More on Worms

Macartan Humphreys, Columbia University

1 August 2015

It's been a disturbing thing watching the spat this week between economists and epidemiologists over the Miguel-Kremer Worms paper (MK04) and the fraught attempt to replicate it (ADHH). Insults were flung, axes ground, numbers crunched. Journalists took exaggerated positions. Researchers gave contradictory interpretations of the same results. In their overview of their replication results ADHH claimed that they found little evidence of positive indirect effects of a worms intervention on school attendance, Miguel thanked them and noted that in fact the re-analysis "strongly supports the finding that treatment improves school attendance of both children who are treated and those who are not."

Strange times indeed. Some saw hope. My colleague Chris Blattman said that a "hive mind" quickly sprang into action and that the truth (which more or less confirmed the claims of the original paper) was rapidly reached. I was more confused though; it looked from the outside a bit more like a circling of the wagons. Many economists blogging and tweeting about the dishonesty of the epidemiologists did not seem to engage much with the data (Ozler's second post is an exception). Some even said that the criticisms just made them more convinced by the study. One of the most informative responses, by GiveWell, a group that has been endorsing deworming, seemed to say largely that they were not too worried because their support for deworming does not really derive from MK04 after all. With a little distance, many of the arguments produced by the epidemiologists seemed reasonable enough and some of the vitriol that they were met with seemed out of proportion. Moreover the main issue highlighted by journalist Goldacre about the weakness of evidence for externalities seemed worth paying attention to.

To make sense of all this I went back to the data to look at what I think are the core analyses. MK (and adding coauthor Joan Hicks, HMK) have made going back to the data very easy. Miguel is a pioneer of the transparency movement in social science and the amount of work that has gone into documenting the very many analyses around this paper is extraordinary.

A quick summary of what comes next is that (a) Miguel et al did terrific work (b) but ADHH still highlighted important issues that go to the heart of the most important claims of the paper (c) the attempt to rescue these results is on somewhat shaky ground (d) oddly the entire discussion and claims from all sides are based on a model that we now know might not be well suited to capture the effects of interest in the first place. Looking ahead there's going to be a lot more of this error-finding and claims-contesting and we have to get better at recognizing errors quickly and put learning before allegiances.

The Issues

MK04 was written at a time when there was less clarity than today about experimental design and the analysis of experimental data. It was at the cutting edge at the time and in many ways it continues to stand up well. It's detailed, thorough, and innovative. Even still, two developments over the last decade that date it a little are worth discussing. The first is the growth in awareness of "**multiple comparisons**" problems and the related problem of data fishing. The second is greater appreciation of the role of **randomization** for justifying inferences from experimental data. Ted Miguel and the broader JPAL group have played a central role in the first advance; some of their travelling companions such as Angrist, Imbens, and Rubin have played a central role in the second.

As methodological issues become clearer over time, work that once seemed unassailable can suddenly start to look more rocky. Some of that is happening here.

Externalities and Multiple Comparisons

Let's look first at the multiple comparisons problem. Basically the multiple comparisons problem is the problem of figuring out what to believe when results are found on some measures but not others. There's now more general agreement that you run into statistical problems when you ignore the non-significant findings in a given analysis and bank on the significant ones. We'll see in a moment why this matters.

The "pure" replication implemented by ADHH dealt a blow to what I take to be the core claim of the original paper. It highlighted a mistake and it looks like the mistake matters (in the first draft of this comment I had written that they ADHH *found* the mistake, but it seems that this mistake has been known by MK for a long time; see here[1]). The mistake was a small coding error that meant that most schools between 3-6km of a given school were not counted in the variable used to measure schools potentially producing spillovers.

Here is why the mistake found here seems to matter. I take the core claim of the paper to be the following, taken from the abstract:

> Deworming substantially improved health and school participation among untreated children in both treatment schools and neighboring schools, and these externalities are large enough to justify fully subsidizing treatment (p 159)

The first paragraph of the conclusion also ends with the claim:

> A rough calculation suggests that these spillovers alone are sufficient to justify not only fully subsidizing deworming treatment, but perhaps even paying people to receive treatment. (p 208)

[1] MK 2007 does give corrected tables though it describes differences such as the loss of significance on treatment in period 2 and the loss in significance for spillovers in the 3-6 km range as slight changes, and that the new tables show "little substantive differences" and "almost identical" results.

