Rashad: Welcome to this month’s QPI Technical Meeting. This month we are joined by several leaders in the field of improvement science to comment on an article in the Annals of Internal Medicine, written by Frank Davidoff, “Improvement interventions are social treatments, not pills.” This was the article distributed before the meeting, which is a commentary on a recently published randomized control study by Goldman et al. Frank is Editor Emeritus of Annals and an Adjunct Professor at The Dartmouth Institute. We have the pleasure of having Frank join us today to provide a background to this piece. In addition, we are joined by our partner Tom Bossert, a professor of Global Health Policy at the Harvard School of Public Health. John Øvretveit, Director of Research and Professor of Health Improvement, Implementation and Evaluation, at the Karolinska Institutet will also be contributing his expertise. Finally, our very own Edward Broughton, Director of Research and Evaluation for the USAID ASSIST Project will join the other panelists in discussing the pros and cons of different techniques for how to study improvement, and applications to the evaluation and improvement work we do. Thank you all for joining us. I would like to ask Frank to start us off by providing a background to his piece in Annals.

Frank: Thank you, Rashad for the opportunity to be a part of this discussion. I wrote this piece after reading a beautifully designed study by Goldman and colleagues, which was a randomized controlled trial that uses p-value driven statistics. The so-called “p-value driven statistics” require stability of the
study design. Since those study designs have been very powerful in proving the efficacy of interventions such as pills, they are not something that people like to tamper with, understandably. But here we are with a dilemma. That dilemma became apparent in this article that Annals published, which was a very carefully designed, protocol-driven study of an intervention to reduce readmissions in a city in which there was a great deal of ethnic diversity and many different languages spoken. To try and reduce readmissions rates they involved nurses to do intensive education in the language of the patient to try and reinforce ways to avoid having to be readmitted. After the usual considerable time that it takes to analyze the data it became apparent that this intervention did not affect the readmission rate.

In thinking about this, it seemed to me that it was a pretty clear example of a very nice, beautifully designed, protocol-driven study that did tell us that the intervention didn’t work, but it didn’t tell us what improvement can possibly do to change things. I actually suggested a few possible modifications that the study group might have done to the intervention to try to make it work better, but that was not really permitted by the fact they were locked into this protocol. So in essence, we are faced with this dilemma of how to study improvement interventions in a way that can convincingly demonstrate to the researchers, and to the people looking over their shoulder, and to the rest of the world that improvement really works. Which, traditionally, has required protocol-driven, traditional research study designs.

I am not sure that there is a clear answer to how to get out of this dilemma. I suggest that there are two possible ways that I know about. One is to use mixed methods, that is to say, to carry out the protocol-driven study and to try to capture evidence that the improvement is in fact making a positive difference, but to supplement that with carefully done qualitative data collection to try to get at the question of why the improvement intervention worked or why it didn’t work, if it does not work, and to try and explore the implications for how it might be improved. All of which takes an entirely different kind of research approach and technical knowledge that most of us, who were not trained in qualitative research, do not have.

The other, complementary approach would be to switch to a different kind of statistical analysis. That is, to recognize that improvement and healthcare delivery are time-dependent processes. In most traditional research, except for prognostic studies, essentially eliminates the dimension of time. They don’t take time into account. But time is an essential element in healthcare delivery and in improvement. Improvements are changes and changes take place over time. So time-sensitive statistical methods, in particular statistical process control, are more suitable for ways to capture the meaningful data on the effect of improvement interventions. The problem is that as much as those techniques are effective in detecting meaningful changes in outcomes as a result of an intervention, they are not particularly suited for detecting confounders or bias. However, of course part of the rationale for protocol drive study designs is to control for bias and confounders.

So we haven’t quite resolved how to escape from the dilemma, however the fact that the issues are now on the table is helpful, because as I’ve just described, there are some ways to deal with these problems, which involve techniques that we in clinical medicine and research, have not been particularly schooled in or knowledgeable about. These approaches are much more akin to social science, and social science is still something that, curiously enough, most people in medicine have relatively little serious experience in. And, social sciences are a different matter altogether, they struggle with the issue of if they are a science to begin with. And with that I will end it and hope to have questions and good discussion.

