Double-Blinding: The Benefits and Risks of Being in the Dark
Jeremy Howick

Series Editor: Damien Fennell
Double-Blinding: The Benefits and Risks of Being in the Dark
Jeremy Howick∗

Editor’s Note
In evidence based medicine, a great deal of work has been devoted to developing methods that produce reliable, unbiased claims that are well-supported. One highly desirable feature of medical trials is that of being ‘double-blinded’ where patients and those administering treatment in the trial are unaware of who is in the treatment and control groups. But just what are the benefits of double blinding? Does double blinding have any negative consequences? In his paper, Howick carefully scrutinizes the methodological role of double-blinding in clinical trials. As a result, Howick draws out many important (contingent) conditions that need to be met for double-blinding to provide its benefits.

Abstract
The feature of being “double blind”, where neither patients nor physicians are aware of who receives the experimental treatment, is almost universally trumpeted as being a virtue of clinical trials. Hence, trials that fail to remain successfully double blind are regarded as providing inferior evidential support. The rationale for this view is unobjectionable: double blinding rules out the potential confounding influences of patient and physician beliefs. Nonetheless, viewing double blind trial as necessarily superior is problematic. For one, it leads to the paradox that very effective experimental treatments will not be supportable by best evidence. If a new drug were to make even the most severe symptoms of the common cold disappear within seconds, most participants and investigators would correctly identify it as the latest wonder drug and not the control (i.e. placebo) treatment. Any trial testing the effectiveness of this wonder drug will therefore fail to remain double blind. Similar problems arise for treatments, such as exercise and most surgical techniques, whose nature makes them resistant to being tested in double blind conditions. It seems strange that an account of evidence should make a priori judgments that certain claims can never be supported by ‘best evidence’. It would be different if the claims at issue were pseudoscientific – untestable. But so far as treatments with large effects go, the claim that they are effective is highly testable and intuitively they should receive greater support from the evidence than do claims about treatments with moderate effects. In this paper I argue that the two potential confounders ruled out by double blinding are not actual confounders outside placebo controlled trials of treatments with mild effects and that have subjective outcome measures.

The patient, treated on the fashionable theory, sometimes gets well in spite of the medicine. The medicine therefore restored him, and the young doctor received new courage to proceed in his bold experiments on the lives of his fellow creatures”
Thomas Jefferson, letter to Dr. Caspar Wistar

many investigators and readers delineate a randomized trial as high quality if it is “double-blind,” as if double-blinding is the sine qua non of a randomized controlled trial. ... A randomized trial, however, can be methodologically sound ... and not be double-blind or, conversely, double-blind and not methodologically sound.
Schulz, Chalmers, and Altman (2002)

* Department of Public Health and Primary Care, University of Oxford
1. **The Problems with Double Masking as a Requirement for Clinical Trial Validity**

Being ‘double blind’ or ‘double masked’, where neither the participants nor the investigators are aware of who gets the experimental treatment, is almost universally trumpeted as being a virtue of medical experiments. Sir Austen Bradford Hill, for example, states:

By the so-called double-blind procedure, when neither patient nor doctor knows the nature of the treatment given to an individual case, it is hoped that unbiased subjective judgements of the course of the illness can be obtained (Hill and Hill 1991, p. 214)

The official evidence based medicine (EBM) text states:

Blinding is necessary to avoid patients’ reporting of symptoms or their adherence to treatment being affected by hunches about whether the treatment is effective. Similarly, blinding prevents the report or interpretation of symptoms from being affected by the clinician’s or outcomes assessor’s suspicions about the effectiveness of the study intervention (Straus, Richardson, and Haynes 2005, p.122)

Armitage, Berry, and Lindgren state:

In such a trial [i.e. a placebo controlled trial] the mere knowledge that an additional intervention is made for one group only may produce an apparent benefit, irrespective of the intrinsic merits of that intervention. The principle of masking the identity of a treatment may be extended to trials in which two or more potentially active treatments are compared. The main purpose here is to ensure that the measurement of the response variable is not affected by knowledge of the specific treatment administered (Armitage, Berry, and Matthews 2002)

Although as we shall soon see the term ‘double masked’ is often used ambiguously, support for double masking can be found in virtually any account of clinical trials textbook (Greenhalgh 2006) (FDA 2005) (Bland 2000) (Jadad 1998). This is because failure to double mask the study successfully introduces the possibility that participant or investigator beliefs or expectations confound the study. As a consequence, failure to keep trials double masked is regarded as a relative vice – a successfully double masked study is regarded to be of higher quality than one that is not successfully double masked.

Although double masking certainly adds value in many cases, there are problems with viewing double masking as a universal virtue. For one, it leads to the “Phillip’s paradox” (Ney, Collins, and Spensor 1986), that very effective experimental treatments will not be testable in double masked trials – the dramatic effectiveness of
the treatment will make it identifiable by both trial participants as well as dispensing physicians. For instance, imagine a new drug for the common cold was invented that the investigators were intending to test against an (initially) observationally-indistinguishable placebo. If the new pill made even the most severe symptoms of the cold disappear within seconds of swallowing it, then most participants and investigators would correctly guess which treatment was the wonder-drug and which was the placebo. A further problem is that many treatments, ranging from most surgical techniques to exercise, cannot be tested in double masked conditions. For these treatments double masking is an impossible standard. Although we may in the end have to bite the bullet and admit that these treatments are simply unlucky and cannot be supported by ‘best evidence’, it is surely worthwhile first investigating the alleged virtue of double masking very carefully. Perhaps because the terms ‘double masking’, ‘placebo controls’ and even ‘randomized trial’ are often spoken of as if they were necessarily connected, the methodological virtues of double masking have seldom, if ever, been systematically examined in isolation.

In this paper I will evaluate the role of double masking from the fundamental view that good evidence rules out plausible rival hypotheses. To anticipate, I will argue that when investigated this way, it is clear that the methodological value of double masking is far more limited than is usually admitted. After a few clarificatory remarks about the meaning of double masking, I outline the rationale for the view that double masking increases the internal validity of a study. In short, it is thought that two potential confounders, participant and investigator expectations, can be eliminated by successful double masking. If the investigator is aware that a particular participant is in the experimental arm of the trial they may lavish more attention on them. This increased attention could have therapeutic benefits for certain ailments. Similarly, if the participant believes she is receiving the best treatment (as opposed to the placebo), then her knowledge that she is in the experimental arm could lead her not only to report better outcomes, but to experience greater beneficial effects. I then point out that these two potential confounders are sometimes not actual confounders. Then, I contend that there are severe practical limits to the potential success of attempts to keep trials double masked. If so, then there is little value in being described as double masked. Finally, double-masking could impair external validity.
since it contributes to making the trial importantly different from routine clinical practice. In conclusion, double masking, although it potentially increases the internal validity of a study, does not always do so; further, since double masking may not be possible, we may be better off seeking other ways to control for the potentially confounding effects of expectations.

2. The Many Faces of Double Masking: Clarifying the Terminology

The term ‘double masked’ is used in several different ways to describe the masking of various groups involved in a clinical trial. It is therefore necessary to make some clarifying remarks about how I will use the term.

First, however, I will defend my use of the term ‘masked’ instead of the more common ‘blind’. The term ‘blind is ambiguous in trials of blind people, and it is especially abhorred by researchers of eye disease (Bland 2000, p. 19). Second, ‘masking’ someone implies that the concealment procedure could be imperfect. As I will argue later, the process of concealing knowledge to study groups is less successful than most of us believe, and indeed may be inherently difficult to achieve. Third, the term ‘masking’ is more in line with the historical meaning. Early trials that concealed the nature of the treatments from participants literally used masks (Kaptchuk 1998).

Masking is the act of concealing the nature of the intervention from one of the groups involved in the study. For example, in a single masked randomized trial of vitamin C versus placebo as a cure for the common cold, the participants in the trial could be prevented from knowing whether they were taking the placebo or real vitamin C.

Six groups involved in a trial that are sometimes masked, namely:

1. **Participants**

2. **Intervention dispensers (henceforth “dispensers”):** This group administers the intervention, whether it is medical or not. Doctors and nurses performing a surgical intervention are dispensers. Psychiatrists and exercise trainers might also be dispensers.

3. **Data collectors:** The individuals responsible for collecting the data for the study outcomes. This could include taking a blood pressure measurement, reading an X-ray, administering a questionnaire, or “recording symptoms potentially compatible with a transient ischemic attack” (Montori et al. 2002, appendix).
4. **Outcome Evaluators:** This group decides if the participant has suffered (or enjoyed) the outcome of interest. For example, they decide whether a participant has died, or whether a patient has high blood pressure. The step of evaluating the outcomes often goes together with collecting data. For example, in a trial of an antidepressant drug, the data collector might administer the Hamilton Rating Score for Depression (HRSD) and then analyze the data. However, they are separate in many cases, and could be separated in many others, so for purposes of analysis I will follow Devereaux (2001) and use a separate category for outcome evaluators.

5. **Data Analysts:** These are the statistical analysts who make decisions about the type of statistical tests to perform and then perform the tests.

6. **Personnel writing the paper:** This group is very rarely masked. The personnel writing the paper are those who might write alternative versions of the manuscript before the trial is unmasked. In the simple case where the trial has one experimental and one control group – call them A and B respectively, one paper is written as if A is the experimental and another is written as if B is the experimental intervention. There is some evidence that whether the experimental intervention was found to be effective or not can influence how the manuscript is written. This is usually but not necessarily done with the collaboration of the data analysts.

