A meta-analytical answer to the crisis of confidence of psychology

Juan Botella[1,*], and Juan I. Duran[2]

1 Universidad Autónoma de Madrid (Spain)
2 Centro Universitario Cardenal Cisneros, Madrid (Spain)

Título: Una respuesta meta-analítica a la crisis de confianza de la psicología.

Resumen: El meta-análisis es una metodología firmemente establecida y una parte integral del proceso de generación de conocimiento en las ciencias empíricas. El meta-análisis también se ha centrado en la metodología y se ha convertido en uno de los principales críticos de las deficiencias metodológicas. Destacamos varios puntos problemáticos sobre cómo investigamos en psicología: exceso de heterogeneidad en los resultados y dificultades para la replicación, sesgo de publicación, calidad metodológica subóptima y prácticas cuestionables de los investigadores. Estos y otros problemas condujeron a una "crisis de confianza" en psicología. Discutimos cómo la perspectiva meta-analítica y sus procedimientos pueden ayudar a afrontar la crisis. Una perspectiva más cooperativa, en lugar de competitiva, puede ayudar a cambiar para que consideremos la replicación como una contribución más valiosa. El conocimiento no puede basarse en estudios aislados. Dada la naturaleza del objeto de estudio de la psicología la unidad natural para generar conocimiento debe ser la distribución estimada del tamaño del efecto, no la decisión dicotómica sobre la significación estadística en estudios específicos. Se ofrecen algunas sugerencias sobre cómo redirigir la investigación y las prácticas de los investigadores, de modo que sus intereses personales y los de la ciencia en sí estén más alineados.

Palabras clave: meta-análisis; crisis de confianza; tamaño del efecto; prácticas cuestionables.

Abstract: Meta-analysis is a firmly established methodology and an integral part of the process of generating knowledge across the empirical sciences. Meta-analysis has also focused on methodology and has become a dominant critic of methodological shortcomings. We highlight several problematic issues on how we research in psychology: excess of heterogeneity in the results and difficulties for replication, publication bias, suboptimal methodological quality, and questionable practices of the researchers. These and other problems led to a "crisis of confidence" in psychology. We discuss how the meta-analytical perspective and its procedures can help to overcome the crisis. A more cooperative perspective, instead of a competitive one, can shift to consider replication as a more valuable contribution. Knowledge cannot be based in isolated studies. Given the nature of the object of study of psychology, the natural unit to generate knowledge must be the estimated distribution of the effect sizes, not the dichotomous decision on statistical significance in specific studies. We make some suggestions on how to redirect the research and the researchers' practices, so that their personal interests and those of science as such are better aligned.

Keywords: meta-analysis; crisis of confidence; effect size; questionable practices.

Introduction

Meta-analysis arose from the need for a rigorous and systematic review of a body of scientific evidence. The objective assumed in its development was to propose methods for the synthesis of evidence while dealing with any heterogeneity of the effects (Botella & Sánchez-Meca, 2015; Cooper, Hedges & Valentine, 2009; Glass, 1976). Although it originated in the social sciences, halfway between education and psychology, its use has spread to practically all empirical sciences. We can say that nowadays it is a firmly established methodology and an integral part of the process of generating knowledge.

Let us examine the use of meta-analysis through some data from the field of educational psychology. We will take as reference the five journals with the highest impact factor in Journal Citation Reports in 2017 within the category "Psychology, Educational"[1]. In the volumes corresponding to that year, those journals published 357 articles, of which 252 are empirical studies, 11 are meta-analytic reviews (3.1%) and the rest are other types. That is, there is a ratio of 23:1 between the primary empirical studies and the meta-analyses. If

we look back, to what happened with the journals of the five highest positions in the same category 20 years before (1997[2]) we observe the following: 241 published articles, of which 186 were primary empirical studies, 2 were meta-analysis (0.8%) and the rest were other types. The ratio between primary studies and meta-analysis was 93:1. The presence of meta-analysis among the journals with the greatest impact has grown substantially. Meta-analysis serves very often as the base of empirical arguments repeated frequently.

The data that seem most revealing about the trends derived from the consolidation of the meta-analysis is that of those 186 primary studies one third (34.4%; 64/186) have subsequently served as a primary source in some meta-analysis. Many of them in more than one. That is, today it is quite frequent (in some areas it is almost routine) that research provides not only the evidence primarily analyzed by the authors who have generated it, but end up integrated into some subsequent synthesis. That is why when we publish primary studies we should think not only about how our colleagues will receive and value our results and about how they can contribute in a singular way to the advancement of knowledge. We should also think in that the results will be a piece of a larger theoretical understanding of phenomena based on the evidence provided by many others.