Rashad: Wonderful, thank you so much Frank. What a great introduction. As you can imagine, I was struck while reading your editorial as this is the issue we face all the time in our work. We have a plethora of time series charts that show significant improvements with multiple, iterative interventions, some of which work and some of which do not, and many variations and adaptations to the
interventions. And then, when we show these, the questions we get asked are, “How do you know if these improvements are due only to what you are doing?” which is really a big question we get asked all the time, and that gets us into the arena of well, we never actually set any controls, and we modify all these interventions as we went along. But we’ve almost come to some level of thinking that we need to have some kind of control site of sorts that allow us to say the bundle of things we worked on over time and changed over time actually led to the improvement, because we could not see them somewhere else that is comparable that didn’t have them. But we are not standing on very solid ground as we say that, and that is one of the issues that comes up all the time. There are other issues that come up, but let me start with this one: that in a controlled trial, like in the paper Frank is commenting on, this intervention with these very particular parameters, implemented with this fidelity, led us to say whether this worked or not. But when we are intervening in a messy situation, without any controls, and we are continually adapting and changing what we are doing to make it work, we get asked the question, “How do you know?” Maybe this is the first question we can take up: how can we deal with this continuous criticism of the improvements that we make, which is the inference to, can we attribute the improvements that we saw at these sites to what we did? I will open it up to the panel. Let me ask Tom to comment on it first.

Tom: Thank you, Frank, for the great introduction. Unfortunately I have not been involved in any clinical trials like this one, and I am a social scientist, although I don’t make much claim for the science part. But, I think we do a lot of work on the two kinds of options that Frank proposes. We do a lot of mixed methods and we are very aware of the limitations that some of the methods that have been used, especially on the qualitative side, have had. So, we do not get around the problem of causality, even in a mixed methods approach. And in a time-series method, we usually do this in terms of time series of an intervention area and controls, and do something we call difference in differences analysis, to show that the changes that occurred in the intervention would be different from what changes occurred in the area that did not have the intervention. This allows us to get around, a little bit, of the problems that Rashad was raising, but if you don’t have a good set of controls then you have a serious problem. You also will have the general problem of not controlling for some things that might have had an effect or an observed thing. Therefore, the emphasis that we work toward is having some kind of randomization of both the controls and the interventions to try and get around some of those things.

One of my colleagues in the undergraduate school, Gary King, has developed an innovative way to approach the randomization problem, and one that might lead to, at least, shorter periods and possible changes in the protocol over time. And that is his evaluation of a major health reform in Mexico. He selected matched pairs of districts, that along all the dimensions that he could use, or could have for those, were matched as closely as they could be. Then he randomized which of the matched districts would get the intervention, the reform. He was able to do this study, which given the kind of things we normally do, in about 8 months, he was able to do an analysis of this. Then the pair that did not get the intervention was then given intervention after this. So there are other methods of randomizing, and he [Gary King] challenged doing this. He challenged the prevailing wisdom of the statisticians and showed that it was in fact better to do that in the matched pair than in the randomized control/ randomized intervention, because if you had—and we often do, in social sciences especially—one of the districts might drop out. When it drops out, it undermines the power of the randomization.

In preparing for this panel and thinking about this issue, I communicated with Gary King, and asked him, because he had mentioned in a seminar that we had for the TRAction Project, that he thought it would be good to develop a method of evaluation that would give real-time response. This would fit into what both Frank and Rashad are looking for. I asked Gary if he had published anything on this, he said unfortunately he hasn’t, but he referred to some of the methods that Amazon uses to pull together real-
time information for making recommendations for what you should buy. Although that probably
depends on big data issues that in most of the countries that we are working with, is not feasible, it
seems to me, that if we move in that kind of direction of thinking of how you can utilize larger datasets
or some kind of modified randomization, then we get around some of the problems and criticisms that
Rashad raised.

One more comment—I think part of the problem Rashad was talking about, about these time-series run
charts, is that you don’t do enough baseline data first. So you have zero at the beginning, because you
just haven’t collected data. Edward brought this up in one of the previous discussions, and that really
undermines the credibility of a lot of that kind of work. My underlying argument is yes, we shouldn’t be
working towards a clinical, protocol model, but we shouldn’t give up the attempt to try to get a whole
lot better data and try to find some ways of at least having a control, and if not, having some kind of
randomization in order to try to answer these real, fundamental questions about causality. Are the
things we are doing actually causing the improvements or not? We have to try to think of ways within
that set of standards to get the answers closer to real time.