In a study of physicians, 5 major journals and several textbooks, Devereaux found that the term double masked is used in over a dozen ways (Devereaux et al. 2001). About a third defined “double masking” as masking of the participants and dispensers. The remaining definitions included various combinations of 2, 3, and 4 masked groups. Because of this ambiguity, both the CONSORT Statement (Moher, Schulz, and Altman 2001) and Jadad (1998) recommend identifying the particular groups that have been masked rather than using the terms “single masked”, double masked, or “triple masked”, etc.

Although specifying exactly which groups have been masked is always useful, there are good reasons to reserve the term double masked for trials that mask the participants and the dispensers. For one, only the knowledge or beliefs of these two groups can be considered as features of the treatment process *per se*: the participant’s or dispenser’s belief that a particular patient is being treated with the experimental treatment may be a feature of the treatment process that could have direct effects on
the target disorder, while knowledge on the part of the other groups are not. My belief that I am getting the ‘real’ as opposed to placebo, or older treatment, can directly affect the target disorder – or at any rate how I feel about my recovery. Similarly, the belief of the dispenser that a particular participant is getting the ‘real’ or best treatment can translate into both stronger participant belief as well as different treatment. Precisely how the participants and dispensers can be part of the treatment will become clearer in the next section. Knowledge of the data collectors, and certainly the knowledge of the other groups, does not affect the treatment process in any straightforward manner. Moreover, the only two groups that cannot be masked in an observational study are the participants and dispensers. Doctors and patients in routine clinical practice are supposed to know what treatment is being administered. It is, on the contrary, unproblematic, even methodologically desirable, to mask the data collectors, outcome evaluators, data analysts, and even manuscript writers in observational studies. Relatedly, the participants and dispensers are most difficult to mask in trials of treatments that are not easily imitated by legitimate placebo controls. For example, in a trial of exercise versus “placebo” (say supervised flexibility) for depression, although it is problematic to mask participants and dispensers, there is no reason why the other groups cannot be masked. Hence I will use the term double masked to refer to trials where the participants and dispensers are masked.

Reserving the term double masked for trials where the participants and dispensers are masked emphatically does not mean that masking the other groups is unimportant. I will not get into a debate about the merits of masking these other groups. Suffice it to note that masking the other groups may well rule out confounders and that it is therefore important to attempt to achieve masking of these groups. More relevantly, any arguments I present about the limited value of double masking do not bear on the importance of masking the other groups.

2 For example, assessing outcomes on X-Rays or taking blood pressure can be influenced by beliefs of the outcome assessor (Sackett 1991). In conversation, many researchers admit that they hire several data analysts and choose the results of the one they like best. Masking the statisticians and manuscript writers would prevent this. Then, if the results were not what the authors expect, the way they write the article could colour the data. In an example I will discuss in detail later, it could be argued that Hróbjartsson and Peter Gotzsche colour their conclusions of the magnitude of the placebo effect (see below). The potential confounding of these unmasked groups remains to be studied in any detail. “The frequency and magnitude of ascertainment bias introduced after data collection have not been studied at all” (Jadad 1998, p.55).
Concealed allocation is sometimes confused with masking. In fact, concealed allocation “occurs when the person who is enrolling a participant into a clinical trial is unaware whether the next participant to be enrolled will be allocated to the intervention or control group” (Straus et al. 2005, p. 279). For example, in a RCT with 10 participants, some process could allocate the 2nd, 3rd, 6th, 8th, and 9th participant to group A and the 1st, 4th, 5th, 7th, and 10th participant to group B, but not indicate whether A was the experimental or control intervention. Subsequently (post-allocation), the same or different investigators as well as the participants could be made aware of which was the experimental intervention and which was the control. Note further that, as in the above case, the allocation need not be random in order for it to be concealed. Although random allocation might make concealment easier, and concealed allocation might make masking easier, the concepts are distinct.

To sum up, although the term double masked is currently used in conflicting ways, there are good reasons to reserve the term double masked for trials whose participants and dispensers are masked. Why should double masking as thus characterised be considered important?

3. Participant Expectation and Pygmalion Effects as Confounders

In this section, I will explain why it is commonly held that beliefs of participants and dispensers that they are receiving/dispensing the experimental intervention can confound a study. First, however, I will clear up a confusion between what are commonly called ‘placebo effects’ and participant expectations.

---

3 Concealed allocation plays a role in preventing selection bias, while masking the participants and dispensers helps prevent performance bias; masking the other groups attempts to rule out assessment bias. Although randomization, concealed allocation, and masking are conceptually distinct, they often confused in the medical community. The prominent medical statistician Stephen Senn, for example, states: “I am interested in showing how randomization is required to support blinding, but it is worth noting that the converse may also be the case” (Senn 1994, p. 221). Senn goes on to cite a study that suggested that “treatment allocation was biased unless the trials were randomized and blind” (ibid, p. 221). As is quite clear, however, there is no sense in which concealed allocation is required by subsequent blinding. It may be the case, of course, that trials which tend to be unblind also tend not to used concealed or random allocation and generally to be of poorer quality. But this, if true, is a merely contingent matter and not necessitated by logical connections between these notions.
3.1 Mistaken Estimates of the Placebo Effect

If participants and dispensers have beliefs about the effectiveness of the experimental treatment then since these are part of the placebo effect, it is likely that there will be placebo effects. The placebo effect is typically estimated by measuring the effects of placebo treatments. Although placebos have been used knowingly and unknowingly by physicians for at least several centuries, Henry K. Beecher was the first person to attempt to quantify placebo effects in his work which was immortalized in his famous article “The Powerful Placebo” (1955). Beecher’s article is still the most cited work about placebos. His method was to measure the effectiveness of the placebo control in the control group, and take that to be the effects of the placebo control treatment. In his study he reviewed 15 studies of treatments for post-operative pain, cough, pain from angina pectoris, headache, seasickness, and anxiety. The studies had a total of 1082 participants, and found that overall, 35% (± 2.2%) of the patients’ symptoms were relieved by placebo alone (Beecher 1955, p. 1604).

However, it is a mistake to attribute any change in the placebo control group to the placebo control treatment. Beecher has been criticized for failing to consider and rule out explanations for the improvement in the placebo control group other than expectation or placebo effects (Kienle and Kiene 1997). In particular Beecher failed to consider the natural history of the disease, which may of course include spontaneous improvement⁴.

Many ailments, including the common cold and post-operative pain usually go away quite quickly without any treatment at all. More generally, the natural history of the disease, and spontaneous remission, are all potential causes of apparent recovery that have nothing whatsoever to do with placebo effects. Many diseases vary spontaneously with time – the “natural history of the disease”. Spontaneous remission is a common reason for symptoms of an ailment to improve that has nothing to do with the placebo effect.

In order to estimate the effect of the placebo treatment, the effects of these other potential causes of recovery in the placebo control group must be taken into account, something Beecher failed to do. When, for instance, 35% of patients with mild common colds felt better within 6 days (2 days after the onset of placebo administration), Beecher concluded that the effects were due to the placebo

⁴ ‘Observer bias’ which is bias of investigators observing the outcomes, is often included as a potential explanation for placebo effects. Because the genealogy of observer bias is distinct, I discuss it separately below.
administration. Beecher “did not consider that many patients with a mild common cold improve spontaneously within 6 days” (Kienle and Kienle 1997, p.1312). In another similar case,

Beecher referred to patients with diseases such as ulcer, migraine, muscle tension, or headache who suffered from anxiety and tension and were treated for eight 2-week periods alternatively with mephenesin and placebo. Beecher claimed a placebo [expectancy] effect of 30% since “roughly” 20-30% of the patients improved” (Kienle and Kienle 1997, p.1313).

In these examples, Beecher makes the error of failing to take into account spontaneous remission – it is wrong to conclude from the fact that 20-30% of patients improved, that the improvement was due to the effect of expectation.

Kienle and Kienle examined all the studies upon which Beecher based his estimate of the placebo effect and concluded that “none of the original trials cited by Beecher gave grounds to assume the existence of placebo effects” (Kienle and Kienle 1997, p.1316). Kienle and Kienle conclude that “the extent and frequency of placebo effects as published in most of the literature are gross exaggerations” (Kienle and Kienle 1997, p.1316). I will not repeat Kienle and Kienle’s detailed study of each of the trials considered by Beecher. Suffice it to say that Beecher failed to rule out natural history of the disease as a plausible hypothesis for effects measured in the placebo control group and that his commonly cited estimate of the magnitude of the placebo effect may have been exaggerated.

In defense of Beecher, it might be argued that he had more evidence for placebo effects than Kienle and Kienle allow, and the scope of his conclusion was not as far reaching as is sometimes assumed. Supporting the hypothesis that placebos have effects, he cites (qualitative) evidence of a dose-response effect in different placebos –this evidence is not called into question by any of the points that Kienle and Kienle make. Also, he limited the scope of his argument to treatments with subjective responses. “It is evident that placebos have a high degree of therapeutic effectiveness in treating subjective responses” (Beecher 1955, p. 1604).

3.2. Participant belief

A participant’s belief that she is being treated with an effective drug could, at least in theory, translate into effects for the outcome of interest. For example, if I believe I am being given the latest and best treatment for the common cold, I may well recover
more quickly than had I not taken the latest treatment, or I may well report that I have recovered more quickly and this is all that is at issue when the outcome is subjective.