In short, the paper saw spillovers (or "externalities") as a big deal and important for the cost/benefit calculations made in the paper. The paper is cited and taught in large part because it is one of the first to really take the analysis of spillovers seriously. ADHH cause trouble because they find a flaw that affects the reported results and then refuse to engage in *ex post* test selection to save the result.

Here is the gist of the problem. Using the corrected data and models used by MK (available here), the Table below shows what are from my read the most important marginal effects for estimates of externalities in the 0-3 km range around schools as well as in the 3-6km range for the primary outcomes (as reported in Column 1 of Table VII and Column 3 of Table IX of the original paper). The shaded rows are the published results, the unshaded rows are the corrected results. You will notice that every single number changes.

Table 1: Core Results; Original and Corrected

I	II	III	IV	V	VI	VII	VIII
Outcome	Direct Effect (sd)	0-3km marginal effect	3 – 6 km marginal effect	Estimate	Source	Linear combination of coefficients in cols III and IV (*p* value)	*p* value from test that both coefficients cols III and IV are 0
Any moderate-heavy helminth infection, 1999	-0.25 (0.05) *p*<0.001	-0.26 (0.09) *p* = 0.004	-0.14 (0.06) *p* = 0.022	MK04 Original	MK 2004, Table VII Col 1	-0.228 *p* = 0.001	*p* =0.004
Any moderate-heavy helminth infection, 1999	-0.31 (0.06) *p* < 0001	-0.21 (0.10) *p* = 0.043	-0.05 (0.08) *p* = 0.515	MK14 Corrected	MK: PSDP-REP_2014-11.pdf p31 Tab A7 Col 1	- 0.151 *p* =0.184	*p* =0.126
Average individual school participation	0.060 (0.015) *p*<0.001 0.034 (0.021) *p* = 0.111	0.044 (0.022) *p* = 0.044	-0.014 (0.015) *p* = 0.339	MK04 Original	MK 2004, Table IX Col 3	+0.020 / +0.017 *p* =0.14 / *p* =0.24	*p* =0.122
Average individual school participation	0.062 (0.014) *P* < 0.001 [0.033 (0.021) *p* = 0.121]	0.040 (se=0.022) *p* = 0.075	-0.024 (se =0.015) *p* = 0.102	MK14 Corrected	MK: PSDP-REP_2014-11.pdf p33, Tab A9 Col 3	-0.017 / -0.016 *p* =0.58 / *p* =0.61	*p* =0.022

Note: See code in **Annex A** for replication. MK and ADHH appear to agree on most of these numbers. Indeed ADHH suggest that MK uncovered many of these discrepancies in 2007. The final column is new and tests the null that both spillover effects are zero. The first column in the last two rows has numbers for the direct effect for each year of intervention separately. The last two rows on the second to last column contains two sets of numbers, one following MK and the replicators, the second showing my inferences which use Model 3 of Table IX.

Note that for this table I am relying on the data cleaning and aggregation and so on as provided by MK and I am using their models (though it looks like both sides agree on these numbers now).[2] At the end I provide some relatively compact STATA code to run all this directly. Here I simply synthesize, provide a little more information from their own analyses, and show the *p* values from two tests for each analysis. [Recall a *p* value is the estimated probability that the pattern seen could have occurred by chance; by convention results with *p* values about 0.05 are considered "not-significant."]

The first is the test for the hypothesis that the overall spillover effect is zero (Col VII). This overall between-school spillover effect is estimated by MK and ADHH by multiplying the spillover effects by numbers of students in treated schools within the 0-3km and 3-6km bands.

[As an aside, although HMK and ADHH seem to agree on column VII here, unless I misunderstand the underlying model, by my calculations these numbers *underestimate* the spillovers by a factor of two or three since they add up the effects of externalities for the directly treated schools only instead of for all schools exposed to the externality.[3] If I am right, the increases in externalities that would be estimated after fixing this mistake would more than offset (in a substantive but not statistical sense) the losses in some specifications from the variable definition mistake.]

The second test, which I will come to later, assesses the hypothesis that both the coefficient for the 0-3km and 3-6km effects are zero.[4]

What's the bottom line? Essentially if we stick to the MK's original inference strategy, things don't look so good for the externalities claims. The strategy in the paper was to assess spillover effects within 0-3km and within 3-6km and then add them up for a total externality effect. For the health results the original analysis had significance for this overall externality effect but that gets lost with the corrected data, even at the 10% level. Things actually never looked too good for the education externalities. The sum of externalities was reported as 0.02 with a standard error of 0.013; in the paper the authors appear to say that this is marginally statistically significant though it is not significant by conventional standards (the *p* value is around 0.14 using the author's approach) and even farther from it using my calculations using model 3 in the original text (this gives a *p* of 0.24; code below). In any event, this jumps up to 0.61 with the corrected data. Note too that even the coefficient on the spillovers in the 0-3km range is no longer significant at the conventional 5% level.

