**Crook:** I'm very pleased to introduce Matthew Inglis. He's professor of mathematical cognition at Loughborough University. We're going to talk about a paper that Matthew recently published with a colleague from York University. The paper is provocative, I think, because it questions the performance of a central method used in educational research. So I guess today we're going to discuss this challenge of errors, what it involves, how it came about and what we can take from it. I'll just say a little word of framing before we start a conversation. I guess research is usually concerned or said to be concerned with what is in the best interests of teachers and learners. Some people like to say researchers are particularly concerned with the question 'what works'? And a highly favored method for finding out what works in education is the RCT or the randomized controlled trial. In fact, the RCT is often called the gold standard for methods of finding out what works. So can I just start by asking you whether you think that's a fair claim regarding how our stand in the research community now?

**Inglis:** Well, obviously there's I think there's a variety of opinion in the research community, and some people are very sceptical of RCTs. But I think it's broadly fair to say certainly amongst funders, that people who focus on the efficacy of interventions in particular, certainly. I think it's fair to say that the majority view in that community is that acts are what you have to do if you want to find evidence of a causal relationship between some kind of intervention and some kind of educational outcome.

**Crook:** Yeah, okay. I think that makes the whole discussion very important. I think because what you said kind of corresponds to my reading of how things do stand. So as far as your paper evaluates this method, as we currently find it in use, can you kind of share with us what you took as a definition of an RCT?

**Inglis:** Yeah, I mean, to some extent, we devolve that responsibility to to the funders we were focusing on. So we looked at assets that were run by the EEF in the U.K. the Education Endowment Foundation. And the American equivalent was that the NCC and they both ran what we called rigorous large scale activities. So there's some I mean, they're pretty similar Research funders. They Have pretty similar things they fund, but there are some differences. But the general principle is that if you want to evaluate the efficacy of an intervention, you randomly split up a group of students, possibly by school, possibly by individual. It might depend on the nature of the intervention, but you randomly split into two groups. The randomization is important because it allows you to conclude causality at the end, and then you give one group the intervention, one group some kind of control activity, perhaps business as usual, or perhaps some other intervention. And then at the end, you, you give them some test or some measure of the educational outcome. And then you just see if they're different and if they are different, then you can try to conclude that the difference was caused by the intervention. And I think it's important because obviously that's just a, you know, a standard experiment, what I've just described there. But I think what makes RCTs, these particular rigorous large scale studies, distinctive is that they do that with very large numbers of participants. The EEF, in particular, are very, very rigorous in their research design. So they they insist that the analysis of preregistered, so they say exactly what they're going to do in advance. They have quite strict criteria about what type of test is appropriate. And they have an independent evaluator, so the person actually running the RCT is not the same person as the person who designed the intervention. So I think those features will perhaps distinguish RCTs slightly from the types of experiments that perhaps psychology researchers might be more commonly associated with. But in essence, it's the same research design.

**Crook:** Okay, and that's very clearly put because it reminds us that it is a sort of scaled up experiment, really, which is a method where perhaps many people listening to this will be familiar with. So, your paper is not considering the efficacy of any particular RCT study, but whether the go to method for what works is itself working. So I mean, that leads to what seems to be quite an intriguing question, which is what is the research method for evaluating a research method?

**Inglis:** Yeah, it's an interesting way of putting it, that question I guess. Yeah, I hadn't really thought of it in those terms before, but I think that is a helpful, helpful way of thinking about what we've done. I mean, what we did is we said to ourselves, Well, look. I mean, particularly in the UK context, the EEF is quite a new organisation, you know, it was set up in 2010 and 2011 or something, and it's rapidly become easily the biggest education research funder. You know, with orders of magnitude more funding than than other education research funders. So I think it's important given that context to think through. Has it has this way of so that essentially what's happened in the UK is there's been quite a significant change to the way education research is done or at least funded in the last decade. Now has that been a success? Is essentially the the question we asked or what's been the effect of that? So one way I think you can begin to approach your question is to think to yourself, Well, what was the stated goal? What's the stated goal of running all these studies has has the actual outcome met that stated goal? So in particular, we were interested in the point of a what works approach to education is is to provide information about whether and whether interventions are going to lead to more learning or not. So we were really interested in what proportion of trials produces that information as intended. Yeah. So I guess that's the way I think of it is the purpose of this method is set by the people who advocate for it. So really, our goal is as it's currently being implemented to think about as it's currently being implemented. Is it meeting that goal or not?