I will call the effects of knowledge that one is being treated with something one believes at least may be effective “belief effects”\(^5\). To measure the effect of participant belief, we need a trial where one group of participants knows they are receiving the intervention, while another group does not believe they receive the intervention. Recent studies of analgesics employed such a design. Using a truncated version of the balanced placebo design\(^6\), Benedetti and a team of researchers at the University of Turin treated patients “overtly” and “covertly” for postoperative pain, Parkinson’s, and anxiety. I will focus on the case of postoperative pain.

In a study of pain (Benedetti et al. 2004) Benedetti’s team used four common painkillers - buprenorphine, tramadol, ketorolac, and metamizol - on a total of 278 patients who had undergone thoracic surgery for different pathological conditions. The postoperative patients were taken to have provided their ‘informed consent’ when they were “told that they could receive either a painkiller or nothing depending on their postoperative state and that they will not necessarily be informed when any analgesic treatment will be started” and they agreed to be in the study. “In this way, patients [did] not know if or when the treatment [was] given.” (Benedetti et al. 2004, p. 680).

The patients were then, of course unbeknownst to them, randomized into “overt” and “covert” groups with sex, age, weight, and pain baseline-balanced. The “overt” group was treated by doctors who “gave the open drug at the bedside, telling the patient that the injection was a powerful analgesic and that the pain was going to subside in a few minutes” (Benedetti et al. 2004, p. 681). Then, one dose of analgesic\(^7\) was administered every 15 minutes until a 50% reduction of pain (from baseline) was achieved for each patient. The “covert” group, on the other hand had the analgesic delivered by a pre-programmed infusion machine (already attached to the patient) without any doctor or nurse in the room. The pain reduction for both sets of patients

\(^5\) These are also commonly called “expectation effects”. However, there is a debate about whether it is the expectation, conditioning, or meaning of the treatment that are responsible for the effects (Moerman and Jonas 2002; Kirsch 2004; Benedetti et al. 2004). I will not delve into this debate here. The salient feature of all these possible mechanisms is that the participant must have some kind of awareness that she is being treated.

\(^6\) See (Howick, forthcoming) for a full description of the balanced placebo design.

\(^7\) Doses of different analgesics were standardized according to a method described earlier (Benedetti et al. 1998).
was measured every 15 minutes on a 10-point subjective pain scale where 0 = no pain and 10 = unbearable pain.

The results were that over 30% more analgesic was required by the patients who were treated covertly ($p$-values ranging from 0.02 – 0.007 depending on drug). See the figure below for details.

**Figure 1.** The amount of analgesic required to reduce pain by 50% for buprenorphine (A), tramadol (B), ketorolac (C), and metamizol (D). From (Amanzio et al. 2001, p. 209).
Benedetti’s study has been criticized on the grounds that the patients in the covertly treated group may have detected when they were getting treated in spite of the attempt that it was done ‘covertly’. Some experimental drugs could be identifiable from their side effects quite independently of its effect on pain (Kirsch 2003). If some of the participants in the hidden’ group had strong suspicions that they were receiving an analgesic, this would tend to enhance the effects of the ‘hidden’ administration and make it more difficult for the effect of open administration to be greater, and hence to demonstrate a belief effect. If Kirsch’s worry is well-founded, then we would expect a reduction in the difference between open and hidden administration. Therefore (again, if Kirsch’s worry is well-founded), since the study already provided evidence for a difference between open and hidden administration (and hence expectations effects), we can conclude that the study provides even stronger evidence for expectation effects than is indicated by the results.

3.3. Beliefs of the Dispensers: When the ‘Pygmalion Effect’ is a Confounder

A classic, though non-medical example of how dispenser beliefs may have effects is the ‘Pygmalion experiment’

Pygmalion was the name of a Greek artist who sculpted a statue out of ivory and fell in love with it. Subsequently, the statue came to life. Likewise, it is thought that dispensers who seek a particular outcome can influence, perhaps in unconscious or subtle ways, whether it comes about.

In the spring of 1964, in a real public (state funded) elementary school that Rosenthal and Jacobsen call the ‘Oak School’ (the real name is withheld), experimenters administered the “Harvard Test of Inflected Acquisition” to all (>500) students in grades 1 to 5. Teachers were told that the test “predicts the likelihood that a child will show an inflection point or “spurt” [i.e. point of rapid academic improvement] within the near future” (Rosenthal and Jacobson 1992, vii). Teachers administered this test, but the tests were scored separately by two blind assessors. Then, the teachers were then given names of the students who were most likely to “spurt”.

As a reason for their being given the list of names, teachers were told only that they might find it of interest to know which of their children were about to bloom. They were also cautioned not to discuss the test findings with their pupils or the children’s parents” (Rosenthal and Jacobson 1992, p.70)

---

8 I am grateful to Dr. Rupert Sheldrake for bringing this example to my attention.
After a year, the same IQ test was administered by the teachers and graded by independent, blind assessors. The “spurters” improved significantly more than the others (see table below). The top 20% of the students named by the test improved in all areas significantly more than the other students (results summarized below).

### Table 2. Mean gain in Total IQ after One Year by Experimental- and Control-Group Children in each of Six Grades

<table>
<thead>
<tr>
<th>GRADE</th>
<th>CONTROL N</th>
<th>GAIN</th>
<th>EXPERIMENTAL N</th>
<th>GAIN</th>
<th>EXPECTANCY ADVANTAGE</th>
<th>IQ POINTS</th>
<th>ONE-TAIL p &lt; .05*</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>48</td>
<td>+12.0</td>
<td>7</td>
<td>+27.4</td>
<td>+15.4</td>
<td>0.002</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>47</td>
<td>+7.0</td>
<td>12</td>
<td>+16.5</td>
<td>+9.5</td>
<td>0.02</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>40</td>
<td>+5.0</td>
<td>14</td>
<td>+5.0</td>
<td>-0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>49</td>
<td>+2.2</td>
<td>12</td>
<td>+5.6</td>
<td>+3.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>26</td>
<td>+17.5 (+)</td>
<td>9</td>
<td>+17.4 (-)</td>
<td>-0.0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>45</td>
<td>+10.7</td>
<td>11</td>
<td>+10.0</td>
<td>-0.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TOTAL</td>
<td>255</td>
<td>+8.42</td>
<td>65</td>
<td>+12.22</td>
<td>+3.8</td>
<td>0.02</td>
<td></td>
</tr>
</tbody>
</table>

* Mean square within treatments within classrooms = 164.24

The results become more dramatic in a chart (see below)

**Chart 3. Effect of Teacher Expectancy Measured as IQ Score Improvement**

---

9 This is surely a major criticism of the IQ test used since IQ is supposed to be inherent and not subject to environmental influence - the argument would be that even if we assume that this is a case where the expectations of the teachers had a real effect this could not have been on ‘real’ IQ and if it seemed to be such an effect on this particular test then the test was not a real test of IQ.

10 Recreated from table 7.1 in (Rosenthal and Jacobson 1992, p. 75). Note that the number of participants in the experimental group is not exactly 20% of the total. This is because “it was felt more plausible if each teacher did not have exactly the same number or percentage of her class listed” (Rosenthal and Jacobson 1992, p.70).
It is interesting that the “spurters” also scored higher in other tests, such as verbal IQ, reasoning IQ. Even their behaviour and social adaptability was improved relative to the “non-spurters”.

In fact the test was a standard IQ test, and the 20% of students who were predicted to “spurt” were chosen completely at random.

The Oak School experiment suggests that the expectations of teachers (and students) can have objective effects on student performance. More generally it suggests that “one person’s expectation for another person’s behavior can quite unwittingly become a more accurate prediction simply for its having been made” (Rosenthal and Jacobson 1992, vii).11

The mechanism of Pygmalion effects is not necessarily mysterious. A teacher, believing that a student was ready to ‘spurt’ might pay special attention to that student which could easily translate into accelerated rates of improvement. At the same time, the scarce resources spent on the ‘spurters’ is not ‘wasted’ on those less likely to improve12.

If there are ‘Pygmalion effects’ in medicine, then if a dispenser believes that she is administering the best experimental treatment (as opposed to placebo) to a patient, then her belief may translate into improved outcomes of the experimental treatment that have nothing to do with its characteristic13 features. A caregiver, believing that an extremely ill participant was being given a great new treatment,

---

11 Interestingly, Popper uses the name Oedipus Effects for the same phenomenon as Pygmalion effects (it is unclear whether Popper was familiar with the Pygmalion experiments): “Years ago I introduced the term ‘Oedipus effect’ to describe the influence of a theory or expectation or prediction upon the event which it predicts or describes: it will be remembered that the causal chain leading to Oedipus’ parricide was started by the oracle’s prediction of this event. This is a characteristic and recurrent theme of such myths, but one which seems to have failed to attract the interest of the analysts, perhaps not accidentally” (Popper 1969, 1.II, footnote 3).

12 The Pygmalion experiments are an interesting comment on IQ tests, which are supposed to be a measure of ‘innate intelligence’. Quite obviously if the tests are sensitive to teachers’ attention then they do not measure ‘innate’ intelligence accurately.

13 The term ‘characteristic’ is borrowed from Grünbaum (1986) and is used to describe features of a treatment process that are sometimes referred to as ‘specific’ or ‘active’. For example, fluoxetine would be the characteristic feature of treatment involving Prozac. Other features, including the prescription from a sympathetic physician, would be ‘incidental’.
coupled with the belief that the new treatment is effective, might provide that patient with a higher quality of care. On the other hand, if the caregiver believed that a different patient was being given a placebo, the dispenser might not bother providing the highest quality of care – it might be ‘not worthwhile’, especially given that they all have scarce resources to distribute amongst their many patients. An obvious scenario where dispenser knowledge could have effects is if the dispenser has an interest in showing that the experimental treatment works. For example, if the intervention administrator discovered the experimental treatment and stands to become famous if it is proven effective, or if she has a financial stake in the experimental treatment then their knowledge of which is the experimental intervention could affect the quality of care they provide. The role of these personal or financial interests could be either conscious or, more charitably, unconscious.