Rashad: Thank you very much, Tom. Let me go to John. John, you have done so many evaluations of
improvement, and you usually come in not at the beginning when they are designing the improvement,
but more in the evaluation phase. We would like to hear about that. How do you go about evaluating
whether interventions worked, whether they had an impact, and what factors were considered? How do
you do these evaluations in real life, given the numerous experiences you have had doing this.

John: Thank you very much, Rashad. I think it is really interesting and I find it fascinating because what
we have here, I am the third person speaking, representing a different paradigm. I work in an institution
where both the medical controlled trial approach is thought of as, “What else would you do?” There is
no question. And there are few of us using traditional social science methods, like the one just
described. The approach that I take is, or the phrase I use is: the controlled trial is the best of designs
and the worst of designs, depending on who it is for, and which decisions it is meant to inform.

The first of three points I want to make before I get more boring, is that it is all about matching the
method to the decisions it is meant to inform and the time and money available. But absolutely, if we
can do an RCT [randomized controlled trial], then let’s do it and it will give you one type of information,
which is about effectiveness under certain conditions. Where I am coming from, is working with a lot of
policy makers, managers and decision makers that say “well actually in terms of our decisions, to go with
an intervention and then our design and planning of the program in some ways, whether or not its
effective, is the least of our worries.” We know the more we move towards complex, social
interventions the more we have got to reinvent it. So the issue for us is how can we appropriately adapt
it and how can we monitor that we are doing that well, and that it is having as good results, if not better
results, than those in the original test site. So my first point is to match the method to the needs of the
decision makers, what I call the evaluation users, and the time and money you have available. It’s a very
pragmatic approach.

The second point is that some interventions are more context-sensitive than others. Indeed, some
interventions depend so much on context that the idea of standardizing and controlling and obsessing
for fidelities is just a joke. I say that, doing more and more evaluations in digital health technologies
integrated into social care systems. Not only in three years and you have an entirely new technology and
it integrates into the social delivery system in a very different way, so we have to use a different... the
whole field of health informatics—look at the debate in the field of clinical informatics about the
appropriate methods for doing an evaluation. It is much more sophisticated than anything I have seen in
any of the quality improvement... there are a whole set of methods and people are much more flexible
and accepting of other approaches.
The third point I want to make is that I find it useful to distinguish between three types of— and I hate this word— interventions, for what we are doing. Three types of interventions, and the evaluation approach that you need to take is different for each. Remember we are mostly talking about interventions to patients either a medication where you have a natural process and a natural mechanism where what the patient thinks of it has some influence, but really has little choice. For example, however much you struggle against the anesthesia, you will be asleep by the count of ten. We know the placebo [effect], but essentially this is an intervention to the patient and there are actual physiological processes that we are assuming would just kick in and we do not have to worry too much about what the patient thinks. These are standardizable. So long as the patient keeps taking the medicine—and that is a whole other set of issues— then we have got interventions to individual practitioners to change their prescribing habits, or reminders, or things like that. Then we also have interventions to organizations where groups of people change the way they organize their service delivery model, e.g. team-based, those sorts of things. The more you get to service delivery interventions and larger policy interventions, the less easy it is to prescribe and standardize it. However we are interested in some things which can be prescribed and are standardizable. In the field of international health, there has been one approach that has simply said you have to have the equivalent of a trial protocol to implement this and you have to copy it exactly. The point there is that we often do not have the resources that they have available to copy it exactly, or to run a randomized controlled trial where researchers are able to get extra resources both for themselves and for the services, to move heaven and earth to make sure that the protocols are implemented and kept to so everything can be evaluated. So first of all, we have got that standardized and prescribed approach.

A lot of what we do is a second type, where we actually are implementing principles or concepts or ideas. The example here would be the chronic care management model, or the Goldman readmissions program that Frank described so well. These are not prescribed and standardized, but even though Goldman and others have done really interesting work to parse out and specify quite well what is core and what is adaptable— now that’s another story. But essentially, the people on the ground have to figure out how to reproduce the spirit of the idea in their situation. The worst thing they can do is copy exactly, with fidelity, to the trial site. This just won’t work.