**Crook:** Okay, so your core piece of material then will be a large number of studies of this kind, which, as you've explained in terms of defining our RCTs, will be reporting differences. They have an intervention group, they have a, let's say, control group or business as usual group. So you'll be focusing on they will be focusing on differences and commenting on them, and you will be focusing on the way in which their focus is being managed. So I think what comes through here as a typical way of reporting differences of these kinds is something that's called an effect size. And here we have to get a bit technical, I guess, because I think I need to ask you to explain what an effect size is, but particularly how is it different from the more familiar significance level that people who know about experiments will be used to hearing reported as the outcome? So what is it about an effect size that is important?

**Inglis:** Yeah. So in effect, size. Well, I mean, there's lots of different types of effect sizes, but a simple way of understanding effect size is just an attempt to quantify the difference in performance between the two groups. So the group that gets the intervention and the group that doesn't. So, you know, the simplest possible effect size is simply the difference in scores between those two groups. You know, so if the mean on some test in the one group is 10 and the mean of the of the on the test of the other group is eight. You could just simply calculate the effect size as being too just literally the difference between the means. And that's what some people would call an standardized effect size. But most people, particularly in this community, prefer to standardize the effect sizes. So rather than report it as two items different on a test, which of course, is you really need to understand what the test is in order to interpret that. Some people like to divide those the difference by the standard, by some kind of pooled standard deviation. So some measure of spread in the in the groups. And because standard deviation and raw scores have the same unit that leads to to to what people call a standardized effect size, which is a unit less quantity that some people think makes is more comparable across contexts. Now that's a very controversial claim, which I would strongly disagree with. But but often people make the claim that you could compare standardized effect sizes between different studies much more easily than you can compare a standardized effect sizes, which is why people like to focus on on the standardized effect sizes such. So the one I've just described is usually called a Cohen's D or something similar to that. Now, in terms of your other question about how that differs from a significance level or a p value, they really give quite different information. So the effect size is an attempt to quantify the difference between the observed difference you see in the study. So it's a it's a quantity that is about about the people who took part in your study, whereas the significance level or p value is an attempt to generalize beyond the people who took part in your study. So you're trying to say to yourself, Well, look, I see what happened in my study. Does that tell me anything about the rest of the world who could have been in my study but words? So they're really quite different. And, you know, people make strong criticisms of significance levels because it doesn't take into account the size of the difference. But of course, that's not really what it's for. So I've always thought those criticisms were slightly unfair.

**Crook:** Ok. I mean, this sounds I mean, as it stands, a very useful metric because in terms of the way you've explained, it is a kind of single shared metric. Now, I wonder if you can anchor these numbers to something more familiar. I mean, for example, what kind of number effect size? This is just an example you may have your own, but what would promise a one grade difference in GCSE, for example?

**Inglis:** Yeah. So people do come up with claims of that sort. So one one commonly used one is based on comparing the performance of students of different years on the same test. So, for example, one commonly one claim that is commonly made is the difference between year five and year six students might be 0.2 or three. One study found it was 0.23 of a standard deviation. So, year six years education and maturation might lead to a 0.25 ish increase in scores as a standard deviation. So 0.25 Of a standard deviation or thereabouts. Now the problem is I I really don't approve of that kind of talk because it varies a lot depending on the type of test you're using and indeed the characteristics of your of your participants. So for example, because a standardized effect size is a raw difference divided by a standard deviation, you can kind of manipulate the size of it by changing the standard deviation. So, for example, there's lots of different ways of doing that. So if you've got a more precise measure, there'll be less noise in your so more precise test. Perhaps your test has got more more items, for example, that will typically make it a more precise test. Then that will reduce the noise in your research design, which will reduce the standard deviation, which will therefore increase the effect size. Similarly, if you do a study on a group of suppose, you're doing it in an education system that has a lot of ability grouping and you just happen to do your study in high ability. One group of in a high ability group rather than a mixed group that will reduce your standard deviation because there's a narrower range of performance and that will increase your effect size. So I think it's a mistake to try and make strong claims about translating between standardized effect sizes and sort of one GCSE grade or one year of education. I think it's really hard to do that reliably or validly. I try and avoid it as much as possible.

