John Hattie is Wrong

[Robert Slavin's Blog](https://robertslavinsblog.wordpress.com/author/sdavis1000/) [Effect sizes](https://robertslavinsblog.wordpress.com/category/effect-sizes/), [Evidence for ESSA](https://robertslavinsblog.wordpress.com/category/evidence-for-essa/), [Evidence of Effectiveness](https://robertslavinsblog.wordpress.com/category/evidence-of-effectiveness/), [Research measures](https://robertslavinsblog.wordpress.com/category/research-measures/), [Research methods](https://robertslavinsblog.wordpress.com/category/research-methods/), [Research use](https://robertslavinsblog.wordpress.com/category/research-use/), [Scientifically-based research](https://robertslavinsblog.wordpress.com/category/scientifically-based-research/), [What Works Clearinghouse (WWC)](https://robertslavinsblog.wordpress.com/category/what-works-clearinghouse-wwc/) June 21, 2018 6 Minutes

John Hattie is a professor at the University of Melbourne, Australia. He is famous for a book, *Visible Learning,*which claims to review every area of research that relates to teaching and learning. He uses a method called “meta-meta-analysis,” averaging effect sizes from many meta-analyses. The book ranks factors from one to 138 in terms of their effect sizes on achievement measures. Hattie is a great speaker, and many educators love the clarity and simplicity of his approach. How wonderful to have every known variable reviewed and ranked!

However, operating on the principle that anything that looks to be too good to be true probably is, I looked into *Visible Learning*to try to understand why it reports such large effect sizes. My colleague, Marta Pellegrini from the University of Florence (Italy), helped me track down the evidence behind Hattie’s claims. And sure enough, Hattie is profoundly wrong. He is merely shoveling meta-analyses containing massive bias into meta-meta-analyses that reflect the same biases.



Part of Hattie’s appeal to educators is that his conclusions are so easy to understand. He even uses a system of dials with color-coded “zones,” where effect sizes of 0.00 to +0.15 are designated “developmental effects,” +0.15 to +0.40 “teacher effects” (i.e., what teachers can do without any special practices or programs), and +0.40 to +1.20 the “zone of desired effects.” Hattie makes a big deal of the magical effect size +0.40, the “hinge point,” recommending that educators essentially ignore factors or programs below that point, because they are no better than what teachers produce each year, from fall to spring, on their own. In Hattie’s view, an effect size of from +0.15 to +0.40 is just the effect that “any teacher” could produce, in comparison to students not being in school at all. He says, “When teachers claim that they are having a positive effect on achievement or when a policy improves achievement, this is almost always a trivial claim: Virtually everything works. One only needs a pulse and we can improve achievement.” (Hattie, 2009, p. 16). An effect size of 0.00 to +0.15 is, he estimates, “what students could probably achieve if there were no schooling” (Hattie, 2009, p. 20). Yet this characterization of dials and zones misses the essential meaning of effect sizes, which are rarely used to measure the amount teachers’ students gain from fall to spring, but rather the amount students receiving a given treatment gained *in comparison to gains made by similar students in a control group*over the same period. So an effect size of, say, +0.15 or +0.25 could be very important.

Hattie’s core claims are these:

* Almost everything works
* Any effect size less than +0.40 is ignorable
* It is possible to meaningfully rank educational factors in comparison to each other by averaging the findings of meta-analyses.

These claims appear appealing, simple, and understandable. But they are also wrong.

The essential problem with Hattie’s meta-meta-analyses is that they accept the results of the underlying meta-analyses without question. Yet many, perhaps most meta-analyses accept all sorts of individual studies of widely varying standards of quality. In *Visible Learning,*Hattie considers and then discards the possibility that there is anything wrong with individual meta-analyses, specifically rejecting the idea that the methods used in individual studies can greatly bias the findings.

To be fair, a great deal has been learned about the degree to which particular study characteristics bias study findings, always in a positive (i.e., inflated) direction. For example, there is now overwhelming evidence that effect sizes are significantly inflated in studies with small sample sizes, brief durations, use measures made by researchers or developers, are published (vs. unpublished), or use quasi-experiments (vs. randomized experiments) (Cheung & Slavin, 2016). Many meta-analyses even include pre-post studies, or studies that do not have pretests, or have pretest differences but fail to control for them. For example, I once criticized a meta-analysis of gifted education in which some studies compared students *accepted into gifted programs to students rejected for those programs,*controlling for nothing!

A huge problem with meta-meta-analysis is that until recently, meta-analysts rarely screened individual studies to remove those with fatal methodological flaws. Hattie himself rejects this procedure: “There is…no reason to throw out studies automatically because of lower quality” (Hattie, 2009, p. 11).

In order to understand what is going on in the underlying meta-analyses in a meta-meta-analysis, is it crucial to look all the way down to the individual studies. As a point of illustration, I examined Hattie’s own meta-meta-analysis of feedback, his third ranked factor, with a mean effect size of +0.79. Hattie & Timperly (2007) located 12 meta-analyses. I found some of the ones with the highest mean effect sizes.

At a mean of +1.24, the meta-analysis with the largest effect size in the Hattie & Timperley (2007) review was a review of research on various reinforcement treatments for students in special education by Skiba, Casey, & Center (1985-86). The reviewers required use of single-subject designs, so the review consisted of a total of 35 students treated one at a time, across 25 studies. Yet it is known that single-subject designs produce much larger effect sizes than ordinary group designs (see What Works Clearinghouse, 2017).