Benedetti’s pain study and the Pygmalion study show that at least in some circumstances participant and dispenser beliefs seem to have genuine effects. Double masking would clearly rule out these effects if they are confounding.

The rationale for double masking can therefore be summarized as follows. At least at the start of a double-blind trial, agents are told (and presumably believe) that they have an equal chance of being in the experimental or control group. This prevents the potential effects of participant or dispenser beliefs from confounding the study.

I will now examine cases where expectations in the experimental and control group might not be confounding, and therefore where double masking might add methodological value.

4. Why Double Masking is Less of a Worry for Active Rather than Placebo Controlled Trials

A confounding factor is one that

1. is unrelated to the experimental intervention,
2. is a determinant of the outcome, and
3. is unequally distributed between experimental and control groups

In this section I will argue that the worry about patient expectations and dispenser attitudes being unequally distributed in experimental and control groups is not as great in active controlled trials as it is in placebo controlled trials.
If participant beliefs about effectiveness are the same in the two arms of an active controlled trial (and after all neither treatment is presented as mere placebo) then there would of course be no confounding by them if the trial were open. My expectations regarding aspirin and a new NSAID\textsuperscript{14} for mild headache, for example, are likely to be the same and any trial comparing the two would not be confounded by failing to remain single masked.

But are those beliefs likely to be identical in general? Research (Chalmers 1997) suggests that they are not. People seem to believe that \textit{ceteris paribus}, the latest intervention is more effective than the older standard intervention, in spite of the fact that the new interventions are more likely to be \textit{less effective} than the older interventions. In an open active controlled trial, if everyone in the trial believed that the latest intervention was best, then the beliefs and effects of beliefs could well be different in the test and control groups. Or, take an imaginary trial that compared an alternative therapy and a conventional therapy for asthma. If, for some reason, the beliefs about the positive effects of alternative therapies were greater than the beliefs about the positive effects of conventional therapies (perhaps because of the often extended consultations provided by alternative therapies), then knowledge of which group a participant were in could confound the study.

I will call beliefs about the effectiveness of treatments that have nothing whatsoever to do with the characteristic effects of the intervention ‘prejudice’. Once prejudice is recognized as a potential confounder, it can be controlled for, at least partially. For example, participants with preferences for the experimental intervention could be introduced in equal measure to both groups. In the imaginary trial of a conventional versus alternative treatment, those who preferred alternative therapies could be divided equally among the ‘conventional’ and ‘alternative’ groups. Restricted randomization is probably the easiest way to achieve this, but it is also possible to adjust \textit{post hoc} for prejudice even in observational studies.

Or, a more sophisticated way to control for prejudice is to employ the so-called ‘patient preference’ design. Here

Patients may be placed in one of three groups according to preference and willingness to be randomised: (a) patients who have no strong preferences and therefore consent to randomisation; (b) patients with a preference who still consent to randomisation; and (c) patients who refuse randomisation and opt for their treatment of choice (Torgerson and Sibbald 1998)

\textsuperscript{14} Non-Steroidal, Anti-Inflammatory Drug, such as ibuprofen.
In short, patients with explicit preferences would be given the opportunity to opt out of the randomized trial and receive their preferred intervention.

To be sure, people’s ‘prejudices’ may well not be apparent to them. If not then it will be impossible to control for them. Still, controlling for participants’ ‘conscious’ prejudice will reduce any potential confounding effects of prejudice even it is not altogether eliminated.

Moreover it wouldn’t be patients’ prejudice that would worry in an open actively controlled trial of regular (non-alternative) treatment, but rather the beliefs of the investigators (of course then subtly conveyed to the patients). Controlling for patient prejudice at the outset of the trial would not reduce the potential confounding effect of the dispensing investigators. However, even dispenser beliefs could be controlled for, at least partially, in the same way that participant beliefs are controlled for. For instance, dispensers with strong beliefs about the positive effects of the experimental intervention could be balanced out by investigators with strong beliefs about the positive effects of the standard control. Then, of course, masked assessment should be employed wherever possible especially if the assessor is performed by the dispensers.

These strategies for reducing confounding of participant expectation and dispenser attitudes, although not perfect, are nonetheless surely useful. These strategies will not generally be successful for placebo controlled trials, where it would be difficult to find any participants at all who have a prejudice for placebo treatment. Prejudice may be exacerbated or caused by what I will call ‘hype’. Using Jefferson’s word quoted at the outset of this paper, some treatments are more fashionable than others. This could be due to the fact that the older, standard treatments have accumulated an extensive side effect profile while the new treatment represents hope. This hope is sometimes boosted by aggressive marketing. For example, Prozac, an SSRI antidepressant, was marketed to consumers in the United States before it was approved for marketing (a practice illegal outside the United States and New Zealand). Eli Lilly, the company who developed Fluoxetine, chose the name Prozac, which was unrelated to the characteristic chemical compound that was supposed to be responsible for the effect on serotonin levels. Previous to Prozac, drug names were often related to the characteristic chemical. The name Prozac was probably chosen because of the positive connotations of its prefix. The end result was that the number of ‘depressed’ people demanding prescriptions (some even before they were approved
for marketing) for SSRIs such as Prozac skyrocketed to levels unheard of in the past (Block 2007; Healy 2004, introduction).

Two other historical facts are worth noting. First, the side effects of the older class of antidepressants, known as tricyclics, that had been developed in the late 1950’s, were well known. Second, depression as a disease was becoming more widely recognized, some argue because of the marketing of depression by the same companies that manufactured the new generation of antidepressant drugs (Gardner 2003). As a result, more people were thinking of themselves as depressed, and they had good reason to be wary of the older drugs. It would be reasonable to expect an open active controlled trial of Prozac versus an older tricyclic to be biased because people had been persuaded to believe that Prozac was (potentially) a much better option than the older treatment. If the vast majority of potential participants in a trial had been ‘hooked’, then it would be difficult to have an equal number of Prozac and standard treatment supporters in each group. However assuming that there were at least some who were not ‘hooked’ on Prozac before the trial began, it would be possible to adjust, post hoc, for this potential confounder.

Note that although I have discussed the case where the new, ‘fashionable’ treatment has been hyped and that this hype may have effects that have nothing whatsoever to do with the characteristic features of the treatment, the same phenomenon could occur with an older intervention. For simplicity, I will continue to discuss the more common case where the hype has the potential to exaggerate the effects of the new experimental intervention.

Further, although it is difficult to control for the inevitable effects of hope that some new intervention will be more effective or less harmful than an older intervention or than no intervention without double masking or placebo controls, it is important to separate this hope from unjustified advertising of new interventions. If the advertising is conducted before there is sound evidence to support the view that that new interventions are safe and effective\textsuperscript{15}, then it is false that we have any grounds to believe that an experimental treatment is more effective than an older one.

Indeed there are equally good reasons to fear new treatments as there are to expect that they will be better. Because early trials are usually relatively short, or at

\textsuperscript{15} I leave a discussion of whether this trust is justified to another study. Suffice it to note that at least in the case of the FDA, the bias is towards approval of the new treatments. Indeed until 1962 a new drug merely had to demonstrate safety, and not effectiveness. They still do not need to demonstrate greater effectiveness than the best existing treatment.
any rate often too short to pick up long-term side effects, there are equally good reasons to fear newer experimental treatments as there are to hope that they are better than standard care.

The problem with hype that I have described is thus partly a problem with direct to consumer advertising (DTA) as currently practised, either before or after approval for marketing. Surely the just thing to do is to provide complete information about all available alternatives to consumers. Nonetheless, in cases where hype surrounds a new treatment, it is surely best to attempt to keep the trial double masked or to carefully adjust for this potential confounder.

To sum up this section, there are good reasons to believe that participant expectation and dispenser attitudes have less potential for confounding an active controlled trial than they do in placebo controlled trials. This is because the choice in an active controlled trial is between two potentially effective non-placebos, rather than between potentially effective non-placebo and placebo. Nonetheless, people often seem to have greater expectations regarding the effectiveness of the newer treatment, in spite of the fact that newer treatments are not usually more effective. In some cases these differential beliefs can be controlled for explicitly at the outset of the trial, or in post hoc adjustment.

Another circumstance where participant expectations and dispenser beliefs will may not confound a study as much as we think is where these potential confounders have no actual effects. I will consider this case now.

5. Where Participant Expectation and Dispenser Attitude Do Not Confound a Study

In this section I will outline cases where participant and dispenser beliefs do not affect the outcome of the study in a way that can be called confounding. These situations are where the belief effects are on the ‘causal pathway’ of the characteristic features, and any trial where the characteristic effects swamp the possible effects of participant and dispenser beliefs.