**Crook:** Yeah, I mean, your arguments make sense. The problem is, is that when you're addressing these kind of findings to practitioners, there needs to be a language in which the significance of a score of some kind, which is a naked number can be translated into something that's meaningful and something that's persuasive. So I mean, is there a way of doing that?

**Inglis:** So that's actually an unrelated paper, a paper that we're not talking about today, but Hugo and I, the co-author of the paper we are talking about and another colleague recently did a study on this actually where we compare different ways of reporting the results of our citizens. So one way, for example, is literally just to report the raw score on the test. And actually, that's not a silly thing to do at all, especially if your outcome measure is some tests that teachers are going to be quite familiar with. So, for example, if you use a key stage maths test, for example, teachers are pretty familiar with that and they they will well understand that, you know, a three point difference on the key stage two months test or whatever. You know, that's not a meaningless thing to say to a group of teachers at all. That's quite an insightful thing to say. And I think that's doing that is really the only way, the only valid way. You know, if you really want to make straightforward comparisons between experiments of different interventions, you just have to use the same test and then you have to report raw scores on that test. I think that's the only. You know, the only robust way of doing it. But that's I don't think that's a widely accepted point of view, but it seems to me to be the correct one.

**Crook:** But you accept the point that it's important to find a way of communicating the significance of these kind of interventions.

**Inglis:** I mean, to some extent, yes, I mean, I just think sometimes it might not be possible to do, and in that case, it's better not to try then to do it badly.

**Crook:** Okay.

**Inglis:** So yeah, I mean, I agree. If it's possible to have a meaningful way of communicating those sorts of differences, we should do it. But I don't think we should run away from the possibility that in some cases, it might just not be possible.

**Crook:** Okay, let's turn to your study and how you went about it scrutinizing these kind of interventions. First of all, I mean, you've characterized where the sample came from. It came from the funded projects from two major agencies, one in the US, one in the UK. Just a kind of aside, really. And it's an intuitive answer you probably have to give. But do you feel that the trials that you harvested in that way were representative of the kind of research that's going on? I mean, I seem to remember about 60 percent were primary schools, for example. So is that the state of play within the field?

**Inglis:** Yeah, it's a good question. I mean, in some sense, they were. I mean, it was, you know, you could think of it as not being a sample. You know, it was the whole population. We took every trial that these two funders had ever commissioned. Okay. But so in sum, so whether or not it's representative of the wider body of education, research is less clear, but it's certainly representative of what these influential funders want to fund and what they consider to be important. So there certainly is a bias towards English science and math outcomes, for example, and fewer on. There were a few other odds and ends, but so, you know, in some sense, it had to be representative if you consider it to be a whole population study. And I think these sorts of RCTs are quite rare. They require a lot of money. So I think it's unlikely that there's a large number of equivalent kind of RCTs going on in the UK and the US that weren't funded by these people.

**Crook:** Yeah, some, but not many. Yeah, that's why I didn't realize there was a sort of comprehensive survey, but that's pretty persuasive. So just move to the notion that you use a lot in the paper, which is informativeness, and some people might find that a little bit of a difficult concept to get their head around. I don't I'm not seeking a highly technical. I think it was going in the area of precision of measurement, maybe to more deeply than we're able to. But can you explain in accessible terms as you can, what informative, how informative is pinned down from this exercise?