* Correspondence address [Dirección para correspondencia]: Agustín Juan Botella. Universidad Autónoma de Madrid, Facultad de Psicología, Campus de Cantoblanco, c/ Ivan Pavlov, 6, 28049 Madrid (Spain).
E-mail: juan.botella@uam.es

1 *Journal of Educational Psychology*, *Child development*, *Learning and Instruction*, *Educational Psychology Review*, *Educational Psychologist*.

2 *Journal of Educational Psychology*, *Child development*, *Journal of Counseling Psychology*, *Reading Research Quarterly*, *Educational Psychologist*.

The original objectives of meta-analysis remain valid, and in fact have been normalized in most sciences. However, great changes usually bring unforeseen consequences. The appearance of meta-analysis and its progressive incorporation into normal science has served as a catalyst for the evaluation of our methodology. It is not a coincidence that meta-analysis was developed from research traditions in the intersection of psychology and education (Botella & Gambara, 2006). These research areas maintain strong internal debates about their epistemology and research methods. As we will see, the perspectives provided by meta-analysis led to a focus on methodology. On many occasions, meta-analyses found unexpected, scattered, inconsistent, and even contradictory results in the primary studies (Botella & Eriksen, 1991). These results can obviously be related to the way in which the meta-analyses are carried out. As in the case of primary studies, some meta-analyses may also show inconsistent results due to reproducibility deficits in the procedure used to collect and process the information, as some recent studies have pointed out (Gøtzsche, Hróbjartsson, Marić, & Tendal, 2007). However, in addition to the difficulties of having such a complex object of study as human behavior, the question was inevitably asked whether these results were because we have developed models and theories of insufficient power. Perhaps the lack of control of relevant environmental and genetic variables has led to a focus on irrelevant and non-explanatory variables, leading to explanations based in part on mere methodological artifacts or circumstances beyond the relevant scientific domains.

Many meta-analyses that currently published are not limited to the combined estimates of the effect size and the analysis of potentially moderating variables. They also add other analyses related to variables outside the substantive issues under study (e.g., Quiles-Marcos, Quiles-Sebastián, Pamies-Aubalat, Botella, & Treasure, 2013). Step-by-step, the idea is growing that meta-analytic tools can be useful for the study and analysis of our basic scientific practices themselves. One intended purpose is to determine the degree to which the observed heterogeneity in the results is genuine (even though it is a consequence of moderation effects not yet identified) and to which degree it is due to factors external to what is being studied, especially the practices of the scientists themselves.

Once we entered the 21st century, psychology (or at least some of its main fields) has become immersed in a severe crisis, visible in multiple publications and conferences, which has been identified as a "crisis of confidence". The main idea that we want to convey in the present article is that one way to address this crisis and lead to a positive outcome is to apply the meta-analysis perspective, extending its methods and its way of understanding science to the various fields of psychology. Thanks to the fact that meta-analysis focuses on methodology and has become a dominant critic of methodological shortcomings, it can offer very useful tools to face the crisis.

A science of science: How we do research

Researchers assume the postulates of scientific methods. We have many sources with guidelines, protocols, technologies and many practical tools and resources where we can learn how to carry out correct and efficient scientific research. Certainly, we continue to debate and discuss the facets that can be improved, but that does not mean that there are not clear evaluative criteria. However, what the results of some meta-analyses tell us is that research methods are not always correctly implemented and that external factors that strongly affect the final results are frequently involved. We can say that we already have some guidelines, derived from epistemology, on how to make correct and credible scientific advances, but now we observe that they are not always implemented, and such breaches have negative consequences. We are concerned with collecting and organizing relevant empirical evidence regarding how research is actually being done. The great heterogeneity sometimes observed in the results has raised many doubts about the robustness of the methodology.

One of the first victims has been the null hypotheses testing, once a ubiquitous tool that was now being questioned (Cumming, 2012, 2014, Frick, 1996, Gigerenzer, 1993, Harlow, Mulaik, & Steiger, 1997; Nickerson, 2000; Tryon, 2001). The fact that meta-analysis typically begins with estimates of effect size provided by the primary studies, and not its dichotomous decision on statistical significance, is slowly leading to undermine the influence of p values. For a meta-analyst, this is trivial, since we rarely are interested in whether the primary studies showed statistically significant effects. That detail is not usually relevant to our objectives. Undoubtedly, hypotheses testing continues to predominate in primary studies, but it is no longer as decisive in theoretical arguments. Meta-analysis shows that a typical body of evidence includes independent estimates in which the result is not significant with respect to the null hypothesis. In fact, it is rare that there is statistical significance in all the primary studies. It is the combined studies and the combined evidence that provide us with a realistic and credible picture. In the context of meta-analysis, we do not consider a non-significant result as a failure of replication, but one more of the values, perfectly compatible with a typical distribution of effects. Likewise, some synthesis studies have shown that a non-negligible part of the variability observed in the results can be explained by variations in the methods, designs and other methodological aspects (e.g., Lipsey & Wilson, 1993).