We have discussed how to evaluate prescribed and standardizable interventions and principle type interventions. I have found it useful to parse out a third category of interventions, where we are evaluating a process and methods that people use to do improvement. An example would be doing a process or pathway analysis or using PDCA. The question there is — is what we tried to implement, is that people uses a standard method for adapting and testing an intervention, and did they copy this method and apply it appropriately, and you would evaluate that. You could say that is an evaluation of an implementation.

I think those three types of interventions need different approaches to evaluate, and I also think that one needs to think about how context-dependent it is. One approach to evaluation usefully excludes and controls the effect of context. I think this then provides information that is of limited use to people on the ground who do not have the resources, and are in a very different situation and need to know if the evaluation will work in their setting. There is a fantastic paper written, called “Will it Work Here?” by Brach, which I can also send you.

Rashad: John, thank you very much. You really hit very well on the issue of whether we can test something in another setting with the same fidelity or not. You have eloquently spoken to that and it is so consistent with the issues.

John: One last thing, the real challenge for those of us working in implementation science and doing hybrid studies—I am currently working with VA information center and we use a lot of hybrid designs. A
quick way to summarize would be in addition to the two approaches we have heard, although Frank isn’t advertising controlled trials but he has well described and summarized the approach and assumptions, what I am pointing to is really option 3, 4, 5, 6, 7, 8, 9, 10, which comes from the field of program evaluation and implementation science, and a history of 50 years in welfare service evaluations in public health and approaches used enormously in public health that I don’t understand why improvement science folks and medical researchers don’t know about or recognize all of that literature, where people have faced and dealt with all of the issues that have been raised for 50 years. There is a whole literature and background on this. I think some of the forefront work is, if we accept that there is going to be local adaptation, what we need to understand is what helps and hinders people to implement what is the change that looks promising and how they can collect data to assess whether they have implemented it well or whether they have adapted it appropriately. That type of action evaluation is very interesting, but it has its own challenges and approaches. There are no ready packaged, sure-fire ways to answer the problems and dilemmas but there are a lot of people working from a different perspective and I believe they have a lot to contribute to this debate. I think those of you working in international health could be at the forefront of these types of developments.

Rashad: Wonderful, John, this is terrific. As there are no hands raised, I would like to move to Edward. We have had multiple conversations about how can we improve our learning about how we improve better and you have gone through a number of changes to our programming and our research that allow us to study it better and maybe reflecting on some of what John, Tom and Frank have said, you can tell us where we are now and where you think we could be heading based on that.

Edward: Thank you very much Rashad and thanks to Frank, Tom and John for their previous comments. Going through what Frank was saying in his article, I agree with the idea that the article with the randomized trial was flawed in that the definition of the intervention itself was too narrow. Going off what John said just a moment ago, I think in terms of fidelity to a treatment or an intervention, I think if you very narrowly define what fidelity to that intervention is, then you are going to end up in trouble. The trouble being that these are social interventions, they are very complex, highly variable and context specific. Again, going off what John said a moment ago, if we are thinking about improvement as a set of principles or a philosophy and a certain methodology used and adapted quite significantly to the context, then it should be fairly easy to get those things through IRBs [institutional review boards] and also get the scientific world to pay attention to the results that are achieved. It is no different than drug trials, because if you consider doing a randomized controlled trial for a seizure medication, a person who takes an oral seizure medication in general, the protocol for those kinds of studies doesn’t control for the rest of their life. So that person, in that trial, is free to eat whatever they want, exercise as little or as much as they would like, have the emotional life that they want, but they have to take the medication. That is the only part of their life being controlled. This basic philosophy and these basic principles in research are already adhered to in randomized trials - they do not control people’s lives, only one aspect of it. With these interventions for improvement, we are not controlling every part of the intervention, we are just saying the principles need to be applied. The principles of identifying problems that are occurring in facilities and addressing those and using PDSA cycles, or whatever, to determine whether they are working or not. This is the way that it is done.

Going back to this seething dichotomy between p-value statistics and run chart analysis, I think if we collect time-series data in a trial setting we can still use standard statistics for analysis. We can still use p-value statistics and I think it is important to do that because there is great value in knowing the level of certainty we have in the results and I think it’s important not to throw out standard statistics just because we are using run charts. I think there is an area where we can use both of them. Using standard statistics but having time as a variable in a regression equation would allow us in improvement work to
control for confounders that are present in the facility, if that is where we are conducting the improvement work. I have a lot more to say on the topic, but I will leave it at that for now. Thank you.