**Inglis:** I can try. Let me let me try. So, yeah, I mean, it seems to me when you run an RCT, there's two possible outcomes that you're trying to distinguish between. Basically, you want to know whether the intervention is effective or ineffective. That's that's essentially what you care about. There's a large majority, really. You're running the study to answer that question. So you then you have to think to yourself, right? Let's try and work out what we mean by effective and ineffective. And ineffective is really straightforward, that just means there's going to be a zero effect size. So if you if you ran that on, if you ran the RCT with everybody in the world, you know the full population study. So every child and every possible child, you know, some some strange conceptualization like that, then you would end up with a with an effect size of zero. Now that's so that's quite straightforward to understand what that means. Then you have to think of what is what is it an effective intervention? What is it that need? And that's a bit harder to pin down. And the way that the Bayesian statisticians like to do that is they say, well, we can think about an effective intervention as having a true effect size. So the effect size, if you run it on everyone in the world from some distribution that goes beyond zero. So it's might have the true effect might be somewhere between point two between zero point zero five and zero point two, or perhaps a bit higher or something like that. So they define some distribution where so some range of possible effect sizes, which are non-zero. So what we meant by informative was simply, does the data collected in these trials allow us to distinguish with some confidence between those two possibilities? So normally, the way a Bayesian would do that would be to calculate something called a Bayes factor, which quantifies the level of evidence in favor of one of those hypotheses compared to the other one. And what's nice about doing that is you can find evidence both for the effective result, but also for the ineffective result. And what we said and followed various standard ways of doing this was we would consider three to one evidence, so evidence of three times more in favor of effectiveness over ineffectiveness or the other way round as being informative. And if the trial didn't give you three times more and more evidence in favor of one or the other, we would consider that to be ineffective. And that's a sort of fairly standard cut off that people in the literature have suggested for many years. So we would following fairly standard practice there. I'm not sure. Hopefully, that was not too technical.

**Crook:** Yeah, no. It's helpful. Okay. So let's let's move towards what your actual findings were from this exercise. So I think if I can without putting words into your mind, summarize the way I read it is that you find when looked across the piece that the typical size of effects is very small, let's say much smaller than the natural optimism that the funders probably had and maybe the researchers had. So. Is that fair that the outcome of scrutinizing effect sizes across a very large sample of projects is a very disappointing level of influence or impact?

**Inglis:** Yeah. So I would phrase it slightly differently. So I think that's correct. But so the average effect sizes is .06 or something like that, which is very close to zero. But you can also think about the precision of those effects. So not only were they quite small, but they would not estimate it very precisely. So if you calculate the confidence interval around those effects, the mean confidence interval width was something like .3 , as I recall. So the problem with that is because the effect is quite small, but sat within quite a large range confidence interval. The typical trial was consistent with both a zero effect, so consistent with the intervention being zero being ineffective, but also with an effect that most people would interpret to be a real positive, useful effect. So when we did the the informativeness analysis, we found that an alarmingly high proportion of these very large, rigorously conducted trials were not informative in that in that technical sense of the word, they couldn't distinguish between giving us evidence that the intervention was ineffective and that the intervention was effective. So there's nothing inherently wrong with finding small effects. Now if you run a study and you can reliably say, Well, actually, we've ruled out the possibility that this intervention was effective. That's a useful, interesting thing to find. The problem comes when you can't rule that out, and you can't rule out the fact that the result that it is effective, you've not really found anything at all. Right.

**Crook:** So observing is a kind of outsider, one is bound to feel this is a sobering kind of conclusion. Here are all these studies heavily funded projects with great optimism, probably that they will make a difference. And now we're finding that the effects, as reported statistically, are very modest now. Can I just put you one possibility? That is I've heard sometimes teachers say in talking about these things, they might say, Oh yes, it appears that there is no influence of this intervention. There has been no effect, there's been no impact. But I know that within my class or within my school, there are a small group of children for whom this has been a very powerful intervention. Now, to what extent is it realistic to think that within a large sample intervention there might be hidden these subgroups, or some contexts in which impact could be there? But but apparently not there?