Even if the effects of beliefs are different in the test and control groups, if beliefs are characteristic rather than incidental features\textsuperscript{16}, then they cannot be

\textsuperscript{16} I am referring to the characteristic/incidental distinction originating from Grunbaum (1986). According to my earlier discussion this distinction is made by an underlying therapeutic theory. However for present purposes what I mean by ‘characteristic’ and ‘incidental’ matches the more common ‘active’ or ‘specific’ used in the medical literature to distinguish between placebogenic and non-placebogenic features of a
considered confounders. To see why, imagine there were a drug whose only characteristic feature was a chemical that was dramatically effective at making any depressed person who took it believe that the intervention had powerful characteristic effects that cured depression. Call the characteristic feature of this new drug the \( x \)-factor. Because depression is probably particularly sensitive to beliefs, the drug containing the \( x \)-factor may well prove very effective for the treatment of depression. However the drug has no other characteristic features for changing the chemicals, such as serotonin, that are currently believed to be correlated with, or cause, depression. Imagine that this drug demonstrated significantly superior effectiveness to standard SSRI antidepressants. The only reason that the new drug was more effective was because of participant belief. Adopting the rule that all belief effects are confounding would lead one to claim that the imaginary study was confounded by the different beliefs even though the beliefs were a direct result of the characteristic feature. In short, if the increased expectations in the experimental group arise from the characteristic features of the experimental treatment, then they cannot be considered confounding.

Participant and dispenser beliefs might not be a worry where their potential confounding effect is large relative to the size of the characteristic effects of the test intervention. Acute appendicitis or meningitis might well be influenced by beliefs, but it is unlikely that the effects of belief are significant relative to the effect of the treatment. For this reason, appendectomies and antibiotics for meningitis have of course never been tested in double masked conditions. The effects of a participant believing he is receiving an effective intervention may well have some effects, but it is unlikely that these effects would be strong enough to explain avoiding death from acute appendicitis. As Smith and Pell (2003) imply, you don’t need a double masked RCT to know that parachute use prevents death in someone falling from a plane.

The case where the characteristic effects appear dramatic explains the Phillip’s Paradox stated at the outset of this paper where dramatically effective treatments cannot be tested in double blind conditions. At least according to the view that double masked studies are of higher quality than open studies, paradoxically, dramatically effective treatments, are not supported by the best possible evidence. However where the treatment effect is dramatic (such as appendectomy for acute appendicitis), it is treatment process. For instance, fluoxetine in a Prozac pill will be the only characteristic feature in treatment with a Prozac pill. The incidental features in this case might be the manner in which the pill is delivered, etc.
safe to assume that expectations or attitudes could not account for the entire effect. Hence, although participant expectations and dispenser attitudes might have confounded the study, they were not sufficiently powerful to present a rival hypothesis for the entire effect. In short, the potentially confounding effects of expectations and attitudes in trials of dramatically effective treatments are relatively insignificant. Therefore, the Phillip’s Paradox dissolves once we abandon universal adherence to the rule that double masking increases quality and instead evaluate studies based on ‘scientific common sense’, namely the view that good evidence rules out plausible rival hypotheses.

In yet another circumstance, participant and dispenser beliefs might not have significant effects; if not, then double masking will not add methodological value. The interesting studies of participant expectation or belief conducted by Benedetti’s team, along with the Pygmalion studies and studies of different coloured placebos, might lead one to believe that participant and dispenser attitudes can, and often do, play a significant role. However, the dispenser, or teacher effects in the Pygmalion studies tapered off as the students aged (see table 5.1). Likewise, Benedetti’s study, although it showed significant effects of participant belief, was restricted to studies of a few ailments and did not show clinically relevant effects. In this section I will argue that current evidence suggests the magnitude of participant and dispenser beliefs varies quite widely. In some cases, they might not have any effects at all while in others they may well have clinically relevant effects. Double masking will help in the latter, but not the former cases. The structure of my argument is as follows:

1. If there are participant and dispenser beliefs then there are placebo effects
2. There is evidence that placebo effects are insignificant outside treatments whose outcomes are subjective and continuous outcomes.
3. Therefore, there is evidence that participant and dispenser belief effects are insignificant in some cases (such as for certain objective outcomes).

If we identify the set of incidental features what produces any overall placebo effect that there may be in particular circumstance, then if participant and dispenser beliefs

\[17\] In theory, if participant and dispenser belief effects are offset by the effects of other incidental features, there would be no overall placebo effects even if there were participant and dispenser belief effects. However in practice this is unlikely to be the case.
have effects it follows that there are *placebo effects* – since no ‘regular’ therapeutic theory gives a ‘characteristic’ role to such beliefs (at any rate in regular pharmacological treatments, where the characteristic features are exclusively the ‘active’ chemicals. In two recent meta-analyses, Asbjørn Hróbjartsson and Peter Gøtzsche (2001) looked at 3-armed trials that included experimental, placebo control, and untreated groups and found no overall significant placebo effect. If we assume that the three groups were always free from selection bias, that the “untreated” groups were actually untreated, and the placebo controls were legitimate, the placebo effect could fairly be estimated as the difference between the average effect in the placebo group less the average effect in the untreated group.

Defining placebos “practically as an intervention labelled as such in the report of a clinical trial” (Hróbjartsson and Gøtzsche 2001, p. 1595) Hróbjartsson and Gøtzsche searched several major medical databases\(^\text{18}\) for 3-armed RCTs. They excluded studies where participants were paid or were healthy volunteers, where the outcome assessors were unmasked, where the dropout rate exceeded 50%, and when “it was very likely that the alleged placebo had a clinical benefit not associated with the ritual alone (e.g. movement techniques for postoperative pain)” (Hróbjartsson and Gøtzsche 2001, p. 1595). (I will discuss whether their exclusion criteria were justified after outlining the study.)

After identifying 727 potentially eligible trials, they excluded 404 for not being randomized, 129 for failing to have a placebo group or an untreated group (although they were described as having one), 29 for being reported in more than one publication, 11 for using unmasked outcome assessment, 24 for meeting other exclusion criteria such as high dropout rates. Sixteen trials did not include relevant outcome data. This left 114 trials for the meta-analysis. Typical pill placebos were lactose pills, typical ‘physical’ placebos were procedures performed with the machine turned off (e.g. sham transcutaneous electrical nerve stimulation), and typical psychological placebo was theoretically neutral discussion between participant and dispenser. Over 40 clinical conditions were included in the analysis, ranging from hypertension and compulsive nail biting to fecal soiling and marital discord.

They classified the trials according to whether the outcomes were binary (“yes” or “no”) or continuous, and whether the outcomes were subjective or objective. For binary outcomes, they calculated the relative risk of an unwanted outcome, which

\(^{18}\) Medline, EMBASE, PsychLIT, Biological Abstracts, and the Cochrane Controlled Trials Register up to 1998.
is the ratio of the number of participants with an unwanted outcome to the total number of patients in the placebo group divided by the same ratio in the untreated group. A relative risk below 1 therefore indicates a positive placebo effect. With continuous outcomes, the authors calculated the standard mean difference, the difference between the mean value for an unwanted outcome in the placebo group and for the no treatment group, divided by the pooled standard deviation. A value of $-1$ indicates that the mean in the placebo group was 1 standard deviation below the mean in the untreated group. The results are summarized in the two tables below.

### Table 1. Effect of Placebo in Trials with Binary or Continuous Outcomes

(From Hrobjarsson and Gotzsche 2001, p. 1596)

<table>
<thead>
<tr>
<th>Outcome</th>
<th>No. of Participants</th>
<th>No. of Trials</th>
<th>Pooled Relative Risk (95% CI)</th>
<th>Pooled Standardized Mean Difference (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binary</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>3795</td>
<td>32</td>
<td>0.95 (0.88 to 1.02)</td>
<td></td>
</tr>
<tr>
<td>Subjective</td>
<td>1928</td>
<td>23</td>
<td>0.95 (0.89 to 1.06)</td>
<td></td>
</tr>
<tr>
<td>Objective</td>
<td>1867</td>
<td>9</td>
<td>0.91 (0.80 to 1.04)</td>
<td></td>
</tr>
<tr>
<td><strong>Continuous</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Overall</td>
<td>4730</td>
<td>82</td>
<td>-0.28 (-0.38 to -0.19)</td>
<td></td>
</tr>
<tr>
<td>Subjective</td>
<td>3081</td>
<td>53</td>
<td>-0.36 (-0.47 to -0.25)</td>
<td></td>
</tr>
<tr>
<td>Objective</td>
<td>1649</td>
<td>29</td>
<td>-0.12 (-0.27 to 0.03)</td>
<td></td>
</tr>
</tbody>
</table>

*CI denotes confidence interval.

†The relative risk was defined as the ratio of the number of patients with an unwanted outcome to the total number of patients in the placebo group, divided by the same ratio in the untreated group. A value below 1.0 indicates a beneficial effect of placebo.

‡The standardized mean difference was defined as the difference between the mean values for unwanted outcomes in the placebo and untreated groups divided by the pooled standard deviation. A negative value indicates a beneficial effect of placebo.

Although there was significant ($p=0.003$) heterogeneity among trials with binary outcomes, placebo did not have a significant effect... (overall pooled risk of an unwanted outcome with placebo 0.95; 95 percent confidence interval, 0.88 to 1.02). For continuous outcomes, there was a significant placebo effect for trials with subjective outcomes (-0.36; 95% confidence interval $-0.47$ to $-0.25$), but not for trials with objective outcomes (-0.12; 95% confidence interval $-0.27$ to 0.03). There was also significant heterogeneity ($p=0.001$) for trials with continuous outcomes. However, there was significant ($p=0.05$) relationship between size of trial and placebo effect, indicating that bias due to small trials played some role. Of all the ailments that
were treated in more than 3 trials, only pain showed a significant effect (-0.27; 95% confidence interval -0.40 to -0.15 – see table below).