A sample of problematic aspects

The image that the mirror returns when observing what meta-analyses show us about methodological practices has many faces and not all are positive. We highlight here only some of them, those that seem most relevant for our argument.

(a) *High heterogeneity* in the results. The persistent observation of this heterogeneity reveals one of the main challenges of psychology. In meta-analysis the estimated variance of true effects or *specific variance*, is represented by τ^2. It is not the square of the standard error, which would be the error of estimation of the value corresponding to the mean effect size. Rather, it is the parametric variance of the effects. For example, suppose that we are going to do an intervention that consists of students' self-assessment throughout some course. We assess its effect by comparing the academic performance with those of a control group that does not self-assess their performance. We can think that the impact of this intervention is a concrete value, which could be specified in terms of Cohen's *d* index (standardized mean difference). Actually, the true parametric value is δ, while with the data of the study we obtain an estimate of it, *d*. If we collect *k* independent estimates from independent studies and combine them meta-analytically the calculated value, $d_\bullet$, is the most efficient estimate of the parametric value. However, what we have learned is that the variability observed in the values of *d* often exceeds that which would be expected as a consequence of the mere sampling of individuals. Even taking into account some moderating variables that we have been able to identify as moderators of the effect, there is still too much heterogeneity. The conclusion is obvious. As much as we attempt to operationalize the interventions, there are always unforeseen factors that modify their real parametric value. Surely, the intervention carried out does not have the same effect according to the teachers' experiences, their personalities, the cultural contexts of the schools, the personal history of the participants, and many other factors that we are not able to identify. That is, we must accept that the effects of our interventions have a variable component that cannot be ignored. In fact, today random-effects models are routinely employed in meta-analyses, the kinds of models in which this characteristic is assumed (Borenstein, Hedges, Higgins & Rothstein, 2010, Botella & Sánchez-Meca, 2015).

But let's see how important such heterogeneity can be. The square root of τ^2 is the standard deviation of these effects. It allows us to calculate the coefficient of variation (*CV*; Botella, Suero & Ximénez, 2012), which is defined as

$$CV = (\tau / mean\_effect\_size) \times 100$$

High values indicate that the variability is very large with respect to the average effect. Figure 1 shows two cases with normal distributions (mean effect size = 1). In A, the standard deviation of the effects is equal to a quarter of the effect size (*CV* = 25), while in B, it equals half (*CV* = 50). In the first case, it must be interpreted that in any future study, it should be expected (with probability around .95) a parametric effect size between 0.50 and 1.50. In the second, the effect size would be in the range between 0 and 2. Many meta-analyses in the field of educational psychology indicate that the *CV* values in the examples shown in the figure are not uncommon. This means that although the size of the expected parametric effect is equal to 1, a relatively large value, we cannot rule out that in our next self-assessment intervention the effect will be much smaller. In other words, when we use a certain educational intervention, *we cannot guarantee an impact reasonably close to the mean effect estimated in the meta-analysis*.
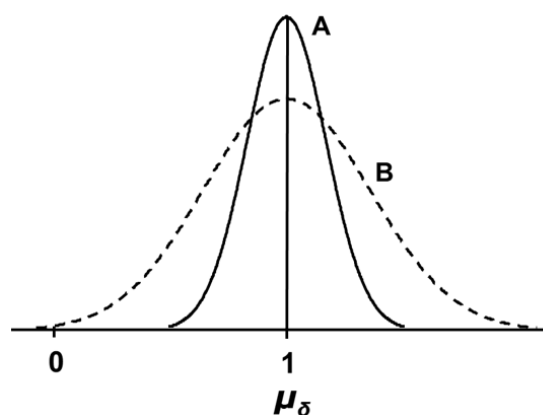

Figure 1. Hypothetical distributions of parametric effect sizes of a given intervention with a large mean effect size (1). Curve A shows a distribution with medium heterogeneity, whereas that of curve B shows a large heterogeneity.

