ASSESSING BLINDING IN RANDOMIZED CLINICAL TRIALS

A Thesis

Presented to the

Faculty of

California State Polytechnic University, Pomona

In Partial Fulfillment

Of the Requirements for the Degree

Master of Science

In

Mathematics

By

Jesse C. Waite

2017

SIGNATURE PAGE

THESIS:	ASSESSING BLINDING IN RANDOMIZED CLINICAL TRIALS
AUTHOR:	Jesse C. Waite
DATE SUBMITTED:	Summer 2017
	Department of Mathematics and Statistics

Dr. Adam King
Thesis Committee Chair
Mathematics & Statistics

Dr. Alan Krinik
Mathematics & Statistics

Dr. Robin Wilson
Mathematics & Statistics

ACKNOWLEDGMENTS

I'd like to thank my adviser Dr. King for his direction and insight, my committee for their interest and insight towards my work, my parents for their example as to the importance of education and hard work, my wife Victoria for her loving patience in dealing with me through this process, and my children for not burning down the house during my frequent absences.

ABSTRACT

In the realm of randomized clinical trials, protocols intended to protect the knowledge pertaining to which treatment assignment each participant actually receives are usually employed. These protocols promote what is known as *blinding*. When these protocols are meant to obscure the assignment from the clinicians as well as the participant, this is known as a double blind study. It is widely held that successfully employing protocols to insure blinding will help to insure the results of the study are not subject to bias. This thesis will discuss some of the methods commonly included in the protocols regarding blinding and the assessment of its success as it pertains to randomized clinical trials. Three methods which have been and could be used to assess the success of blinding protocols will be analyzed. A simulation study comparing the three methods for assessing blinding using R will show the differences between these methods, and their strengths and weaknesses will be discussed. Finally the development and employment of a method for determining when unblinding occurs because appropriate protocols are not enacted or followed is discussed.

Contents

List of Tables

List of Figures

Chapter 1

A Gentle Introduction to Blinding

Randomized clinical trials are conducted with the goal of determining whether the particular treatment under investigation is superior to an existing, established treatment (or sometimes no treatment at all). To this end, participants are randomly assigned to either receive the treatment being investigated, a treatment whose efficacy is already well established, or a placebo. In order to protect the validity of the trial protocols are generally put in place to minimize fraud and bias. Our focus will be on assessing the success of the protocols employed to promote blinding.

1.1 The What of Blinding

The term *Blinding* is used to describe the intentional hiding of a particular participants treatment assignment when speaking of a clinical trial. The practice of employing blinding techniques is usually designed to reduce bias that might skew the final results of the study. When researchers are conducting a randomized clinical trial, i.e. a trial in which the assignment of a participant to the treatment group or the control group is done randomly, usually with the probability of assignment to either group being equal, a double-blind study is generally preferred. The phrase *double blind* refers to the participants not being given any indication whether they are assigned to the treatment group or the control group, as well as the same information being withheld from the clinicians who administer the treatments and monitor participant progress throughout the trial. A *triple blind* study would include the blinding of the statisticians responsible for performing the final data analysis. In this case the data would indicate only that a particular participant was assigned to treatment group A or B, etc. but it would not indicate which of those groups is the control or treatment group.

1.2 The Why of Blinding

In a randomized clinical trial, it is important to have a group of people that can be monitored that will not receive the treatment, but who are afflicted with the malady that the treatment is designed to treat. This group of people are commonly referred to as the control group. Those selected to actually receive the treatment are usually called the treatment group. Generally speaking, the participants in a randomized clinical trial, as well as the clinicians administering the treatments and assessing patient outcomes, are not made aware of which group individual participants are assigned to. Protocols generally indicate that the placebo which will be given to the control group, when appropriate, should be made to resemble the actual treatment as closely as possible. Some placebos are even constructed to produce side effects common to the actual treatment in order to help preserve blinding Freidman et al. (2010). In some cases an active control could be administered. In this case, a well established treatment is given to the participants assigned to the control group. The efficacy of the study treatment is then compared to the efficacy of the treatment which is already in common use.