**Table 2. Effect of Placebo on Specific Clinical Problems** (From Hróbjartsson and Gøtzsche, 2001, p. 1597)

<table>
<thead>
<tr>
<th>Outcome</th>
<th>No. of Participants</th>
<th>No. of Trials</th>
<th>Pooled Relative Risk (95% CI)†</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Binary</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nausea</td>
<td>182</td>
<td>3</td>
<td>0.94 (0.77 to 1.16)</td>
</tr>
<tr>
<td>Smoking</td>
<td>887</td>
<td>6</td>
<td>0.88 (0.71 to 1.09)</td>
</tr>
<tr>
<td>Depression</td>
<td>152</td>
<td>3</td>
<td>1.03 (0.78 to 1.34)</td>
</tr>
<tr>
<td><strong>Continuous</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pain</td>
<td>1602</td>
<td>27</td>
<td>-0.27 (-0.40 to -0.15)</td>
</tr>
<tr>
<td>Obesity</td>
<td>128</td>
<td>5</td>
<td>-0.40 (-0.92 to 0.12)</td>
</tr>
<tr>
<td>Asthma</td>
<td>81</td>
<td>3</td>
<td>0.34 (0.83 to 0.14)</td>
</tr>
<tr>
<td>Hypertension</td>
<td>129</td>
<td>7</td>
<td>-0.32 (-0.78 to 0.13)</td>
</tr>
<tr>
<td>Insomnia</td>
<td>100</td>
<td>5</td>
<td>-0.26 (-0.66 to 0.13)</td>
</tr>
<tr>
<td>Anxiety</td>
<td>257</td>
<td>6</td>
<td>-0.06 (-0.31 to 0.18)</td>
</tr>
</tbody>
</table>

*Only problems addressed by at least three trials are included. CI denotes confidence interval.
† The relative risk was defined as the ratio of the number of patients with an unwanted outcome to the total number of patients in the placebo group, divided by the same ratio in the untreated group. A value below 1.0 indicates a beneficial effect of placebo.
‡ The standardized mean difference was defined as the difference between the mean values for unwanted outcomes in the placebo and untreated groups divided by the pooled standard deviation. A negative value indicates a beneficial effect of placebo.

In sum, there were significant placebo effects for trials of pain (where the outcome measure was subjective) and in general for trials of ailments with subjective continuous outcomes. Even in these cases Hróbjartsson and Gøtzsche even question whether these studies provide evidence of placebo effects.

Patients in an untreated group would know they were not being treated, and patients in a placebo group would think they were being treated. It is difficult to distinguish between reporting bias and a true effect of placebo on subjective outcomes, since a patient may tend to try to please the investigator and report improvement when none has occurred. The fact that placebos had no significant effects on objective continuous outcomes suggests that reporting bias may have been a factor in the trials with subjective outcomes” (Hróbjartsson and Gøtzsche 2001, p. 1597).

Indeed the so-called ‘Hawthorne Effect’, which is part of what we mean when we talk about placebo effects, is, very briefly, the positive effect of being in an experiment no
matter what the intervention is. If there are Hawethorne Effects, then we might expect them to be greater in the placebo control group than in the ‘no treatment’ group. If so then we would expect the placebo effect enhanced relative to no treatment.

Hróbjartsson and Gøtzsche conclude that there is:

little evidence that placebos in general have powerful clinical effects. Placebos had no significant pooled effect on subjective or objective binary or continuous objective outcomes. We found significant effects of placebo on continuous subjective outcomes and for the treatment of pain but also bias related to larger effects in small trials. The use of placebo outside the aegis of a controlled, properly designed clinical trial cannot be recommended (Hróbjartsson and Gøtzsche 2001, p. 1599).

There are several problems with Hróbjartsson and Gøtzsche’s meta-analysis. First, it can be questioned whether the statistical results of the study warrant their concluding statements. Overall, there was a significant placebo effect in trials with continuous outcomes, and there were more participants in trials with continuous outcomes than in trials with binary outcomes. This would surely have tempted some authors to claim that there are often significant placebo effects rather than conclude that there is little evidence for significant placebo effects.

Then, Hróbjartsson and Gøtzsche’s warning that significant placebo effects for pain and subjective continuous outcomes may be due to bias is contrived. If treated in a way that made them feel they were in a trial, the ‘untreated’ group may have experienced Hawethorne Effects as well, in which case the apparent placebo effects would be reduced. Hróbjartsson and Gøtzsche admit that if participants in the ‘untreated’ groups sought treatment outside the study, the apparent placebo effects would have been reduced (Hróbjartsson and Gøtzsche 2001, p. 1597). In short, Hróbjartsson and Gøtzsche’s claim that bias may have affected the studies could work

---

19 The Hawethorne Effect is named not after an author, but after the name of a series of experiments in the Hawthorne works of the Western Electric Company in Chicago between 1924 and 1933. In one study, for example, the illumination remained stable for the control group and was slowly raised for the experimental group. Productivity increased equally in both groups. In another study, illumination was stable in the control group but was slowly lowered in the experimental group. Productivity increased steadily and equally in both groups until the lights were so low in the experimental groups that the experimental workers protested and production fell off (Roethlisberger and Dickson 1939). The “Hawthorne Effect” has been interpreted many ways. Here I will take it to be the potential (sometimes temporary) positive effects of being in an experiment.

20 Masking the assessors and manuscript writers may have prevented this possible bias.
either to enhance or reduce the estimate of the true placebo effect, and not simply to reduce it as they suppose.

Further evidence of Hróbjartsson and Gøtzsche’s bias is evident elsewhere. They state that “It surprised us that we found no association between measures of the quality of a trial and [increased] placebo effects” (Hróbjartsson and Gøtzsche 2001, p. 1598). One could just as easily be surprised to find no association between measures of the quality of a trial and decreased placebo effects. Poor trial quality is usually associated with exaggerated treatment effects. One way to exaggerate treatment effects is for the placebo effect to decrease. On the other hand, increase of trial quality could mean that a trial is visibly (to the participant) more tightly controlled. This increase in control could lead to increased Hawthorne Effects, and hence increased placebo effects. In short, the quantitative results of the meta-analysis do not support the conclusion that there are no placebo effects.

Third, Hróbjartsson and Gøtzsche imposed insufficient normative constraints on what they took to be placebos and ‘no treatment’, which may have led to erroneous results. Recall that the estimate of the placebo effect is based on the average difference between ‘placebo’ and ‘no treatment’. But if the placebo controls were illegitimate, i.e. they had characteristic features or they didn’t have all the potentially effective incidental features, then the estimated ‘placebo’ effect could be erroneous. Or, if what they counted as ‘no treatment’ in fact amounted to some form of placebo or treatment, then the ‘placebo effect’ will have been underestimated. A superficial glance at what the authors counted as ‘placebos’ or ‘no treatment’ controls suggests that their criteria for what counted as ‘placebo’ controls or ‘no treatment’ may well have led to mistaken results.

Initially the authors attempt to skirt around the thorny issue of defining placebos and claim to define placebos as whatever is called a placebo in a report of a clinical trial. Yet later they exclude the trial if “it was very likely that the alleged placebo had a clinical benefit not associated with the ritual alone (e.g. movement techniques for postoperative pain)” (Hróbjartsson and Gøtzsche 2001, p. 1595). That is, they rightly excluded the trial if the placebo control was illegitimate because it included characteristic features of the test treatment (see my ‘Placebo Controls’, forthcoming). Although they can be forgiven for giving up their initial claim that they would define placebos ‘practically’ as whatever was used as a placebo control in a trial, the fact that they had no coherent rules for deciding what counts as a legitimate
placebo made their choice unsystematic and possibly biased. In fact, they jumble (along with placebo pills and injections) relaxation (described as a placebo in some studies and a treatment in others), leisure reading, answering questions about hobbies, newspapers, magazines, favourite foods and sports teams, talking about daily events, family activities, football, vacation activities, pets, hobbies, books, movies, and television shows as placebos (Kirsch 2002). In short, although their exclusion criteria are legitimate, they need to go further before we can accept that the placebos they examined are legitimate. If they are not, then the results of the meta-analysis cannot be relied upon.

A parallel problem that has yet to be pointed out in the literature is the failure to impose restrictions on the untreated groups. If the ‘untreated’ groups did something having a clinical benefit, this will reduce the difference between ‘no treatment’ and placebo and hence the estimated placebo effect. On the other hand, if the ‘untreated’ participants were closely monitored, then placebo effects, such as Hawthorne Effects could have resulted. Either way, the effects of being left ‘untreated’ may have been exaggerated, which would have led to an underestimation of placebo effects.

Further, the highly significant heterogeneity of the interventions studied calls into question what conclusions can be drawn from the meta-analysis. An overall finding of insignificant placebo effects does not count against the reasonable view that placebo effects are common and powerful for certain disorders (those whose outcomes are measured subjectively) but not for others. Indeed had the Hróbjartsson and Gøtzsche study used masking of the data collectors, statisticians, and manuscript writers it is unclear whether their conclusions or the title of their work would have been the same.

There is also a suggestion implicit in their work that ‘subjective’ means ‘unreal’. An outcome such as decreased pain can be very important clinically even though it can only be measured (at any rate directly) ‘subjectively’. This also ties in with what they call bias. Although ‘bias’ in the reporting of subjective outcomes may be a real worry, it is not even clear to what extent it is possible to make sense of claims that a patient reports decreased pain to ‘please’ the investigator, but allegedly, in fact, is not experiencing ‘real pain’. In short, the subjective nature of the outcome does not make it ‘unreal’.