So what to infer? Here we come to the multiple comparisons issue. There are two things you might do now.

---

[2] Although there are some modest differences between these numbers and some of the numbers I have seen HMK present (see in particular their useful discussion around Table 1 in the 3ie responses here), I don't think they will disagree much with these numbers below.

[3] I think this must be a mistake, in this case one that *underestimates* the magnitude of spillover effects overall by a factor of two (for education) or three (for health). Note that although the abstract and discussions talk about the spillover effects on untreated students, it looks like the results in the original Table VII, model 3, do not support this; the spillover action appears to be operating largely for the *treated* students. Also note that estimates of uncertainty of the spillover effects do not take account of uncertainty around the number of neighboring students.

[4] Technically the two tests used are chi-squared tests and *F*-tests respectively.

You might stick with the inferences from the strategy that appeared reasonable until the errors were found. Even then, the results of the pure analysis show strong, in some cases stronger, direct gains from deworming. These direct effects seem to be quite solid. The discussion by Ozler at the World Bank clarifies that a number of the critiques from the re-analysis (not to be confused with the "pure replication") are not as strong as they seemed at first (see robustness code below also). So you could take the hit on the indirect effects and do the re-calculation and figure out what the policy lesson is.

Or you might start looking to other models and measures to support your original claims. MK seemed to opt for the latter, preferring a new specification in which the 3-6km variable is dropped. Note that if you adopt this strategy it means accepting a nearly 50% drop in the size of the estimated externality effect. Not a small drop, and a correction worth highlighting. AMHH basically refused to accept this kind of strategy however. In refusing, they were following a growing trend to resist the temptation to hunt for positive results in large datasets.

Once you start changing strategies after seeing the results, there are many ways to go. That's the problem in a sense. MK prefer to drop the combined analysis and look only at the significant results for the 0-3km estimates. That's a hard sell these days, thanks in part to other work by Miguel. Would they have done the same had there been significant results at 3-6 km (where there is more data) but significance was lost at 0-3km? Perhaps. If so, then choosing the significant result is stacking the deck for positive findings. The point is a little subtle: in focusing on 0-3km, MK are not simply pointing out that a subset of the spillover results survive, they are changing the estimation strategy and the estimand.

One simple way to account for the multiple comparisons implied by the strategy is to test the hypothesis that there are no spillover effects at *either* distance. That's the test in the last column of the Table above. Surely if there are spillovers, even if they are only over a short range, that hypothesis will be rejected. Using this test one can easily reject this null on the erroneous data for the health outcome but you cannot reject it with the corrected data. For the educational outcomes the null could not be rejected even with the original data. With the corrected data the null of no externalities is rejected, though awkwardly only because the evidence for *negative* externalities is moderately strong. One could also use various corrections for multiple comparisons but these would also be punishing. One could also assess the effect of spillovers in the 0 - 6 range combined. That approach (in the code below but not in the table above) doesn't produce good news for spillovers either though.

Of course, if that's the game then as well as focusing on the positive you could also start focusing on the negatives. What is with that large and very strongly significant adverse effect on attendance in the 3-6km range in column 5 of the updated school attendance tables for example (Table A9 / Updated Table IX here)? Or you could start to rethink the analysis more generally. Why is a linear model used again? Why aren't distinct spillover effects for each year estimated? Why are total spillover effects calculated for treatment units only? And so on. Or maybe rethink the approach to estimating

spillovers more fundamentally. There might be reasons to do that and it's the second issue I wanted to get to.

Randomization and the Estimation of Spillover Effects

The second issue is around randomization and causal inference. At many points the worms paper describes a randomized intervention and random assignments to treatment, but in fact as pointed out by ADHH it seems that there was no actual randomization. Rather a sort of alternating procedure was used. How problematic this is depends on details of the ways units happen to be ordered.[5] As a practical matter it makes it hard to employ design-based inference – such as randomization inference, or introducing propensity weights – without making some form of as-if random assumption regarding the list order.