**Inglis:** Yeah, no, that's a good question. And I think that's not a foolish thing to say at all. But what the conclusion I would draw is if you're. You know, if the intervention has been designed to, you know, the I guess it depends on what was predicted in advance, you know, if if in advance people were saying, well, actually this intervention is only going to help this small subgroup of people, then an RCT with a large group of people is simply the wrong way of evaluating it. You know, it's it's a pointless study to do. You should do a study that evaluates it just on the subgroup because you'd have much more. You'd have a much. I mean, you know, that's the hypothesis you're interested in. There are some. Obviously, you would be a little bit concerned if people find small effects. So conclude that this intervention doesn't work and then just go looking for subgroups because you'll always find some subgroups where apparently had an effect. But if that wasn't motivated in advance by some feature of the intervention, you would be. I mean, I would certainly be worried that actually you were just going to be finding some statistical noise rather than. So I think, yeah, absolutely. If there are interventions designed for subgroups, then they should be tested on those subgroups. Not not on bigger groups.

**Crook:** Yeah. Okay. Point taken, I suppose, because I'd like to move to, you know, the way in which you interpret this disappointment. We can call it that or this rather surprising, really poor level of performance by these methods. I think that came into my mind because I think it relates possibly to your first kind of hypothesis about what might be going wrong here. And that is in my example, there was a failure to anticipate that there may be contextual effects or there may be individual difference effects. So if that had been reasoned in advance, then you wouldn't have been, you know, subject to the risk of a post hoc discovery of it. So have I got it right about your first hypothesis about what's going wrong here, which is that really, these interventions have not been conceived on an adequate grounding of theoretical exploration or exploratory exploration of the thing being studied? Is that a fair?

**Inglis:** Yeah, exactly. I mean, so the way I conceptualize it is that you can think about education very, very simplistic model of education research. Well, maybe I should rephrase that very, very simple conception of the sort of path from which you can take some education research and have some impact. you might think of it as well you do some kind of basic research on how we learn, you know, some kind of more pure stuff on the processes of learning, and you develop a theoretical understanding of that which you then translate into some kind of design work to turn into some kind of intervention that you think it's going to help people, people, students and then you evaluate it. So what you're talking about is the way I think about it is maybe the theoretical grounding on which a particular intervention is based. Maybe the the theoretical understanding which is driven, it's design is just broad. You know, maybe it's not been the the literature is not sufficiently well developed or possibly worse. The literature might simply be wrong. There may be problems with the robustness of the research in that area. And of course, if you're designing interventions based on a flawed understanding of what's going on, probably they're not going to be very effective. And one way that could be certainly is the kind of subgroup contextual thing you mentioned. That's certainly a possibility. Maybe this way of thinking about how learning works works in some contexts and not others, but the underlying research is not studied that properly.

**Crook:** Right? Ok, so there's an invitation here to be less preoccupied with our RCts and to invest more in the kind of underpinning that would lead to, you know, not wasting money on a central. Okay, so that would be your hypothesis number one. Your second one again, help with this one. You talk about the notion of translation. So there's a risk a design flaw, if you like, that arises from mis-translating what is being discovered informally or through observation or through experimental work into an actual intervention. Now what does that mean that notion of translation?

**Inglis:** Yeah, I mean, I think I suffer from this. I think quite this always crops up. And so suppose I do some study where I find in a lab based study, I think I've got a pretty decent understanding of what's going on. So this happens to me a while back where some colleagues and I had done some work on applying self explanation training to advanced mathematics, and this is a very well researched area, which lots of people have done work on, and we applied it to a slightly new area and we did a lot of what I could loosely call lab studies found positive results. So then you think, OK, so I think I have this this idea, which will help students learn. But then of course, you've got to turn it into some product or some pedagogical activity or some something which can actually just be sort of rolled out in some way. You can give to, in this case, maths lecturers to use in their classes, or you can give to students and then this how would you do that right? And the answer is, in my case, I don't really know because it's not a skill I've ever developed. And I think, you know, obviously I've got ideas. I can, you know, we did some versions of this. We put some advice in some people who are writing textbooks, added things to their textbooks and so on. But it seems to me that actually turning a an insight from more basic research into an actual educational design which can be used in a Real-World setting is a is a highly skilled job and requires some. Certainly, it requires skills that I don't possess in abundance by any stretch of the imagination. But then I think when you think through who does, who does possess those skills, I think it's probably an undervalued. Aspect of education, I mean, it's not really the job of most, I think most education researchers would say that's not their main role, and probably most teachers would say that too. So then you think, Well, who is it? And we can all think of people who do that kind of work and do it very skillfully. But I think there aren't many of them, and probably they're undervalued by the wider community. And it seems to me it's quite plausible that one thing that's going wrong is you've got some basic research that's reliable. But as it's been turned into an intervention, sort of mutations of emerge that are going to lead it to be less effective.