In reaction to criticisms, Hróbjartsson and Gøtzsche dug in their heels. They updated the 2001 meta-analysis in 2004, in a study which supposedly confirmed the results of the earlier meta-analysis (Hróbjartsson and Gøtzsche 2004). Then, in 2006
they reviewed several meta-analyses of apparent placebo analgesic effects and concluded that the analyses had been poorly done. In the meta-analysis update, Hróbjartsson and Gøtzsche recognize the problem that the ‘untreated’ participants may have been treated: “Patients in a no-treatment group also interact with treatment providers, and the patients are therefore only truly untreated with respect to receiving a placebo intervention” (Hróbjartsson and Gøtzsche 2004, p. 97). They also mention the problem of heterogeneity: “we cannot exclude the possibility that in the process of pooling heterogeneous trials the existence of such a group [of trials that showed a significant placebo effect] was obscured” (Hróbjartsson and Gøtzsche 2004, p. 97).

In spite of recognizing at least a few of the problems with their analysis, their conclusions about apparent placebo effects were not at all tempered. The conclusion of the updated paper was similar to that of the first:

In conclusion, we reproduced the findings of our previous review and found no evidence that placebo interventions in general have large clinical effects, and no reliable evidence that they have clinically useful effects. A possible effect on patient-reported continuous outcomes, especially on pain, could not be clearly distinguished from bias (Hróbjartsson and Gøtzsche 2004, p. 98).

Problems with the Hróbjartsson and Gøtzsche studies notwithstanding, there are two important points they highlight. First, their method, and not Beecher’s, for measuring the magnitude of belief effects is correct, if done properly. Measuring the difference between placebo and no-treatment, is the correct way to measure placebo effects. Second, it is possible, indeed probable, that the placebo effect varies depending on the intervention, the ailment, and the type of outcome.

Another empirical study needs mentioning before concluding this section. In a study of 33 meta-analyses Schulz et al. claim that the unmasked studies odds ratios were exaggerated by 17%. This study seems to suggest that participant expectation and dispenser attitudes (whose effects were supposedly neutralized in the masked

<table>
<thead>
<tr>
<th></th>
<th>Outcome event</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Control Group</td>
<td>a</td>
<td>b</td>
</tr>
<tr>
<td></td>
<td>a + b</td>
<td></td>
</tr>
<tr>
<td>Experimental Group</td>
<td>c</td>
<td>d</td>
</tr>
<tr>
<td></td>
<td>c + d</td>
<td></td>
</tr>
</tbody>
</table>

Odds ratio = (a/b) / (c/d) = ad / bc. The odds ratio is a measure of effect size. An increase in 17% of the odds ratio, however, does not mean an increase of 17% absolute effect size. A 17% increase in odds ratios can be due to a much smaller increase in absolute effect size.
studies) have a 17% effect. I will leave aside the fact that other studies of the effects of unmasking have had conflicting results (Moher et al. 1998; Miller, Colditz, and Mosteller 1989) to focus on a problem they all share. Schulz’s team failed to define which groups were masked. If, for example, it was the outcome assessors and not the dispensers that were masked in the masked studies and unmasked in the unmasked studies, their beliefs could have been responsible for the different effects. The authors recognized this shortcoming, admitting that the meagre information on the approaches used for double masking made it difficult to interpret what the source of exaggerated effects are (Schulz et al. 1995). In light of the ambiguity with which the term double masking has been used, it is difficult to interpret this study as evidence for participant and dispenser belief effects.

To sum up this section, double masking will not reduce the potentially confounding effects of participant and dispenser beliefs as much in active controlled trials as it will in placebo controlled trials. Then, there are cases where beliefs are not confounders. These cases are (a) where the beliefs are characteristic features of the treatment, (b) where the belief effects are small relative to the characteristic effects of the experimental treatment, and (c) where there are no belief effects.

I will now question whether double masking is a realistic goal.

6. The Near-Impossibility of Successful Double Masking

Attempting to double mask a study is one thing; keeping it successfully double masked for the duration of a trial is another. There is strong evidence to suggest that even when the best efforts are made, double masking is rarely successful for the duration of the trial. This is because to keep the trial successfully masked means that the appearance, smell, taste, and side-effects of the experimental intervention must be mimicked by the control. Otherwise, given the (usually enforced) requirement of full informed consent whereby participants are made aware of the nature of the experimental intervention (potential effects, side effects, etc.), participants and dispensing physicians will correctly guess whether a particular intervention is the experimental treatment or control. If it is the case that double masking is unlikely to be successful, then the apparent advantage of being double masked, this refers to attempts to keep the trial masked, is nullified and open trials are equally good.

In a recent study, Fergusson, Glass, Waring, and Shapiro (2004) investigated whether trials described as double masked were successful. Whether a trial is
successfully double masked can reasonably be ascertained by asking participants and dispensers to guess which participants were in the treatment or control groups—if guesses do significantly better than chance then there is a degree of ‘unmasking’. These authors conducted a Medline search of randomized, placebo controlled trials published from 1998 to 2001 in 5 top general medical and 4 top psychiatry journals. Their search turned up a total of 473 medical trials and 192 psychiatry trials. From this group they randomly selected 100 trials in each group. Nine of the randomly selected trials were excluded because they were not placebo controlled in spite of being described as such. They ended up with 97 medical trials and 94 psychiatry trials.

Of the 97 medical trials, only 7 provided evidence of the success of double masking. Of those, only 5 reported that the masking was unsuccessful. Five trials reported that the success of blinding was imperfect. The trial that did not present blinding data described blinding as successful without further comment. The trial that reported aggregated data did not comment qualitatively, or provide statistical tests of success of blinding (Fergusson et al. 2004).

Of the 94 psychiatry trials, 8 reported evidence of testing for successful masking. Four of these reported that the masking was unsuccessful. Overall:

15 of the 191 trials (8%) provided such information, be it qualitative or quantitative. Of the 15 trials, only five trials reported that blinding was successful and of these, three did not present any quantitative data analysis to support their claim (Fergusson et al. 2004).

This study suggests that masking is rarely tested for and, where tested for, rarely successful. It incited a flurry of responses on the BMJ website by prominent medical researchers such as Stephen Senn, David Sackett and Douglas Altman. Describing the

---


23 In spite of noting that double masking is important for active controlled trials, the authors limited their evaluation to placebo controlled trials.

24 The authors don’t define success. I will assume that it means that significantly more participants correctly guessed which group they were in than would be predicted by chance alone.

25 Fergusson et al. proceed to question whether the studies which provided quantitative data on the success of double masking were methodologically sound, and claim that they weren’t.
well-known ‘Tea Lady Experiment’ Senn claim that ‘unsuccessful’ masking is not a methodological failing as long as it is a result of the efficacy of the treatment:

The classic description of a blinded experiment is Fisher’s account of a woman tasting tea to distinguish which cups have had milk in first and which cups have had tea in first in support of her claim that the taste will be different. Here the efficacy of the treatment, order of milk or tea, is “taste” and the lady’s task is to distinguish efficacy. Fisher describes the steps, in particular randomization, that must be taken to make sure that the woman is blind to the treatment. But if he were to adopt the point of view of Fergusson et al, there would be no point in running the trial, since if the lady distinguished the cups, the trial would have been described as inadequate, as she has clearly not been blind throughout the trial. (Senn 2004)

What Senn claims is that, in cases where the participant (i.e. the ‘Tea Lady’) correctly guess which is the ‘experimental intervention’ (i.e. the milk first cups of tea) because they have identified the characteristic feature (i.e. whether milk was poured first), that it is mistaken to call the trial inadequate.

Although Senn may be mistaken about the classic description of a blinded experiment – the point of the Tea Lady experiment is to demonstrate the value of concealed random allocation – his point that successful masking is difficult to achieve with effective treatments is well made. Sackett echoes Senn’s view that testing for successful masking once a trial is underway is difficult:

Once a trial is under way, I'm not smart enough to separate the effectiveness of blinding from the effects of pre-trial hunches about efficacy, and I've never met anyone who is. Accordingly, we vigorously test for blindness before our trials, but not during them and never at their conclusion (Sackett 2004).

Sackett and Senn are surely correct that in some cases, namely where the characteristic features of the test treatment are so apparent that the trial becomes unmasked, should not count against the methodological value of the trial. On this view, investigators should attempt to make a double masked trial at the outset. Then, if the participants or dispensers correctly guess which intervention they are taking or giving, it is unimportant because it is often the characteristic features of the treatment that cause the unmasking.

Yet even if Sackett and Senn are correct, their arguments only apply in cases where the unmasking is due to the characteristic features of the experimental treatment. As I explained earlier, if the effects of the experimental treatment are
dramatic, then the ‘unmasking’ of the study cannot be said to have impaired its methodological quality. Yet, unmasking can occur because of many reasons other than characteristic effectiveness, dramatic or otherwise. For instance, cases where the side-effects of the experimental treatment are identifiable will lead to unmasking that does impair a study’s methodological quality. Where unmasking occurs for these other reasons, it does in fact call the adequacy of the trial into question. Further, both Sackett and Senn are incorrect that the different reasons for unmasking cannot be disentangled. Perhaps even more relevantly, ignoring other possible reasons for unmasking is unjustified. As Shapiro notes, “[it] seems contrary to an evidence based approach to avoid obtaining data because we have to struggle with its interpretation. (Shapiro 2004).

To illustrate, imagine a trial of an antidepressant drug whose characteristic chemical has no effects yet that has unique and recognizable side effects versus inactive placebo (one that does not mimic the side-effects of the experimental treatment). If this is the case then participants could guess when they are in the experimental arm of such a trial, which could lead to increased beliefs and expectations about recovery, and hence better effects.