Perhaps more interestingly though, even if as-if randomness is assumed, the analysis itself does not make use of the experimental design to estimate the spillover effects. Experimental estimates come from control over assignment to treatment. In the experiment, the researchers had direct control over assignment to deworming; but they did not have the same control over indirect exposures. Indirect exposure, or spillovers, could work in many ways, but even if they worked in the simple way assumed in the paper (as a function of geographic distance) the probability that a given school is assigned to spillover effects is a function not just of the treatment assignment procedure but also of features of the school, in particular how many other pupils are in schools close to that school.

This produces an element of *self-selection* into the spillover treatment. Regression analysis, even regressions that controls for the number of students nearby, and interactions between numbers of students and treatment, does not take account of this self-selection problem very well and can lead to biased results (to get a sense for why this is see the example in **Annex B** below): This is now textbook stuff (it is discussed for example in Gerber and Green's textbook on field experiments; see also the nice treatment by Aronow and Samii). But interestingly the replication team did not question the reliance on regression and they employed the same kind of linear model implemented by MK.[6]

I hope someone will do that design-based analysis properly. When they do it they will have to grapple with dosage issues, which in this case is very important but also tricky as the spillover treatment varies in dosage along at least three dimensions (numbers treated, where they are located, and when they were treated). In the MK treatment, linearity was assumed throughout. But with an increase in

[5] You can see the problem immediately if you imagine lining up a set of married couples and then selecting every second one. You might just end up with only men or only women in treatment.

[6] Note that *share* treated, unlike the number treated, can have uniform propensities under some conditions. For example if all schools had the same number of neighboring schools even if these had different numbers of students. However a school with many neighbors is less likely to have none or all treated than a school with few. As highlighted by GiveWell MK do look at shares as the treatment variable in an Appendix (Table AIII, col 3 for the primary outcome). Not highlighted by GiveWell or MK is that in these tables the 3-6km results are not there (though they are present in Col 7). Interestingly the original analysis focused on the 12 nearest schools within 6km, if indeed all schools had 12 schools within this range then this might have reduced heterogeneity in treatment propensities.

interest in understanding scaleup, the bigger gains may be in understanding how externalities depend on *concentration*. After all, the policy lesson here is not to deworm a third of the world and bet on the externalities for the remainder. Rather what becomes important with scaleup are the externalities *on directly treated* groups when there is wide local coverage. The MK data could speak to this but it would a call for a very different analysis.

What We Learned About Science

Besides the implications for the externalities from a deworming intervention in 75 schools in Kenya two decades ago, there are at least two important lessons from all this. One is about the pressures we place on the results of single papers. The second is about how we handle the discovery of errors as scholars and how we engage across disciplines.

MK04 has been a tremendously important paper and the evidence for positive direct effects, at least, seems to hold up quite well. But it's just one paper, studying one place. The scandal is that so much weight has been placed on this one paper. It has sometimes seemed as if this whole discussion has been about the merits of deworming, but really it is just about one paper. The replicators highlighted many errors in the paper. We can expect similar scrutiny of other papers will also find many errors. Some level of error is probably avoidable. The critical point though is that risks of error makes dependency on one paper even more risky. When inferences for policy are drawn from many independent studies, there is some hope that these sorts of errors will wash out. *So, another reason to invest seriously in more field replication.*

Placing so much weight on one paper is also bad for scholarship. The rallying to defend this result in this case produced mud-slinging, a hardening of positions, and deafness to the claims and arguments made by the epidemiologists. So we learned, if we didn't know it already, that criticism is hard to take. In this case the criticism was slow coming (in fact it took a grant from 3ie for many of these issues to get attention). And the criticism was not very welcome when it came. It would be a pity if this kind of response deters (pure) replication exercises like this in the future. A better response perhaps would have been more curiosity about the standards and expectations of epidemiologists, perhaps gratitude that they found out that the core claims on externalities seem not as strong as we once thought, and redoubled efforts to reassess the importance of the externality claims for the implications of this paper.

I expect that as data becomes more readily available and people get in the habit of replicating, mistakes like this will be found in many of our favorite papers. None of us will be spared. As this happens we need to see these as opportunities to advance knowledge and not as attacks on individuals, let alone disciplines.