Rashad: Thank you. I believe we have a question from Leighann.

Leighann: I am coming from a social sciences background and I think what the common theme is separating controls and context, so my question is what would be the harm in integrating the social sciences and collecting data that is more contextual in order to better understand and supplement the data that is being collected on these trial bases?

John: I think you can do this two ways. First, I think you can try and control for context, as Tom said with matching, that is the practical way. Or, you can actually dive into and embrace context, and collect data that is about context and try and understand which aspects of context help and hinder the implementation, and the proximal, distal and other results down the causal train. You can do a single-site study using that. And so that is two approaches, one of which controls for context using traditional methods, randomization or matching. The second way is to say that we cannot control for context, what we need is to understand how context affects, and in terms of actually implementing, we need to enable people to adapt to context appropriately or even judge when—don’t even think about starting because the context is so hostile—you don’t stand a chance in really getting terribly far. This second approach would build a logframe model that has context as part of the logframe program theory and you would build your theory about which aspects of context may be important to implementing this program, either a promising one or proven elsewhere. You would build that on the basis of previous research on similar things, or you would use a generic implementation science model that theorizes about which contexts are important for these sorts of interventions that you are looking at. It is possible to do it either way. Our units and I have tended to specialize in that second approach simply because of time and money available, and that the information we get about outcomes is much, much less certain, but what we have is a causal change showing that we have an effect and if this effect led to that effect. Those of us who have used logframes for program planning are familiar with that way of thinking and which data to gather. That is an alternative case study type of way. Then there is a whole set of other issues about making it an action evaluation by feeding back every 3 or 6 months to the implementers the results of what you are finding from an independent perspective, and they then adapt the implementation, and then we are into a much more dynamic thing. Lastly, that approach is simply doing PDSA, but the researchers are doing the study bit and feeding back the data. It is an iterative PDSA with the researchers being the data gatherers and feeding the data to help the study bit.

Rashad: John that is wonderful, thank you very much. We have a question from Esther from our Uganda program. Esther, please go ahead.

Esther: I wanted to ask about the role of the controls for use of day-to-day decisions for how one change leads to improvement as we have to decide when we are working on an area whether one change at a time leads to improvement. An example would be if we wanted to decide whether patients are coming back to keep their appointments, we would test separate changes at several sites. So if one site is making phone calls, we would ask them to make calls to 10 patients and find out if those patients came back and do that for several clinics until we have enough information to make a decision using run chart rules. How do controls apply at this level? Do we have to get another 10 patients in the same facility and compare their appointment-keeping behavior compared to those we made phone calls to then to increase the confidence in the decision towards improvement?

Edward: Thanks very much for the question, Esther. That is a great one. I think that in this kind of case—I don’t usually like to make a too much of a difference between research and evaluation as you are going along in an improvement intervention—but in this case I think it is ideal to have a control group and it’s
actually necessary in many ways in research, but in the context of you guys doing the improvement in a certain number of facilities in Uganda or elsewhere, you don’t have the luxury of having control groups because you can’t go to facilities you aren’t working in and collect the same data—it’s just not practicable. Then I think you just have to use the run charts and understand that maybe there are other things going on in the country overall that are changing the outcomes. But, just according to the data that you have and according to what is feasible to collect in these facilities in which you are working, you can see this improvement and then move ahead with a degree of certainty that is consistent with the data you have collected. When you are trying to make generalizations about methodologies, then in particular the situations in which we are working where there is a lot of instability in countries—we have a program in Ukraine, for example, there is a lot going on there, and we do not have a control group in that situation, so we will never really know with confidence if our intervention is causing the changes that we are seeing because there is so much else going on there. So we cannot report to the world if this particular intervention is working because we just do not have that certainty. But in terms of implementing the program on the ground, in the facilities, in Ukraine or in Uganda, we can go by a somewhat lower or more practical level of evidence to go forward with the intervention. Otherwise we would be paralyzed by this need for a high degree of certainty that we would never be able to attain.

**Tom:** It seems to me that the proposal that she made would be stronger if the intervention was not universal, but that you can show that having called the patients and having relatively similar patients that are not called, you would see you most likely have an effect. That is pretty powerful. It does not take a lot of additional research or resources to do that. That is the kind of thing that I think would be quite useful to try to push further toward, having everyone be aware of the need to have some kind of controls for the things you are talking about.