**Crook:** Okay, that's very clear, but then the question that comes to mind, of course, is how do we change the infrastructure of research to bring together different types of intelligence? I mean, is it is it a question of the community of researchers responding to that?

**Inglis:** Yeah, I mean, it's a it's a good question. So people have made suggestions about this so that one of your colleagues, Hugh Burkhard, has been talking about this for many years about how what he calls engineering based education research is really undervalued because it doesn't lead to high impact journal papers and so on. So I think there are suggestions out there for what could be done about this, but it requires some sort of systematic change, I think. And it's also it's not totally clear to me that that is the universities are the best place to do that. And perhaps maybe they are. I don't know. But, you know, probably we could take some inspiration for how the kind of research into practice operates in other contexts in health or engineering or, you know, that kind of design work is not always done by the in the same place that the basic research is done. And maybe that's another reason

**Crook:** That I know. Well, I suppose universities might be the best catalyst for it or the best Platform, maybe.

**Inglis:** Yeah, I guess so. Maybe. I mean, it's to some extent this sort of thing is changing slowly. I mean, I guess you could see the sort of impact agenda that the government has had over the last five or 10 years, including in the research assessment processes we have here in the UK as being an attempt to try and prompt people to do more work of this type. I don't know how successful it's been.

**Crook:** Yeah. Okay. Finally, your third hypothesis, if that's the right way to describe it, but I'm a little less sure what you're getting at here. So it's about the trial design rather than the translation process. So is this just a question of sloppy researchers not being careful?

**Inglis:** I don't think I would characterize it quite like that. I mean, I think one thing I would say is, you know, in the UK, at least, the EU came along in 2010, 11 or whatever it was and rapidly increased the number of trials that were being conducted. So it's perhaps unreasonable to expect them to get everything right straight away, you know, and maybe you could see this paper as part of a process for which the quality of the trials is being improved slowly but surely. So I don't think it's a case of saying researchers are sloppy. But for example, suppose we have to accept that education interventions as as currently thought about are going to lead to small effect, small, standardised effect sizes. So I suppose that's just a fact. I think that does have quite significant implications for how you design trials because there's different things you can do about that one, one way you could you could kind of deal with this low d world. If you think about Cohen's d's are going to be low lower than we thought, perhaps. What can we do about that? But one way is simply to do fewer trials, but to make them much bigger. You know, if you just did half as many trials, but they twice as many people, that would help. I don't think that would actually be enough. That's the slight concern. So that's one way of trying to change trial designs. Another way would be to have more precise measurements. So rather than relying on standardized stage tests or other standardized tests which introduce a lot of noise, You could have more precise measurements. So, for example, suppose I'm developing some intervention that I think is going to help your seven students learn algebra. If I evaluate that with a year seven maths test, which is the sort of standard way of doing things at the moment. Then my effect is going to be much smaller than if I evaluate it with a algebra focus test because the test is going to have other things on it that are just not related to the intervention. Now, normally people in the education evaluation world would say something like, well, you shouldn't use developer to develop design tests because that inflates the effect size because they've chosen a test that's quite close to the intervention. And I think that is precisely the wrong way of looking at it. I think that's a desirable feature of a test because it's closer to the hypothesis that the researcher wants to test. I think this is going to improve on children's algebra learning. My test should be about algebra. It shouldn't be about maths. That's another way of changing the way trials are run, which doesn't require increase in samples, but would lead to higher effect sizes. So it would lead to a more precise way of estimating estimating whether or not these these interventions are effective or not. So it would be more informative. Now, the reason people are sceptical of what some people are skeptical of that suggestion is you get you get further and further away from a kind of metric that teachers are familiar with. So if I'm not using a key stage test about the whole of maths, I'm really quite a long way away from a meaningful and meaningful outcome measure that teachers can engage with. I think that might just be a price that we have to pay in order to have affordable trials. But I mean, it's a discussion that could be had about that.