Our interventions are influenced by a great variety of unknown or uncontrollable factors (Botella & Zamora, 2017). Our constructs are often poorly defined and our object of study is subject to multiple causalities, isolated and interacting with others. We already suspected this and are willing to accept that our discipline works this way, and that behind the reported main effect we will always have to attach a catalog of moderating factors. However, we still have doubts about whether this heterogeneity also reveals methodological artifacts that we could control and reduce.

(b) A clear *publication bias* favors studies with statistically significant results and is a classic case in the study of the external circumstances that influence the development of science. We have long known that the preference for publishing statistically significant results ends up resulting in an excess of these values (Bakker, Dijk & Wicherts 2012, Ioannidis & Trikalinos, 2007, Sterling, Rosenbaum & Weinkam, 1995). The consequence is that the synthesis of the available (published) results produces effect sizes that are overestimated. We have empirical evidence of that in integration studies, such as Lipsey and Wilson (1993), who compared published and unpublished studies, and also in funnel plots and in other sources. The evidence is overwhelming that publication bias has a major impact. The most extreme case is that in which the distribution of published effects, with an appreciable average effect, is no more than the visible part of a mostly hidden iceberg. As an example, in figure 2 we can see on the left an empirical distribution of published *d* values (visible) with an average value around 0.60. This distribution presents us with a very different idea from that which would be conveyed by the complete distribution, which appears on

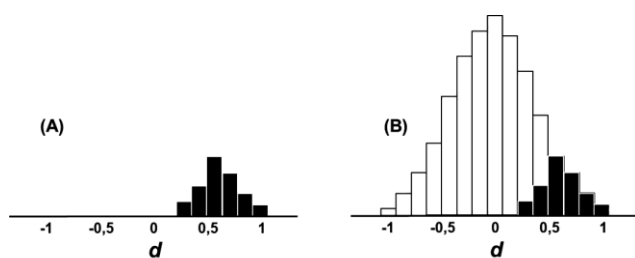the right, with an average value equal to 0. In this, what was hidden (not published) has been made visible.

Figure 2. An extreme case of publication bias. Panel A shows the small sample of studies published, all with moderate to high values of effect size (and probably all statistically significant). Panel B shows the complete distribution of values. The black zone in figure B shows the "visible" (published) part of a kind of iceberg, whereas the white zone is the "invisible" (unpublished) part.

Meta-analysis has attempted to deal with the problem of publication bias practically from the beginning, providing tools designed to assess its impact and trying to correct its consequences (Rosenthal, 1979, Rothstein, Sutton & Borenstein, 2005). This effort has led many researchers to believe that publication bias is a specific problem of meta-analysis, but actually it is not. It is a problem for anyone who wants to get an idea of the "state of the art" regarding any scientific question. Taking what is published as the source, one will have a biased idea of the extent to which what is published (or accessible) is a true sample of the evidence provided by the studies conducted.

(c) A clearly suboptimal *methodological quality* in the average of the published studies. In many meta-analyses, statistics are reported regarding the methodological characteristics of the studies, on aspects such as randomization, the clear operationalization of the manipulations, the blind evaluation, the analysis of the experimental mortality or attrition, the absence of pre-test comparisons, the use of reliable instruments, etc. What was a suspicion became an uncomfortable observation: in much research, not all the methodological resources available to improve the interpretability of the results are implemented (Klein et al., 2012). The great variety observed in the methodological quality of the studies has led to suspicions that part of the heterogeneity of the results is due to the fact that the quality of many studies is low, or at least is very heterogeneous. A positive consequence of these results has been an attempt to develop quality scales of primary studies (Botella & Sánchez-Meca, 2015). The concern about what the meta-analyses reveal regarding this aspect is leading to pay more attention to how the primary studies are carried out and published.

(d) An excessive presence of *questionable practices*, something that had become an open secret (John, Loewenstein & Prelec, 2012). The motivation to obtain and publish statistically significant results has led to opportunistic behaviors (DeCoster, Sparks, Sparks, Sparks & Sparks, 2015). Questionable practices include the elimination of atypical participants without an explicit criterion; the statistical analysis with incomplete data and the decision to add more participants according to the results (Botella, Ximénez, Revuelta & Suero, 2006); the use of several statistical techniques reporting only those that yield significant results; reporting results of analyses aimed to test hypotheses raised after the results are known ("harking"), but presenting them as a priori objectives ("sharking"; Hollenbeck & Wright, 2017); the analysis of multiple dependent variables reporting only the significant ones, etc. This set of practices has led to an obvious excess of significant results in the scientific journals (Ioannidis, 2005; Ioannidis & Trikalinos, 2007). Some practices of scientists have become a problem for the development of science. Their personal objectives interfere with those of science, and their opportunistic practices generate noise.