Participants are not told their actual assignments for several reasons. If a participant is told that they are in the control or placebo group, they would be likely to seek treatment outside the trial for their condition, or to discontinue participation in the study completely, particularly if the study is examining the efficacy of a treatment for a life threatening condition. This would jeopardize the integrity of the trial by not giving an accurate representation as to the efficacy of the treatment. Conversely, if a participant is informed that they have been assigned to the treatment group, they might be inclined to report an exaggerated efficacy of the treatment for various reasons, such as to please the clinicians or to not be removed from the trial. This also jeopardizes the integrity of the trial by introducing a positive bias of the treatment's efficacy. Clinicians who are administering the treatments are not made aware of a participants assignment, when possible, in an effort to keep the participants from discovering which treatment group they were assigned to through the conduct of the clinician. Clinicians responsible for observing the participants in order to determine the treatment efficacy are not made aware of the assignment of individual participants in an effort to eliminate the possible biases a clinician might have, either positive or negative, toward the treatment being studied. In an effort to avoid these and other problems that might bring the validity of a study into question, certain protocols are put into place that define the methods of blinding that will be employed.

Chapter 2

Assessing the Success of Blinding

It is a widely held opinion today that in order to maintain the integrity of a randomized clinical trial, protocols must be in place to eliminate as many opportunities for the introduction of bias as is possible. It is also commonly agreed upon that appropriate blinding should be included in these protocols Freidman et al. (2010). It has been found, however, that only about 45% of studies describe the protocols regarding maintaining a similarity between the treatment and control regimens. Most studies made no reference to any attempt to assess whether the blinding was successful, or to what degree it was successful Bang et al. (2004). There are several methods for gathering the information necessary to assess the degree of success of blinding in a randomized clinical trial. One of these includes asking the participants and clinicians what they thought each participant's treatment allocation might be at or near the end of the treatment. Another method is to ask these same questions shortly after the beginning of treatment and then one or more times during treatment, with the last series of questions occurring near the end of treatment. When the participants are polled more than once, whether the participant changes their response (as well as when that happens) is sometimes taken into consideration. There is not only a great deal of variation in opinion regarding the appropriate time to poll the participants and clinicians, but also what questions should be asked and the allowable responses. Some studies ask the participants whether they believe they are in the treatment or control group, and a *don't know* response is not allowed. In other studies the surveys taken in an effort to asses the degree of success of the blinding allow *don't know* responses, but then disregard these responses in the analysis that follows. Sometimes the *don't know* responses are considered. Participants are sometimes asked to venture a guess if they first gave a *don't know* response, and this additional data may or may not be used in the analysis. Other studies use a five point scale, with responses of *strongly believe treatment*, *somewhat believe treatment*, *don't know*, *somewhat believe control*, or *strongly believe control* being allowed.

There is a complete lack of consensus from one study to the next as to what the appropriate protocols regarding blinding might be, as well as how to assess the success of these protocols (or if this should even be attempted) even though the goal is same: maintaining the integrity of the study. Having clear guidelines dealing with the issues concerning protocols for protecting the integrity of blinding as well as the polling of participants in an effort to assess the success of that blinding would make the assessment of the success of blinding much simpler, as well as giving the ability to confidently compare one study to another with regards to the success of blinding. There is no clear consensus however on which methods are best, or if there are benefits using one method over another depending on the structure of the trial, or if attempting to assess blinding is even a worthwhile endeavor. Some investigators believe that when a participant is polled on the treatment they believe they are assigned to, that this can possibly introduce bias as well, in which case attempting to assess the success of blinding could in and of itself jeopardize the integrity and success of the blinding.

All of these factors combine to make a uniform assessment of blinding highly unlikely in the foreseeable future. While this is true, there are still methods available for statistically measuring the success of blinding, three of which we discuss hereafter. These three indices are Cohen's Kappa statistic, James' Blinding Index developed by Kenneth James and his colleagues, and the NewBI proposed by Heejung Bang and his colleagues. In order to explore the differences in these indices, we will need to familiarize ourselves with the basic structure of data that has been collected for the purpose of assessing blinding. A typical data structure might look something like Table 2.1 or Table 2.2:

Table 2.1: 2 × 3 Data Structure

	Responses (guesses)			
	Treatment	Control	Don't Know	Total
Assigned Treatment	n_{11}	n_{12}	n_{13}	$n_{1.}$
Assigned Control	n_{21}	n_{22}	n_{23}	$n_{2.}$
Totals	$n_{.1}$	$n_{.2}$	$n_{.3}$	N

Table 2.1 shows the data structure that would result in a survey which asks for a participant or clinician to state simply whether they believe they have been assigned to the treatment group or the control group. The first column on the left indicates the actual assignment possibilities for the participants. The Treatment column under the Responses category indicates how many of the participants indicated their belief that they were assigned to the treatment group, while the Control column indicates a declared belief of being assigned to the control group, with those totals being represented by cells $n_{.1}$ and $n_{.2}$ respectively. If *don't know* responses are allowed, they would appear in the don't know column. For example, if a participant was assigned to the treatment group

but guessed that they had been assigned to the control group, their response would be recorded with all similar responses in cell n_{12}. On the other hand, if a participant was assigned to the control group but had no idea which group they had been assigned to, their response would be added to cell n_{23}. In the case where *don't know* responses are not allowed, cells n_{13}, n_{23} and $n_{.3}$ would contain zeros or the column would be omitted entirely.

Table 2.2 demonstrates a more complex data structure. The responses range from 1 to 5, with a possibility of response 1 indicating the participant strongly believes they were assigned to the treatment group, a response of 2 indicating the participant somewhat believed their assignment was to the treatment group, 3 represents a response indicating the participant somewhat believes they were assigned to the control group, and 4 indicates the participant strongly believes they were assigned to the control group. A response of 5 would coincide with a *don't know* response. The cells containing $n_{1.}$ and $n_{2.}$ represent the total number of participants assigned to the treatment and control groups respectively. This particular data structure represents one in which there is a single treatment and a single control assignment. It is possible that a study could have several different treatment assignments as well as a control assignment. In this case more rows would be added to account for the additional possible treatment assignments. If the relative frequency or proportion of each cell is desired, this is obtained by dividing each cell by the total number of participants. These tables represent just two variations of a limitless selection. The basic principles discussed however should readily apply with some minor modification to each of these variations.

Table 2.2: 2 × 5 Data Structure

	Responses (guesses)					
	1	2	3	4	5	Total
Assigned Treatment	n_{11}	n_{12}	n_{13}	n_{14}	n_{15}	$n_{1.}$
Assigned Control	n_{21}	n_{22}	n_{23}	n_{24}	n_{25}	$n_{2.}$
Totals	$n_{.1}$	$n_{.2}$	$n_{.3}$	$n_{.4}$	$n_{.5}$	N

2.1 Cohen's Kappa

Cohen's Kappa was first developed to check the agreement of two different people (or committees) who are rating the same group of items. One example of this would be if two different admissions officers for a university reviewed the same group of applications, with each officer giving a recommendation to either admit or not to admit each individual candidate. Once all of the applications were reviewed, then a two by two table could be constructed to show how many applications the two admissions officers agreed on (either both recommending admission or both recommending rejection of the application), and how many applications the two admissions officers disagreed upon, such as one recommending admission and one recommending rejection. Cohen's Kappa was applied to determining the success or failure of blinding in a randomized clinical trial by calculating the degree of agreement between the actual treatment assignment for a particular participant and the assignment that the participant or clinician supposes was made for that participant. That is, the degree of agreement between the actual and supposed assignments.

In Table 2.3, *a* would be the number of people in the treatment arm who correctly guessed that they were assigned to the treatment group, *b* would be the number of people

Table 2.3: Agreement of True and Supposed Treatment Assignment

Assignment	Treatment (guess)	Control (guess)	Total
Treatment (actual)	a	b	a+b
Control (actual)	c	d	c+d
Total	a+c	b+d	a+b+c+d

assigned to the treatment group who incorrectly guessed they had been assigned to the control group, *c* would be the number of participants assigned to the control group who incorrectly guessed they had been assigned to the treatment group, and *d* would be the number of participants assigned to the control group who correctly guessed that they had been assigned to the control group. This is similar to table 2.1 above, but without any allowance for a *don't know* response.