In order to determine whether unmasking was due to the characteristic or non-characteristic features, the tests for successful masking could ask why participants believe that they were in a particular group. In short, if a trial comes unmasked for reasons other than characteristic effectiveness then it is legitimate to claim that double masking has not achieved the purpose of ruling out confounding participant and dispenser beliefs. Contrary to what Sackett implies, it is possible, at least to some extent, to discover the reasons for unmasking.

Another study was conducted earlier this year by a research team at the Nordic Cochrane Centre that made the case for the difficulty of keeping a trial successfully double masked even stronger. Hróbjartsson and colleagues randomly selected a sample of 1599 clinical trials from the Cochrane Central Register of Controlled Trials that were published in 2001. Thirty-one (2%) reported tests of the success (or otherwise) of masking. It was considered successful in 14 of the 31 studies, unclear in 10, and reported as unsuccessful in 7. The possibility of a biased result due to the unmasking was not addressed or dismissed in 6 of the 7 unsuccessful cases. In short, 2% of the trials were checked for double masking, and of the 2%, less than half reported double masking success.
To test whether the apparent failure to successfully to double mask was a problem with reporting, Hróbjartsson et al. selected a random sample of 200 trials not reported as having conducted tests for the success of masking, and asked the authors whether unreported tests had been done. Of the sample, 130 (65%) responded, and 15 (less than 12%) of these reported having conducted, but not reported, the tests. In short, masking was *reported* as successful in a very small proportion – perhaps as low as 1% - of 1599 trials. The authors conclude that “Blinding is rarely tested. Test methods vary, and the reporting of tests, and test results, is incomplete. There is a considerable methodological uncertainty about how best to assess blinding, and an urgent need for improved methodology and improved reporting” (Hróbjartsson et al. 2007). Trials that perform tests of the success of blinding but do not report them must be viewed with suspicion: successful double masking is something to be proud of. This study, combined with the earlier one suggest either that tests for successful double masking are rarely made, when made are often methodologically problematic. Most importantly, when tested for and reported, the results in a majority of cases reveal that double masking has been unsuccessful.

To make matters worse, there are good reasons to believe that the problem of successfully double masking a study may run deep. The question of whether studies are successfully double masked has been lurking in the background since I laid down the conditions required for legitimate placebo controls (see Howick, 2008). Recall that the placebo control, in order to gain legitimacy, had to be similar in superficial appearance (taste, smell, touch), side effects, and effectiveness to the experimental treatment. The practical difficulty in designing placebo controls that meet these conditions, especially with regards to obscure side effects, may be insurmountable.

For example, one side effect of SSRIs is increased probability of sexual dysfunction. How could this side effect be imitated? People could not. Even if there were some drug which raised the probability of sexual dysfunction that could be added to the control treatment, a further set of problems emerges. First, it would have to be established that the drug did not have effects (either positive or negative) for the target disorder. Second, it would be ethically problematic to deliberately cause harm with a control treatment. Third, there would still be the other side effects of the experimental intervention (most drugs have more than one side effect) and of the drug for sexual dysfunction.

Irving Kirsch and Guy Sapirstein (Kirsch and Sapirstein 1998) investigated whether the characteristic features of *Prozac* and other SSRIs were due to the belief
that SSRIs are effective, or to the characteristic (pharmacological) features of SSRIs themselves. They first noted that the side effects of SSRIs are different from the side effects of other drugs and placebos. Because of informed consent, participants in a trial are aware of these potential side-effects associated with the experimental and control treatments. Therefore, in spite of attempts to keep the trials double masked, participants often correctly suspect (and are in effect encouraged to suspect) when they are taking the experimental treatment.

In one example, Irving Kirsch and Guy Sapirstein compared the effect of established antidepressant drugs versus inactive placebo with other drugs versus inactive placebo for depression (Kirsch and Sapirstein 1998, p.7). They found that the other drugs (amobarbital, lithium, liothyronine, and adinazolam), which had no known characteristic antidepressant features, were at least as effective as established antidepressant drugs. This study could be interpreted to mean that, if we had a ‘complete’ therapeutic theory, we would identify characteristic features of the ‘other drugs’ that have positive effects on depression. Or, the hypothesis preferred by the authors, that the non-antidepressant physiological properties (side-effects) of these drugs lead to unmasking of the study and that the unmasking of the study was what accounted for the superior effectiveness of the test treatment over placebo. The participants, suffering from side effects of the experimental treatment (which they knew about because of fully informed consent or their own research), correctly guessed that they were receiving the experimental intervention as opposed to placebo, which led to increased expectations regarding recovery. Correspondingly, participants not suffering from side effects correctly deduced that they were taking the ‘placebo’ and had lower expectations regarding recovery. Hence, the increased benefit of the experimental treatment could have been due to these differing expectations: “these medications function as active placebos” (Kirsch and Sapirstein 1998, p.7). Although supported by some other independent studies (Moncrieff 2003), the study has not gone uncriticised (Klein 1998). I will not go into a detailed critique of the Kirsch/Sapirstein study here. The idea to be gleaned is that it is important for placebo control treatments mimic the side-effects of the treatments they are supposed to imitate. Otherwise, even if a study is initially carefully blinded, it may later, or, more often, sooner, come unmasked.

It could still be maintained that because it is possible in some cases, trials that attempt to keep the double mask are superior to those that do not. Further, because the problem with retaining the double mask has only been highlighted recently, we might
expect the rate of successful double masking to increase in the near future. Since double masking raises the probability of ruling out some confounding factors, a trial’s being described as double masked is a methodological value that makes it superior to open trials. However, as was pointed out in the previous sections this alleged potential advantage of double masking will not be actualized in all cases.

Furthermore, attempts to keep trials double masked have their costs. Keeping a trial double masked makes the conditions of the trial different from the conditions in routine clinical practice. In an attempted double masked placebo controlled trial, the dispenser administers an intervention that she cannot know for sure is an ‘effective’ (nonplacebo) treatment. The participant receives the intervention with a commensurate degree of doubt. This doubt may affect their beliefs and hence the effectiveness of the experimental or control intervention. In contrast, in routine clinical practice, the dispenser usually offers an intervention with confidence, and the patient believes that they are being treated with a nonplacebo.

It could be argued that the doubt about which treatment patients in a trial have been given does not make a difference to the estimated effects of the characteristic treatment. As long as the doubt is the same in both the test and control groups, then any reduced belief effects will ‘cancel out’ because they are reduced equally in the test and control group. Recall that in a placebo controlled trial the characteristic effects are calculated by subtracting the average effect in the test treatment group from the average effect in the control group. If both these groups have effects reduced by doubt, then as long as the doubt is the same in both groups, the effect estimate of the characteristic features will remain constant. An analogous argument applies to double masked active controlled trials.

The argument that the reduced belief effects ‘cancel out’, however, relies on the assumption of additivity whereby the component features of a treatment (participant and dispenser belief, other incidental factors, and characteristic factors), simply add up to form the composite effect. If additivity does not hold, then changing the effect of participant or dispenser belief could affect the overall effectiveness in a different way. I will discuss the assumption of additivity further in another work.

In sum, it seems safe to say that attempts to keep a study double masked may often fail. This may be because of the inherent difficulty in imitating an intervention’s sensory qualities and side effects. The apparent methodological advantage of being double masked at the outset of a trial may therefore be frequently illusory – trials are rarely successfully double masked. If many trials are not successfully double masked,
then the basis for considering double masked trials superior to open trials becomes
difficult to uphold. Further, the seemingly legitimate aim of double masking studies
where it seems reasonable has its costs.

7. Conclusion

The view that double masking always adds methodological value was exposed by the
Phillip’s Paradox whereby dramatically effective treatments are not supportable by
double masked, and hence ‘best’ evidence, to be wanting. In this paper I examined the
methodological values from the more fundamental, ‘scientific common sense’ point of
view that good evidence rules out rival hypotheses, and hence that better evidence
rules out more rival hypotheses.

Double masking, where successfully executed, rules out the potential
confounding factors of participant expectations and dispenser attitudes. However
there are three situations where these potential confounders are not actual confounders
(or not significant confounders). First, there are some situations where patient and
dispenser expectations do not have effects. Second, there are other cases, perhaps
some ‘active’ controlled trials, where the expectations could be the same for both
groups. Third, in cases where the experimental treatment has a dramatic effect, any
expectations might have (relatively) negligible effects.

My investigation in this paper also revealed that in practice, keeping trials
successfully double masked is usually difficult and sometimes impossible. This means
that the property of being described as double masked does not necessarily mean that
any methodological value has been added. In cases where it is likely that double
masking will not be successful (i.e. where the side effects of the experimental
treatment are common and difficult to imitate), ‘open’ trials lose their relative
disadvantage. Indeed, in these cases it may be better to seek other ways to control for
the potentially confounding effects of expectations and attitudes. For example, by
explicitly controlling for participant prejudice and by placing greater emphasis on
masking the outcome assessors and statisticians.
References


Chalmers, I. (1997), "What is the prior probability of a proposed new treatment being superior to established treatments?" *Bmj* 314 (7073):74-75.


Howick, J. (forthcoming), "The Importance of 'Legitimate' Placebo Controls".


Roethlisberger, F. J., and W. J. Dickson (1939), Management and the Worker: Harvard University Press.


Senn, S. J. (2004), "Turning a blind eye: authors have blinkered view of blinding", Bmj 328 (7448):1135-1136; author reply 1136.

Shapiro, Stan (2004), "Widening the field of vision", BMJ Rapid Responses.