Annex A: Code for Core Results

For simplicity the STATA code below re-runs the main analyses from compiled versions of the data created by running code on rawer data by MK on dataverse. To re-run it should be enough to paste it into STATA.

```
*  Core results from MK04

** T7 ORIGINAL
 use http://www.columbia.edu/~mh2245/w/MK_T7_original.dta, clear
 global x_base = "sap* Istd4-Istd9 mk96_s"
 mean  pop1_3km_original popT_3km_original pop1_36k_original popT_36k_original
[pw=indiv_weight] if (wgrp ==1 | wgrp==2)
 dprobit any_ics99 wgrp1 pop1_3km_original pop1_36k_original  popT_3km_original
popT_36k_original  $x_base [pw=indiv_weight] if (wgrp==1 | wgrp==2), robust cluster(sch98v1)

* Externality test (using numbers from paper); extracted below
 display (-454*.2557  -802*.14)/1000
 test 454*pop1_3km_original  +802 *pop1_36k_original = 0
* Joint test
 test pop1_3km_original pop1_36k_original


** T7 UPDATED
 use http://www.columbia.edu/~mh2245/w/MK_T7_updated.dta, clear
 global x_base = "sap* Istd4-Istd9 mk96_s"
 mean  pop1_3km_updated popT_3km_updated pop1_36k_updated popT_36k_updated  [pw=indiv_weight] if
(wgrp ==1 | wgrp==2)

 dprobit any_ics99 wgrp1 pop1_3km_updated pop1_36k_updated  popT_3km_updated popT_36k_updated
$x_base [pw=indiv_weight] if (wgrp==1 | wgrp==2), robust cluster(sch98v1)

* Externality test (using numbers extracted above) [MK use di (448*0.21 + 1108*0.05)/1000]
display (-448*.2125  -1108*.0504)/1000
test 448*pop1_3km_updated  +1108*pop1_36k_updated = 0

* Full accounting of externalities - not just on the treated
display (-1332*.2125  -3338*.0504)/1000
test 1331.6*pop1_3km_updated  +3338*pop1_36k_updated = 0

* Joint test
 test pop1_3km_updated pop1_36k_updated

* Robustness check: Combine ranges
g pop06 = pop1_3km_updated + pop1_36k_u
g pop06T = popT_3km_updated + popT_36k_updated
dprobit any_ics99 pop06 pop06T wgrp1 $x_base [pw=indiv_weight] if (wgrp==1 | wgrp==2), robust
cluster(sch98v1)


** T9 ORIGINAL
 use http://www.columbia.edu/~mh2245/w/MK_T9_original, clear
 mean  pop_3km_original popT_3km_original pop_36k_original popT_36k_original  [aw=obs] if (t1~=.
& elg~=. & sch98v1~=. & mk96_s~=. & p1~=. & Istd2~=. & pop1_3km_original~=.)
 regress prs t1 t2 elg p1 mk96_s Y98sap* sap* Istd* Isem* pop_3km_original popT_3km_original
pop_36k_original popT_36k_original [aw=obs] if (t1~=. & elg~=. & sch98v1~=. & mk96_s~=. & p1~=.
& Istd2~=. & pop1_3km_original~=.), robust cluster(sch98v1)

* Externality effect: the 0.02 discussed in the paper seems to use a t_any regression, not Model
3.

quietly  regress prs t_any elg p1 mk96_s Y98sap* sap* Istd* Isem* pop_3km_original
popT_3km_original pop_36k_original popT_36k_original [aw=obs] if (t1~=. & elg~=. & sch98v1~=. &
mk96_s~=. & p1~=. & Istd2~=. & pop1_3km_original~=.), robust cluster(sch98v1)
```

```
display .000048*608.305  -.0000127*726.893

test pop_3km_original*608.305 +  pop_36k_original*726.893 = 0
* Compare: estimate in MK code: effect  .01996351, tau_se: .01346655

* If Model 3 coefficients were used, this would give
quietly regress prs t1 t2 elg p1 mk96_s Y98sap* sap* Istd* Isem* pop_3km_original
popT_3km_original pop_36k_original popT_36k_original [aw=obs] if (t1~=. & elg~=. & sch98v1~=. &
mk96_s~=. & p1~=. & Istd2~=. & pop1_3km_original~=.), robust cluster(sch98v1)
display .000044*608.305  -.000014*726.893
test pop_3km_original*608.305 +  pop_36k_original*726.893 = 0

 test pop_3km_original pop_36k_original


 * T9 UPDATED
 use http://www.columbia.edu/~mh2245/w/MK_T9_updated.dta, clear
 mean  pop_3km_updated popT_3km_updated pop_36k_updated popT_36k_updated  [aw=obs] if (t1~=. &
elg~=. & sch98v1~=. & mk96_s~=. & p1~=. & Istd2~=. & pop_3km_updated~=.)
 regress prs t1 t2 elg98 p1 mk96_s Y98sap* sap* Istd* Isem* pop_3km_updated popT_3km_updated
pop_36k_updated popT_36k_updated [aw=obs] if (t1~=. & elg98~=. & sch98v1~=. & mk96_s~=. & p1~=.
& Istd2~=. & pop1_3km_updated~=.), robust cluster(sch98v1)
* Externality effect:
 display .0000395*605.65527   -.0000242*1631.4674
 test pop_3km_updated*605.65527 +  pop_36k_updated*1631.4674 = 0
* Joint Test
 test pop_3km_updated pop_36k_updated

* Full externality effect
 display .0000395*1268.957   -.0000242*3404.18
 test pop_3km_updated*1268.957 +  pop_36k_updated*3404.18 = 0

* If t_any coefficients were used (as in MK code), this would give
quietly regress prs t_any elg98 p1 mk96_s Y98sap* sap* Istd* Isem* pop_3km_updated
popT_3km_updated pop_36k_updated popT_36k_updated [aw=obs] if (t1~=. & elg98~=. & sch98v1~=. &
mk96_s~=. & p1~=. & Istd2~=. & pop1_3km_updated~=.), robust cluster(sch98v1)
display .000038*605.65527   -.0000243* 1631.4674
test pop_3km_updated*605.65527 +  pop_36k_updated*1631.4674 = 0

* Combine ranges
g pop06 = pop_3km_updated + pop_36k_u
g pop06T = popT_3km_updated + popT_36k_updated
regress prs t1 t2 elg98 p1 mk96_s Y98sap* sap* Istd* Isem* pop06 pop06T  [aw=obs] if (t1~=. &
elg98~=. & sch98v1~=. & mk96_s~=. & p1~=. & Istd2~=. & pop1_3km_updated~=.), robust
cluster(sch98v1)



* Robustness check:
* Run a model excluding group2 (as proposed by Ozler), dropping weights entirely, and excluding
spillover analysis. Interact treatment with demeaned year variable so that coefficient on t_any
is then the average effect across the two years, allowing for heterogeneity over time
use http://www.columbia.edu/~mh2245/w/MK_T9_updated.dta, clear
egen groups = mean(t1+t2), by(pupid)

g year = (yr - 1.5 )

xi: regress prs i.t_any*year elg98 p1 mk96_s Y98sap* sap* Istd* Isem* if (t1~=. & elg98~=. &
sch98v1~=. & mk96_s~=. & p1~=. & Istd2~=. & pop1_3km_updated~=. & groups!=.5 ), robust
cluster(sch98v1)
```