The crisis of confidence: an almost unavoidable consequence

The reporting of bad news finally caused a crisis. In psychology it has come to be called *crisis of confidence* (Earp & Trafimow, 2015), and we cannot say that the crisis has yet been overcome. The crisis of confidence has had as its main axis the difficulties, apparently excessive, to reproduce the effects that form the bases of our models (Baker, 2016; OSC, 2015; Pashler & Wagenmakers, 2012; Yong, 2012). We believe that the debate on reproducibility has helped to a better understanding of what effect size means and its role in theory development. On the one hand, some researchers have discovered through this debate that the failure to find a significant result where others have found one is not an anomalous result. Rather, the odd thing would be that there were no such cases (except when the power of the tests approaches 1, its ceiling value). Replicability cannot be evaluated exclusively through consistency in the dichotomous decision on statistical significance (Pashler & Harris, 2012). Obtaining a natural distribution of values of the chosen index for effect size is a good sign, even if this distribution includes small (not significant) or even null or inverse values. This is what is expected from the meta-analytical perspective.

However, we cannot rule out that some well-known effects are mere artifacts, due mainly to questionable practices. Factors external to science can have this unwanted consequence. There are both internal and external factors that complicate the efficient development of the social sciences. Among the *internal factors* are the weaknesses of the constructs we manage. While we have made advances in reliability thanks to the psychometric and assessment developments of the last three decades, we still have difficulties in defining measures with high validity (e.g., the difficulties to define the concept of complexity; Arend, Colom, Botella, Contreras, Rubio & Santacreu, 2003). Sometimes it is said that if we do not measure what is important we will end up giving importance to what we measure. We suspect that something like this happens in our discipline, in which we often have the feeling that we miss crucial elements that we cannot cap-

ture with our instruments. We have a lot of work to do in order to get coefficients of variation that make the effects of our interventions more predictable.

Among the *external factors*, not related to the contents of science itself, some are shared by most disciplines, such as conflicts of interests with organizations which give financial support to specific research programs and, therefore, are concerned by the results obtained, or the pressure for researchers to publish. Regarding this last factor, the healthy idea that it is necessary to evaluate the quality of the work of researchers is having as a consequence the pressure on productivity. That pressure is clearly visible above all through the publication of results in scientific journals. Whereas in the past novel findings of theoretical interest were necessary ingredients for publication, now it is necessary for researchers to publish in order to continue an academic career, through economic incentives (e.g., the Spanish "sexenios") or promotion requirements (e.g., accreditations). This pressure has notably increased the presence of the questionable practices mentioned above, through the promotion of opportunism (DeCoster, Sparks, Sparks, Sparks & Sparks, 2015). The confluence between this pressure and the fascination with low p values has led to practices that are exclusively designed to maximize the probability of obtaining a p value below α. Many investigations, including some meta-analyses, are done because they are easy to carry out and publish, without contributing anything really new (Ioannidis, 2016). Often that apparent uselessness is masked with tricks that allow papers to be entered into the desired journals. The tricks usually include methods designed to produce statistically significant results (Lash, 2017).

The combination of these two groups of factors has led to a situation that could be interpreted as a real scientific bubble. From the meta-analytical perspective, this is not necessarily true. We believe that replication is not useless. It provides new data and evidence. Any repetition of a study is welcome. Even if research provides only new data for an established phenomenon, it can strengthen the body of evidence and facilitate building of a more complete theory of what we are studying. Our main tool is an estimate of the true distribution of effect sizes.

Looking at the future

What has been said so far might well transmit a more negative image than we actually have. It is true that psychology has already suffered several notable crises in the last century, about one every 20 years, from which no agreed and definitive answers emerged, but which show a curious regularity. Perhaps they are only peaks in the oscillations of a permanent crisis. It may be that living in crisis is in the nature of our discipline. It may be that each generation has to go through a crisis to lose its innocence. However, even at the risk of being naive we want to believe that from the current one, also 20 years later, something good can result, so we will point out some suggestions for the future that might help us to enter a new, more productive era.