The second-highest effect size, +1.13, was from a meta-analysis by Lysakowski & Walberg (1982), on instructional cues, participation, and corrective feedback. Not enough information is provided to understand the individual studies, but there is one interesting note. A study using a single-subject design, involving two students, had an effect size of 11.81. That is the equivalent of raising a child’s IQ from 100 to 277! It was “winsorized” to the next-highest value of 4.99 (which is like adding 75 IQ points). Many of the studies were correlational, with no controls for inputs, or had no control group, or were pre-post designs.

A meta-analysis by Rummel and Feinberg (1988), with a reported effect size of +0.60, is perhaps the most humorous inclusion in the Hattie & Timperley (2007) meta-meta-analysis. It consists entirely of brief lab studies of the degree to which being paid or otherwise reinforced for engaging in an activity that was already intrinsically motivating would reduce subjects’ later participation in that activity. Rummel & Feinberg (1988) reported a positive effect size if subjects later did *less* of the activity they were paid to do. The reviewers decided to code studies positively if their findings corresponded to the theory (i.e., that feedback and reinforcement reduce later participation in previously favored activities), but in fact their “positive” effect size of +0.60 indicates a *negative*effect of feedback on performance.

I could go on (and on), but I think you get the point. Hattie’s meta-meta-analyses grab big numbers from meta-analyses of all kinds with little regard to the meaning or quality of the original studies, or of the meta-analyses.

If you are familiar with the What Works Clearinghouse (2007), or our own Best-Evidence Syntheses ([www.bestevidence.org](http://www.bestevidence.org/)) or Evidence for ESSA ([www.evidenceforessa.org](http://www.evidenceforessa.org/)), you will know that individual studies, except for studies of one-to-one tutoring, almost never have effect sizes as large as +0.40, Hattie’s “hinge point.” This is because WWC, BEE, and Evidence for ESSA all very carefully screen individual studies. We require control groups, controls for pretests, minimum sample sizes and durations, and measures independent of the treatments. Hattie applies no such standards, and in fact proclaims that they are not necessary.

It is possible, in fact essential, to make genuine progress using high-quality rigorous research to inform educational decisions. But first we must agree on what standards to apply.  Modest effect sizes from studies of practical treatments in real classrooms over meaningful periods of time on measures independent of the treatments tell us how much a replicable treatment will actually improve student achievement, in comparison to what would have been achieved otherwise. I would much rather use a program with an effect size of +0.15 from such studies than to use programs or practices found in studies with major flaws to have effect sizes of +0.79. If they understand the situation, I’m sure all educators would agree with me.

To create information that is fair and meaningful, meta-analysts cannot include studies of unknown and mostly low quality. Instead, they need to apply consistent standards of quality for each study, to look carefully at each one and judge its freedom from bias and major methodological flaws, as well as its relevance to practice. A meta-analysis cannot be any better than the studies that go into it. Hattie’s claims are deeply misleading because they are based on meta-analyses that themselves accepted studies of all levels of quality.

Evidence matters in education, now more than ever. Yet Hattie and others who uncritically accept all studies, good and bad, are undermining the value of evidence. This needs to stop if we are to make solid progress in educational practice and policy.

**References**

Cheung, A., & Slavin, R. (2016). How methodological features affect effect sizes in education. *Educational Researcher, 45*(5), 283-292.

Hattie, J. (2009). *Visible learning*. New York, NY: Routledge.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77* (1), 81-112.

Lysakowski, R., & Walberg, H. (1982). Instructional effects of cues, participation, and corrective feedback: A quantitative synthesis. *American Educational Research Journal, 19*(4), 559-578.

Rummel, A., & Feinberg, R. (1988). Cognitive evaluation theory: A review of the literature. *Social Behavior and Personality, 16*(2), 147-164.

Skiba, R., Casey, A., & Center, B. (1985-86). Nonaversive procedures I the treatment of classroom behavior problems. *The Journal of Special Education, 19*(4), 459-481.

What Works Clearinghouse (2017). *Procedures handbook 4.0*. Washington, DC: Author.

**Photo credit: U.S. Farm Security Administration [Public domain], via Wikimedia Commons**

*This blog was developed with support from the Laura and John Arnold Foundation. The views expressed here do not necessarily reflect those of the Foundation.*

Samenvatting

* Waarom zulke hoge effectgroottes bij Hattie vraag Slavin zich af?
* Er worden meta analyses gebruikt met fouten in meta meta analyses die dezelfde fouten weerspiegelen. Hattie misleid de boel door van alles en nog wat te accepteren
* Conclusies zijn makkelijk te begrijpen en spreken aan
* Hatties hecht ten onrechte weinig belang aan geringe effectgroottes omdat het niet gaat om leerwinst van die leerlingen, maar leerwinst van die leerlingen tov van leerlingen in een controlegroep, waardoor .15 best wel interessant is
* Het gebruik van “zones” en kleurcodering van effectgroottes mist het punt van het gebruik van effectgroottes
* Hattie accepteert klakkeloos de resultaten van onderliggende meta analyses in zijn onderzoek en negeert bijvoorbeeld dat effectgroottes de tendens hebben hoog te zijn in kleine studies
* Het onderzoek naar feedback is vertekent doordat er onderzoeken in opgenomen zijn met een opzet die erg hoge effecten genereert (leerlingen 1 voor 1 trainen, ipv van als groep)
* Andere individuele studies laten bijna nooit effectgroottes zijn van .40, wat bij Hattie een gemiddelde is
* Slavin gebruikt liever een bruikbare studie met een effectgrootte van .15, dan het prutswerk van Hattie met een effectgrootte van .79 (feedback). Hattie moest er maar eens mee stoppen want zo krijgen we het onderwijs geen stap verder