**Crook:** Yeah, no. I certainly personally find that quite convincing. Ok, so just looking back on those three. Responses to your low d or low effect size outcome, do you have a favorite I mean, do you have one that you think is one that needs particular attention?

**Inglis:** I'm not saying no, I don't think they are mutually exclusive. I mean, I think all of them are true to a greater or lesser extent.

**Crook:** They're equal in their significance. Or, you know, you wouldn't put one above as needing urgent attention.

**Inglis:** I think they probably all need urgent attention. I mean, I think it's certainly true that there is a problem with the quality of basic research. And I think the replication crisis in psychology has shown us that over the last decade and a half. So I think it's not implausible to suggest that there are serious mistakes in the literature, which probably have an impact. So I think that's a definite problem. But I think people are dealing with that slowly but surely. I think methodological reform is helping there. I think. The design, encouraging design, high quality design research is is a problem, and I think that's a hard problem to get around, but we should certainly try and do something of that, and I definitely think that trials should be designed. You know, there's now, I think, enough evidence that the current way that trials are being run is not a good use of money. So some methodological reform is needed, and I think there needs to be a community discussion as to the best what what the best way of of dealing with this problem is. And I think that discussion is quite important that it should be happening.

**Crook:** Yeah, I mean, just to pick up on the sort of instrumentality here, I mean, you mentioned the funders and I think you hinted at the need to influence funders, but what strikes me about this is the sheer investment that's been targeted on our seats. I mean, I think I read somewhere that the typical RCT costs about half a million to run of the kind that is funded from the sources you mentioned. And so this is a big investment and and yet I read your results as being sobering in relation to those costs. Are you getting a response from the funders? I mean, I know your papers quite widely cited. Are they reading it and are they reacting to it?

**Inglis:** Yeah. So the EEF. wrote a response where they made several points, which I mean, I wasn't massively convinced to be honest by their response. So I think they made some. I mean, you know, I absolutely do not want this paper to be seen as an anti eef piece because it's really not that. I mean, indeed, the EEF in some ways are an absolute model of how a funder should insist their researchers behave because they're completely open and upfront with all of the information, you know, be very hard to run this study with. If I wanted to do the same study with the research funded by the other, by most other funders, it just wouldn't wouldn't be possible because they don't insist on the same level of transparency in the same way. So in some, in some cases, in some sense, the EEF really are an excellent organization in terms of improving the quality of research. But yeah, I wasn't totally convinced by their response. One of the points they made was that the size of their trials is getting bigger since they started. Now that's true, but the level of informativeness is not improving, and that might be some issue to do if they start it off by by evaluating all the kind of low hanging fruit and as they've moved through the through the years, okay, their trials have got bigger, so you would have thought the informative ness of them would improve. But if they're studying more ambiguous interventions, then maybe that counterbalances. They also made the point, which I think is fair, that their trials often have what they call process evaluations at the moment as well. So it's not just a case of we do the trial, we find the single D, and that's the end of the trial. They also go and observe how the thing was into what actually was happening in the classroom and in the school and so on. And you can learn things from doing that, too. I think that's right. But the reason for running and asked is to get an estimate of impact. It's, you know, the rest is a secondary outcome. So I think that's true, but I don't think that's A.. But, you know, I've been quite encouraged by the level of engagement with the paper. And I think people who do asked these, you know, people who are commissioned by the EEF to run these sorts of assets have in general recognize that there's an issue. These are the ones that have engaged with me. Maybe maybe the ones that don't agree of just ignored it. But you know, there's been plenty of positive engagement. And I think a recognition that we should think about this as a community.