First of all, we believe that progress involves changing from a competitive to a cooperative perspective. The implicit script of scientific success is to publish articles with ideas that change our way of thinking about an issue, having a decisive influence on their development (Giner-Sorolla, 2012). But this is only carried out by a small minority of scientists. Most of us simply collaborate, contributing with relevant evidence that develops the current theories and models. That is why it is convenient to think of our study as just another one that contributes to a collective task from which important results will only be seen over the years and after the accumulation of abundant and varied evidence. However, that change of perspective can only occur if there is an alignment between the objectives of science and the individual motivations of scientists. If the scientific community only values the contributions that involve revolution and not those that imply mere evolution (Ioannidis, 2016), scientists will try to publish crucial studies intended to play the role of significant steps. For this goal, many will not hesitate to employ questionable practices that end up distorting the evidence with which we build our knowledge. It is necessary that performing non-novel studies is attractive, making replication a valuable, honest and necessary work in itself, and not simply as an opportunity to publish and improve one's personal curriculum (Makel, Plucker & Hegarty, 2012; Nosek, Spies & Motyl, 2012).

Secondly, the change of perspective requires the implementation of practical instruments that favor the realization and publication of replicas that lead to results that produce evolution. Initiatives such as the *Registered Replication Reports* (*RRR*) are examples to follow (e.g., Hagger & Chatzisarantis, 2016). These initiatives begin by launching an open call to replicate a certain effect under a strict protocol. Several laboratories or research groups respond to this call. Only preregistered studies will be taken into account, and their authors commit themselves to contribute their results, whether or not they are significant. Therefore, bases of studies are formed free of publication bias. The techniques usually used to analyze the evidence produced in the *RRR* are those from meta-analysis (Blázquez, Botella & Suero, 2017). Once considerable progress has been made in the synthesis of the evidence published to date spontaneously, through retrospective meta-analysis, the time comes to design, program and perform prospective meta-analyses. These are genuine research programs based on cooperation among groups of scientists. In fact, the idea of the prospective meta-analysis is aligned with the recommendations made to improve the general reproducibility of the methods used in the quantitative synthesis. These recommendations include the previous registration, but also follow some information standards and share openly the data, making them available to other researchers (Lakens, Hilgard, & Staaks, 2016).

Finally, we must maintain constant feedback of scientific practices, by including the methodological aspects in the me-

ta-analyses. This healthy habit will allow us to identify questionable practices, maintain the methodological requirements and develop strategies to improve and to compensate for them.

Conclusions

It is time to return to the starting point. What can we say about meta-analysis 40 years later? In the first place, it has been consolidated as a methodology for the synthesis of scientific evidence and the analysis of heterogeneity of results. Second, the techniques have evolved and have become considerably more sophisticated, reaching a high degree of robustness and efficiency (e.g., Botella, Sepúlveda, Huang, & Gambara, 2013). Third, it has acquired an added function, not foreseen at birth, which refers to the supervision and analysis of scientific practices. The techniques have become essential tools to make an empirical study of the scientific practices themselves. Meta-analyses can detect anomalies, diagnose their origin and introduce elements with which to correct them. The paradigmatic example is publication bias, but other questionable practices can be approached from similar perspectives.

Fourth, we can envision the possibility of turning a problem into an opportunity if we adopt the meta-analytical perspective. The need to publish impressive results, with high levels of statistical significance, leads to sending reports to journals in which the results are forced and disguised as revolutionary when in fact they must be considered as being evolutionary. These trends have led to what looks like a scientific bubble. Nevertheless, maybe we can reconvert it if we change the perspective. The proliferation of scientific publications need not be a problem for science. If we accept that we need exhaustive replications we have to facilitate their re-

alizations and publication, recognizing it as merit where the scientists need it (Koole & Lakens, 2012). We must attune the objectives of scientists as individuals with those of science as a collective task. We can take the symptoms of the scientific bubble as a propitious occasion in the progress of knowledge construction. Meta-analysis has taught us that we cannot make a complete idea of any effect with a single study or with a very few. We need a critical mass of repetitions. We can turn the problem of pressure into publishing into an opportunity for systematic replication.

Even so, we must be aware that facing the alleged bubble does not solve the main underlying problem: the methodological and epistemological weakness of the behavioral sciences (Schmidt & Oh, 2016). We have to accept, simply, that our methodology is weak, learning to coexist with the uncertainty and ambiguity that it entails. In return, we can organize ourselves to conducting mass replications when useful or necessary. If we take advantage of the pressure to publish to launch projects of systematic repetitions of studies, we can end up transforming a problem into an opportunity. Observing the impact of works such as Hattie's (2009) on what counts in learning, we realize how small each individual research contribution is and how unimportant the results of most of them are when we look at them in perspective.

To add a final word on scientific perspective, our opinion is that we should not give in to the temptation to abandon rigor and exigency in the face of what seem to be excessive difficulties in reproducing any phenomenon. The verification of weaknesses in our discipline must have the effect of reinforcing the efforts to make psychology a credible science.