Cohen's Kappa is calculated using the following formula:

$$\kappa = \frac{P_o - P_e}{1 - P_e} \tag{2.1}$$

where P_o is the observed proportion of agreement between two raters, or in our case the percentage of participants or clinicians who guessed their assigned treatment group correctly, and P_e is the probability of a participant or clinician guessing the assignment purely by chance. We calculate P_o and P_e using the following equations:

$$P_o = \frac{a+d}{a+b+c+d} \tag{2.2}$$

$$P_e = \frac{(a+b)(a+c)+(c+d)(b+d)}{(a+b+c+d)^2} \tag{2.3}$$

The standard error for the Kappa statistic, denoted SE(K), can be calculated using the following equation:

$$SE(K) = \frac{SD(K)}{\sqrt{N}} \tag{2.4}$$

with

$$SD(K) = \sqrt{\frac{P_o(1-P_o)}{(1-P_e)^2}} \tag{2.5}$$

A $1-\alpha$ confidence interval can be constructed using the standard normal distribution as follows:

$$K \pm Z_{\alpha/2} SE(K) \tag{2.6}$$

Similarly, Cohen's Kappa statistic can be calculated to measure agreement in situations where there are more than two possible treatment assignments. Here the data would be arranged in a $k \times k$ table, where k is the number of treatment possibilities, or in a $(k+1) \times k$ table if a *don't know* response is allowed. The *don't know* responses will still be disregarded in the calculation. In such a case the Kappa statistic would be calculated using the following formula:

$$K_m = \frac{P_{mo} - P_{me}}{1 - P_{me}} \tag{2.7}$$

where:

$$P_{mo} = \sum_{1}^{k} P_{ii} \tag{2.8}$$

and

$$P_{me} = \sum_{1}^{k} P_{.i} \times P_{i.} \tag{2.9}$$

with $P_{ii} = \frac{n_{ii}}{L}$, $P_{.i} = \frac{n_{.i}}{L}$, and $P_{i.} = \frac{n_{i.}}{L}$ from table 2.1, with L being the number of responses excluding the *don't know* responses.

Cohen's Kappa that returns values of $K \leq 0$ indicates a complete lack of agreement beyond what would be expected simply by chance, and positive values up to and including 1 which would indicate increasing levels of agreement, a value of 1 indicating a perfect agreement. While it is generally agreed upon that a positive value for the Kappa statistic indicates some level of agreement beyond what would be expected by pure randomness, there is no scale to which we can refer as far as the level of significance for a particular non-negative Kappa value, because the marginal probability values are used to calculate P_e, and while the marginal probability of the actual assignment can and should be controlled, the marginal probability of the participant's guess depends upon many varied factors, including but not limited to the possible optimistic nature of the participants and a desire to be "special" or "chosen".

Cohen's Kappa statistic has some major drawbacks when it is used as a method for assessing unblinding in randomized clinical trials. While it is simple to use and relatively easy to calculate, it makes no allowance for the inclusion of a *don't know* response from the participant, because there would be no treatment arm that could be appropriately matched with a response of *don't know* and therefore it lacks the possibility of being included in the calculations. The *don't know* response would need to either be disregarded or disallowed. Disregarding the *don't know* responses would mean making calculations based upon only a subset of the responses, and the size of the subset is based solely upon the way the participants respond. Disallowing a *don't know* response would force participants to submit a guess as to which treatment arm they were assigned to. Forcing a participant to choose between the possible assignments might introduce a bias by not taking into consideration the degree to which the participant believes the response they give as to their assignment. One of the obvious choices in the spectrum of responses which range from a participant who firmly believes they were assigned to the *treatment* arm, to the participant who responds they firmly believe they were assigned to the *control* arm, would be the response directly between those two by a participant who is truly unsure which treatment arm they were assigned to, or *don't know*.