**Crook:** Well, that's good. I am just by way of closing. I just sort of wonder what your views are about the extent to which one would call this a kind of crisis or in terms of what works, not working. I mean, at the beginning, I reminded us that people often talk about the RCT as the gold standard of research. So if we pursue that metaphor, are we dealing with? Are we dealing with this gold as being no longer a viable underpinning for the research infrastructure? Or are we dealing with a kind of tarnished project as it were?

**Inglis:** So I think the way I would see it is that the balance is gone a bit wrong. So because of the way the lot is in the UK, so if you think about the UK funding landscape is so heavily, the balance of funding is so heavily pushed towards assets at the moment. You know, if you compare the amount of money that's invested in education research through a standard funder, it's so much smaller then than is invested through the ETF. And to some extent, this is not the know they were given a mandate. And they were given a large amount of money, and they're carefully getting on with what they were, what their remit is. The problem is that there's just not much funding for the other stuff, you know. So in a world where there is ample education research funding, I think the EEF, I think they should tinker around with their methods and so on to make the seats more informative. But I don't have a problem with them carrying on doing that. I think that's a good thing. Carry on doing RCTs. I think the problem comes is when you've only got seats being funded to a large level of money, then you do have this real supply problem and you end up looking at some of the things that go through our seats and thinking, Well, there's no way that's going to work because I know a bit about the background literature that that led to that intervention. And it's problematic for this reason this week and this reason. What should it be happening there is that there should be money available to someone to tidy up the literature and do some small scale investigations of the theoretical, the proposed theoretical mechanism, and that should happen before it ever gets to trial. And the problem, I think, is that there's no funder who has that remit. So I think the system is a problem, but I wouldn't be in favor of just abolishing the notion of an act altogether. I think I think one, when done well, they're really useful and important, but

**Crook:** They do exist. I mean, it's fair to acknowledge this or to ask you whether you agree with it being acknowledged that they do exist within a bigger political infrastructure of what we might call evidence based education. And it's fair to say, is it not that there are some critics within the discipline who are fairly final in their judgment? These are inherently a bad thing. I take it from what you said, you have no sympathy with the extreme kind of skepticism about the value of this method.

**Inglis:** Yeah, yeah. I think absolutely RCts are useful when they're done well and provide useful when they provide information and value to the community. They're useful. It's almost a tautology. I think the problem comes when when you get this extreme imbalance so that they dominate over all other forms of research and of course, they fit within an ecosystem that they're only ever going to be operating well when you have, well, evidence interventions that have been well designed. Going to trial if you if you've got so much money to run trials, the interventions that that people are proposing and not, you know, being trialled too soon. Trials aren't going to be useful. So I think you've got to think about the whole ecosystem rather than just the asked part.

**Crook:** Yeah, I mean, I guess there's a comparison to be made in the current climate, particularly to medicine. And I do wonder sometimes whether those of us who might express a scepticism about our seats in education are very anxious to have our seats when we take a vaccine or some other medical intervention. And so maybe there's a certain kind of inconsistency in some people's judgments. Asked fits in some domains, but they don't want it in others.

**Inglis:** Yeah. Well, though one interesting, interesting thing about RCTs in medicine is it seems to me that they're much easier to do than they are in education, actually, because especially the measurement outcome, you know, I am no medic So I perhaps medics, if any medic watches this would be shouting at me and denying this. But you know, it seems to me that the outcome measures are much easier to measure. I mean, certainly in a vaccine case, you know, people get the get the get the disease or they don't. Yeah. So these kind of precision of test instrument issues don't appear to me to crop up to the same extent. So I.

**Crook:** Yeah, so perhaps the danger there from your your comparison is that people too readily extrapolate the ease of value for our in medicine or a field like that to something as subtle and nuanced as education. Perhaps that's where you need to be careful in making that, that that link too tight.

**Inglis:** Yeah, I think that's right. And you know, but absolutely I would I wouldn't want this paper to be seen as some anti RCT position, that's certainly not not the view I adopt. It's just a case of hopefully people seeing it as a constructive attempt to analyze how assets are currently being used with the aim of improving them, right?

**Crook:** Well, that's a very good point to close down. So can I thank you very much for indulging this conversation? I found it very interesting, and I'm sure many people watching it will also. So thanks very much.