References

Arend, I., Colom, R., Botella, J., Contreras, M. J., Rubio, V, & Santacreu, J. (2003). Quantifying cognitive complexity: evidence from a reasoning task. *Personality and Individual Differences*, 35, 659-669. doi: doi.org/10.1016/S0191-8869(02)00243-X

Baker, M. (2016). Is there a reproducibility crisis? A Nature survey lifts the lid on how researchers view the 'crisis' rocking science and what they think will help. *Nature*, 533(7604), 452-455. doi: 10.1038/533452a

Bakker, M., van Dijk, A. & Wicherts, J. M. (2012). The Rules of the Game Called Psychological Science. *Perspectives on Psychological Science*, 7, 543-554. doi: 10.1177/1745691612459060

Blázquez, D., Botella, J., & Suero, M. (2017). The debate on the ego-depletion effect: Evidence from meta-analysis with the p-uniform method. *Frontiers in Psychology*, 8, 197. doi: 10.3389/fpsyg.2017.00197

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2010). A basic introduction to fixed-effects and random-effects models for meta-analysis. *Research Synthesis Methods*, 1, 97-111. doi: 10.1002/jrsm.12

Botella, J., & Eriksen, C. W. (1991). Pattern changes in rapid serial visual presentation tasks without strategic shifts. *Bulletin of the Psychonomic Society*, 29(2), 105-108.

Botella, J., & Gambara, H. (2006). El meta-análisis: una metodología de nuestro tiempo. *Infocop*, 29 mayo.

Botella, J., & Sánchez-Meca, J. (2015). *Meta-análisis en ciencias sociales y de la salud*. Madrid: Síntesis.

Botella, J., Sepúlveda, A. R., Huang, H., & Gambara, H. (2013). A meta-analysis of the diagnostic accuracy of the SCOFF. *Spanish Journal of Psychology*, 16, e92, 1-8. doi:10.1017/sjp.2013.92

Botella, J., Suero, M., & Ximenez, C. (2012). *Análisis de datos en psicología I*. Madrid: Pirámide.

Botella, J., Ximénez, M. C., Revuelta, J., & Suero, M. (2006). Optimization of sample size in controlled experiments: the CLAST rule. *Behavior Research Methods, Instruments & Computers*, 38(1), 65-76. doi: 10.3758/BF03192751

Botella, J., & Zamora, Á. (2017). El meta-análisis: una metodología para la investigación en educación. *Educación XXI*, 20(2), 17-38. doi: 10.5944/educXX1.18241

Cooper, H. M., Hedges, L.V., & Valentine, J.C. (Eds.) (2009). *The handbook of research synthesis and meta-analysis (2ª ed.)*. Nueva York: Russell Sage Foundation.

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

Cumming, G. (2014). The new statistics: why and how. *Psychological Science*, 25, 7–29. doi: 10.1177/0956797613504966

DeCoster, J., Sparks, E. A., Sparks, J. C., Sparks, G. G., & Sparks, C. W. (2015). Opportunistic biases: Their origins, effects, and an integrated solution. *American Psychologist*, 70(6), 499. doi: 10.1037/a0039191

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in psychology*, 6, 621. doi: 10.3389/fpsyg.2015.00621

Frick, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, 1, 379-390. doi: 10.1037/1082-989X.1.4.379

Gigerenzer, G. (1993). The superego, the ego, and the statistical reasoning. En G. Keren y C. Lewis (eds). *A Handbook for data analysis in the behavioral sciences: methodological issues*. Hillsdale, NJ: Erlbaum.

Giner-Sorolla, R. (2012). Science or Art? How Aesthetic Standards Grease the Way Through the Publication Bottleneck but Undermine Science. *Perspectives on Psychological Science*, 7, 562-571. doi: 10.1177/1745691612457576

Glass, G. V. (1976). Primary, secondary and meta-analysis of research. *Educational Researcher*, 5, 3–8. doi: 10.3102/0013189X005010003

Gøtzsche, P. C., Hróbjartsson, A., Marić, K., & Tendal, B. (2007). Data extraction errors in meta-analyses that use standardized mean differences. *JAMA, 298*(4), 430-437. doi:10.1001/jama.298.4.430

Hagger, M. S., & Chatzisarantis, N. L. D. (2016). A multilab preregistered replication of the ego-depletion effect. *Perspectives on Psychological Science*, 11(4), 546-573. doi: 10.1177/1745691616652873

Harlow, L. L., Mulaik, S. A., & Steiger, J. H. (1997). *What if there were no significance tests?* (pp. 37-64). Mahwah, NJ: Lawrence Erlbaum Associates.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of 800+ meta-analyses on achievement*. London: Routledge.

