subsequent decline. *Cognition*, 40(1-2), 1–19.

Johnson, S. P., Fernandes, K. J., Frank, M. C., Kirkham, N., Marcus, G., Rabagliati, H., & Slemmer, J. A. (2009). Abstract Rule Learning for Visual Sequences in 8-and 11-Month- Olds. *Infancy*, 14(1), 2–18.

Jusczyk, P. W., & Aslin, R. N. (1995). Infants' detection of the sound patterns of words in fluent speech. *Cognitive Psychology*, 29(1), 1–23.

Kagan, J., Henker, B. A., Hen-Tov, A., Levine, J., & Lewis, M. (1966). Infants' differential reactions to familiar and distorted faces. *Child Development*, 37(3), 519–532.

Kanakogi, Y., Okumura, Y., Inoue, Y., Kitazaki, M., & Itakura, S. (2013). Rudimentary Sympathy in Preverbal Infants: Preference for Others in Distress. *Plos One*, 8(6), e65292.

Kirkham, N. Z., Slemmer, J. A., & Johnson, S. P. (2002). Visual statistical learning in infancy: evidence for a domain general learning mechanism. *Cognition*, 83(2), B35–42.

Lawson, C. A., & Rakison, D. H. (2013). Expectations About Single Event Probabilities in the First Year of Life: The Influence of Perceptual and Statistical Information. *Infancy*, 18(6), 961–982.

Luo, Y. (2011). Do 10-month-old infants understand others' false beliefs? *Cognition*, 121(3), 289–298.

Marcovitch, S., & Lewkowicz, D. J. (2009). Sequence learning in infancy: the independent contributions of conditional probability and pair frequency information. *Developmental Science*, 12(6), 1020–1025.

Marquis, A., & Shi, R. (2008). Segmentation of verb forms in preverbal infants. *Journal of the Acoustical Society of America*, 123(4), EL105–EL110.

Maurer, D., & Young, R. (1983). Newborns Following Of Natural And Distorted Arrangements Of Facial Features. *Infant Behavior & Development*, 6(1), 127–131.

Nazzi, T., Dilley, L. C., Jusczyk, A. M., Shattuck-Hufnagel, S., & Jusczyk, P. W. (2005). English-learning infants' segmentation of verbs from fluent speech. *Language and Speech*, 48, 279–298.

Nazzi, T., Iakimova, G., Bertoncini, J., Fredonie, S., & Alcantara, C. (2006). Early segmentation of fluent speech by infants acquiring French: Emerging evidence for crosslinguistic differences. *Journal of Memory and Language*, 54(3), 283–299.

Nazzi, Thierry; Mersad, Karima; Sundara, Megha; Iakimova, Galina; Polka, Linda. (2014). Early word segmentation in infants acquiring Parisian French: task-dependent and dialect-specific aspects. *Journal of Child Language.*

Nishibayashi, L.-L., Goyet, L., & Nazzi, T. (2015). Early Speech Segmentation in French-learning Infants: Monosyllabic Words versus Embedded Syllables. *Language and Speech*, 58(3), 334–350.

Onishi, K.; Baillargeon, R. (2005). Do 15-Month-Old Infants Understand False Beliefs?. *Nature*

Phillips, A. T., & Wellman, H. M. (2005). Infants' understanding of object-directed action. *Cognition*, 98(2), 137–155.

Polka, L., & Sundara, M. (2012). Word Segmentation in Monolingual Infants Acquiring Canadian English and Canadian French: Native Language, Cross-Dialect, and Cross-Language Comparisons. *Infancy*, 17(2), 198–232.

Poulin-Dubois, D., Polonia, A., & Yott, J. (2013). Is False Belief Skin-Deep? The Agent's Eye Status Influences Infants' Reasoning in Belief-Inducing Situations. *Journal of Cognition and Development*, 14(1), 87–99.

Saffran, J. R., Pollak, S. D., Seibel, R. L., & Shkolnik, A. (2007). Dog is a dog is a dog: Infant rule learning is not specific to language. *Cognition*, 105(3), 669–680.

Salvadori, E., Blazsekova, T., Volein, A., Karap, Z., Tatone, D., Mascaro, O., & Csibra, G. (2015). Probing the Strength of Infants' Preference for Helpers over Hinderers: Two Replication Attempts of Hamlin and Wynn (2011). *Plos One*, 10(11), e0140570.

Scarf, D., Imuta, K., Colombo, M., & Hayne, H. (2012). Social Evaluation or Simple Association? Simple Associations May Explain Moral Reasoning in Infants. *Plos One*, 7(8), e42698.

Seidl, A., & Johnson, E. K. (2006). Infant word segmentation revisited: edge alignment facilitates target extraction. *Developmental Science*, 9(6), 565–573.

Seidl, A., & Johnson, E. K. (2008). Boundary alignment enables 11-month-olds to segment vowel initial words from speech. *Journal of Child Language*, 35(1), 1–24.

Slone, L. K., & Johnson, S. P. (2015). Infants' statistical learning: 2-and 5-month-olds' segmentation of continuous visual sequences. *Journal of Experimental Child Psychology*, 133, 47–56.

Slone, L. K., & Johnson, S. P. (2015). Infants' statistical learning: 2-and 5-month-olds' segmentation of continuous visual sequences. *Journal of Experimental Child Psychology*, 133, 47–56.

Sodian, B., Schoeppner, B., & Metz, U. (2004). Do infants apply the principle of rational action to human agents? *Infant Behavior & Development*, 27(1), 31–41.

Song, H., & Baillargeon, R. (2008). Infants' Reasoning About Others' False Perceptions. *Developmental Psychology*, 44(6), 1789–1795.

Surian, L., Caldi, S., & Sperber, D. (2007). Attribution of beliefs by 13-month-old infants. *Psychological*

Science, 18(7), 580–586.

Traeuble, B., Marinovic, V., & Pauen, S. (2010). Early Theory of Mind Competencies: Do Infants Understand Others' Beliefs? *Infancy*, 15(4), 434–444.

Xu, F. (2003). Numerosity discrimination in infants: Evidence for two systems of representations. *Cognition*, 89(1), B15–B25.

Xu, F., & Denison, S. (2009). Statistical inference and sensitivity to sampling in 11-month-old infants. *Cognition*, 112(1), 97–104.

Xu, F., & Garcia, V. (2008). Intuitive statistics by 8-month-old infants. *Proceedings of the National Academy of Sciences*, 105(13), 5012–5015.

Xu, F., Spelke, E.S., Large number discrimination in 6-month-old infants, *Cognition*, Volume 74, Issue 1, 10 January 2000, Pages B1-B11.

Xu, F., Spelke, E. S., & Goddard, S. (2005). Number sense in human infants. *Developmental Science*, 8(1), 88–101.

Yott, J., & Poulin-Dubois, D. (2012). Breaking the rules: Do infants have a true understanding of false belief? *British Journal of Developmental Psychology*, 30(1), 156–171.

Zmyj, N., Prinz, W., & Daum, M. M. (2015). Eighteen-month-olds' memory interference and distraction in a modified A-not-B task is not associated with their anticipatory looking in a false-belief task. *Frontiers in Psychology*, 6, 857.