Annex B: Regression does not handle heterogeneous propensities well

Regression does not handle heterogeneous propensities well. To see why, imagine a case with three possible levels of exposure to spillovers (0, 1, or 2). Table 1 shows possible "potential outcomes" for each unit for each level of exposure – these are the values on the outcome that a unit *would* take it if received a given level of exposure. In this example, if the groups are equally sized, then there would a zero average treatment effect for all comparisons of treatment groups. That is, the average outcome across units at each level of spillover is 0. The (null) treatment effect is linear in the sense that the average effect of moving from 0 to 1 is the same as the average effect of moving from 1 to 2; though there is unit heterogeneity in effects. See Table 1.

		Group 1 Potential Outcomes	Group 2 Potential Outcomes	Group 1 propensity	Group 2 propensity
Indirect treatment = # Treated within 3km	0	0	2	0.5	0.5
	1	2	0	0.5	0.0
	2	0	2	0.0	0.5

So far so good. Imagine now though these different units were assigned to indirect treatments with different propensities, as given in the Table above. In the case in effect the comparison of outcomes between level 0 and level 1 of treatment would use variation from Group 1 units only; which suggests a 2 point effect. The comparison between level 0 and level 2 of treatment would use data from Group 2 units only for a 0 point effect. No comparison can be made between levels 1 and 2. Combining these estimates (of different treatments) would yield an estimated 1 point average effect. If you ran a regression of observed outcomes on treatment, and you were careful enough to control for group and for the interaction between group and treatment, you would estimate this 1 unit effect. With lots of data of this form, this 1 unit estimated effect could be estimated precisely with as low a p value as you like. But it would be wrong. The true average treatment effect is 0.

The problem is a deep one. In this case the problem is that different units are really taking part in different experiments and are being inappropriately aggregated. With a slightly less extreme example (where eg propensities were (.5, .4, .1) instead of (.5, .5, 0), you can think of all units being in the same experiment but with different assignment propensities. In this case regression would continue to get the wrong answer but you could get the right answer using inverse propensity score weighting.