I would also like to comment on the case study approach. I found John’s description of the different models he’s been able to use very impressive. I would add just a small comment that a part of the job of these kinds of case studies is to make an argument that alternative explanations are not as good as the kind of intervention that you have done. So you would try to get as much information as possible in that case study or in that type of gestalt analysis to say that another possibility that is causing the outcome could be this but I have investigated that and have found that that is not a very powerful explanation and therefore it is more likely that what we have been doing has had the effect.

**Rashad:** Thank you very much, Tom. We are coming to the end of the hour. We have several questions that have not been asked. Can I ask those who have questions to be so kind to write them down in an email and send them to Leighann please. I would like to ask the panelists to respond to any or all of the questions and we will share the responses with everybody. All in agreement? Yes. Can I ask all panelists to take about 30 seconds each for concluding remarks. Frank, would you like to begin?

**Frank:** It has been a fascinating discussion. A couple of things struck me as important. One is that in traditional clinical research, understandably the logic suggests that unless you can demonstrate internal validity, that is that the things works within the experimental situation, then there is not much point in looking at generalizability. On the other hand, a number of researchers have pointed out that if you have a reasonable measure of outcome, and it seems to be working in your original setting and then you carry it out in a number of other settings, and it seems to be working there, then replication in itself can be used as a form of evidence for efficacy. That is worth pondering as sometimes you don’t have much choice as there is not easy access to an appropriate control group.

The second point is the absolutely crucial importance of baseline data. There is always a temptation to rush into an intervention without the baseline data but that is a big mistake.
Finally, the mention was of Plan-Do-Study-Act, PDSA cycles, a favorite tool in improvement. Some recent work has really shown that PDSA cycles are essentially a social tool. They are used in many different ways, you can’t just point to one and say that there is a such thing as a standard PDSA cycle.

**Tom:** This has been very interesting and I hope we can continue this kind of discussion as I am up in Vermont a lot of the time so perhaps Frank and I can get together at Dartmouth. One comment that keeps coming up is that we do not have enough resources to conduct a good study. It seems to me that we need to advocate for more resources for doing this kind of work. I know there is a lot of emphasis at USAID and others at Washington to change the orientation for better evaluation but it does not seem to have trickled down to the mission levels and to levels of people designing interventions that can improve quality. At the same time, they are starving us from the kinds of things that we could do to show that there is something that will work. I think one of the things we should do is advocate for more resources for doing this sort of thing and not just see it as a limitation on what kind of methods we can use. We should try to use these newer methods, and as John pointed out, many of these older methods that we have been using for many years. We should try to show that you do not need a very sophisticated long term randomized controlled trial to show good results if you have a good baseline, if you have decent controls, and if you have carefully designed. One of the things that struck me about Frank’s article was that he made some suggestions that I would have thought the initial designers of the intervention would have paid attention to, just for example, including the family issue. It seems to me that part of the job of intervention people is to try to anticipate the kind of changes that might improve quality. Those things should be included more in the earlier stages in order to make sure you are not missing something you could have done and would have improved. This is part of that process of improvement, but still also something that why don’t you think about it more carefully at the beginning and do some more pilots, to show you where you could be doing more work.

**John:** Amen. Great comments, absolutely agree. You are more likely to get funded if you do action evaluation that feeds back the data while doing it, but that opens up a whole other set of issues. Thanks.

**Edward:** Thank you. I definitely agree with all previous speakers as well. I think given the way USAID funding is these days, and the way things are in the field I think that we still need to do the best possible designs we can, given the circumstances and we should really hold ourselves to account for doing that. Trying to get the best science out there possible to prove that this works or doesn’t work and why. It is incumbent on us to do that.

**Rashad:** Fascinating. Wish there was more time. Please email your questions to Leighann and we will distribute them to everybody. Once we aggregate responses, I can see us bringing them back and distributing them to not just those that asked the questions but others. I hope this is the basis for much more conversation for how we learn about improvement, how we design our improvement in such a way that we achieve the ability to confidently say that the changes we made led to the improvements that we see and the rigor of the data and the results is maintained through the good work all of you are doing. Big thanks to everyone and meeting adjourned.