Hollenbeck, J. R., & Wright, P.M. (2017). Harking, Sharking, and Tharking: Making the case for post hoc analysis of scientific data. *Journal of Management*, 43(1), 5-18. doi: 10.1177/0149206316679487

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS medicine*, 2(8), e124. doi: 10.1371/journal.pmed.0020124

Ioannidis, J. (2016). The mass production of redundant, misleading, and conflicted systematic reviews and meta-analyses. *The Milbank Quarterly*, 94(3), 485-514. doi: 10.1111/1468-0009.12210

Ioannidis, J. P., & Trikalinos, T. A. (2007). An exploratory test for an excess of significant findings. *Clinical trials*, 4(3), 245-253. doi: 10.1177/1740774507079441

John, L., Loewenstein, G. F., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth-telling. *Psychological Science*, 23, 524–32. doi: 10.1177/0956797611430953

Klein, O., Doyen, S., Leys, C., Magalhães de Saldanha da Gama, P. A., Miller, S., Questienne, L. & Axel Cleeremans, A. (2012). Low Hopes, High Expectations: Expectancy Effects and the Replicability of Behavioral Experiments. *Perspectives on Psychological Science*, 7, 572-584. doi: 10.1177/1745691612463704

Koole, S. L., & Lakens, D. (2012). Rewarding replications: A sure and simple way to improve psychological science. *Perspectives on Psychological Science*, 7(6), 608-614. doi: 10.1177/1745691612462586

Lakens, D., Hilgard, J., & Staaks, J. (2016). On the reproducibility of meta-analyses: Six practical recommendations. *BMC psychology, 4*(1), 24. doi: 10.1186/s40359-016-0126-3

Lash, T. L. (2017). The harm done to reproducibility by the culture of null hypothesis significance testing. *American Journal of Epidemiology*, 186(6), 627-635. doi: 10.1093/aje/kwx261

Lipsey, M. W., & Wilson, D. B. (1993). The efficacy of psychological, educational, and behavioral treatment: Confirmation from meta-analysis. *American psychologist*, 48(12), 1181-1209. doi: 10.1037/0003-066X.48.12.1181

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in Psychology Research: How Often Do They Really Occur? *Perspectives on Psychological Science*, 7, 537-542. doi: 10.1177/1745691612460688

Nickerson, R. S. (2000). Null hypothesis significance testing: a review of an old and continuing controversy. *Psychological Methods*, 5, 241-301. doi: 10.1037/1082-989X.5.2.241

Nosek, B. A., Spies, J. R., & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science*, 7, 615–631. doi: 10.1177/1745691612459058

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), aac4716. doi: 10.1126/science.aac4716

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? Three arguments examined. *Perspectives on Psychological Science*, 7(6), 531-536. doi: 10.1177/1745691612463401

Pashler, H., & Wagenmakers, E.J. (2012). Editors' Introduction to the Special Section on Replicability in Psychological Science: A Crisis of Confidence? *Perspectives on Psychological Science*, 7, 528-530. doi: 10.1177/1745691612465253

Quiles-Marcos, Y., Quiles-Sebastián, M., Pamies-Aubalat, L., Botella, J., & Treasure, J. (2013). Peer and family influence in eating disorders: A meta-analysis. *European Psychiatry*, 28(4), 199-206. doi:10.1016/j.eurpsy.2012.03.005

Rosenthal, R. 1979. The "file drawer problem" and tolerance for null results. *Psychological Bulletin*, 86, 638–641. doi: 10.1037/0033-2909.86.3.638

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (Eds.) (2005). *Publication bias in meta-analysis: Prevention, assessment, and adjustments*. Nueva York: Wiley.

Schmidt, F. L., & Oh, I. S. (2016). The crisis of confidence in research findings in psychology: Is lack of replication the real problem? Or is it something else? *Archives of Scientific Psychology*, 4(1), 32-37. doi: 10.1037/arc0000029

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician*, 49, 108–112. doi: 10.2307/2684823

Tryon, W. W. (2001). Evaluating statistical difference, equivalence, and indeterminacy using inferential confidence intervals; an integrated alternative method of conducting null hypothesis statistical tests. *Psychological Methods*, 6, 371-386. doi: 10.1037/1082-989X.6.4.371

Yong, E. (2012). In the wake of high-profile controversies, psychologists are facing up to problems with replication. *Nature*, 485, 298-300. doi: 10.1038/485298a