This fact was one of the driving factors in the development of the blinding index put forward by James James et al. (1996). When calculating Cohen's Kappa, any response of *don't know*, which James believed was the strongest indication of successful blinding, must be disregarded if it was not disallowed during the survey of the participants. Cohen's Kappa has another weakness when used to calculate agreement between more than two categories or treatment possibilities. When more than two possible categories are compared for agreement, the probability of chance agreement is reduced and the overall agreement indicated by the responses for any particular category become mathematically more significant, since a participant has more categories to choose from. Since this is the case, a particular kappa statistic calculated for a study that had more than two treatment assignments (such as varying levels of the medication, or a study that compares the efficacy of a treatment involving a new medication to a placebo as well as to the efficacy of an established medication), would likely indicate a much lower level of agreement between the actual treatment assignment and the supposed assignment than it would indicate if there was an equal level of agreement but only two options for the participant to choose from. This fact leads us to understand that the interpretation of Cohen's Kappa statistic depends also upon the format of the study that it is being used to assess, and is therefore somewhat subjective and not easily used to compare results from one study to the next unless the study parameters are precisely the same. The greatest advantage of Cohen's kappa is the relative ease with which calculations can be made to obtain the kappa statistic. The calculation of the confidence interval is also relatively simple. The following is an example of some of the values we might use to calculate the Cohen's Kappa statistic:

Table 2.4: An Example of Values for Calculating Cohen's Kappa

Assignment	Treatment (guess)	Control (guess)	Total
Treatment (actual)	42	8	50
Control (actual)	21	29	50
Total	63	37	100

Table 2.4 shows a two by two table that summarizes the agreement and disagreement of a hypothetical trial, which could then be used to calculate Cohen's Kappa. This table represents a trial involving one hundred participants in which fifty were assigned to the treatment group and fifty were assigned to the control (or placebo) group. The indication is that 42 participants that were assigned to the treatment group and 29 participants assigned to the control group guessed their assignment correctly, while 21 participants assigned to the control group guessed that their assignment was to the treatment group, and 8 participants who were assigned to the treatment group guessed that they were assigned to the control group.

Using the data from our example in table 2.4, we would calculate Cohen's Kappa by first calculating P_o and P_e:

$$\begin{aligned} P_o &= \frac{42+29}{42+8+21+29} \\ &= \frac{71}{100} \\ &= 0.71 \end{aligned} \tag{2.10}$$

$$
\begin{aligned}
P_e &= \frac{(42+8)(42+21)+(21+29)(8+29)}{(42+8+21+29)^2} \\
&= \frac{(50\cdot 63)+(50\cdot 37)}{100^2} \\
&= \frac{3150+1850}{10000} \\
&= \frac{5000}{10000} \\
&= 0.50 \qquad (2.11)
\end{aligned}
$$

Then Cohen's Kappa for our example is calculated as:

$$K = \frac{0.71-0.50}{1-0.50} = 0.42 \qquad (2.12)$$

There are some situations in which Cohen's Kappa gives slightly different kappa statistics for tables of data that show the same level of agreement. As an example consider Table 2.5 and Table 2.6:

Table 2.5: Example A

Assignment	Treatment (guess)	Control (guess)	Total
Treatment (actual)	45	15	60
Control (actual)	25	15	40
Total	70	30	100

We can see that the participants in each case correctly guessed their treatment assignment sixty percent of the time, which would lead us to believe that the kappa statistic for each should be the same. We can see however that this will not be the case once we begin to calculate the kappa statistic for each case. We begin by calculating P_o for each:

Table 2.6: Example B

Assignment	Treatment (guess)	Control (guess)	Total
Treatment (actual)	25	35	60
Control (actual)	5	35	40
Total	30	70	100

$$\begin{aligned} P_{oa} &= \frac{45+15}{45+15+25+15} \\ &= \frac{60}{100} \\ &= 0.60 \end{aligned} \tag{2.13}$$

and

$$\begin{aligned} P_{ob} &= \frac{25+35}{25+35+5+35} \\ &= \frac{60}{100} \\ &= 0.60 \end{aligned} \tag{2.14}$$

We see that the observed agreement, or P_o, is the same in both cases. Now we calculate the expected probabilities, or P_e for each example:

$$\begin{aligned}P_{ea} &= \frac{(45+15)(45+25)+(25+15)(15+15)}{(45+15+25+15)^2}\\ &= \frac{(60\cdot 70)+(40\cdot 30)}{100^2}\\ &= \frac{4200+1200}{10000}\\ &= \frac{5400}{10000}\\ &= 0.54\end{aligned} \tag{2.15}$$

and

$$\begin{aligned}P_{eb} &= \frac{(25+35)(25+5)+(5+35)(35+35)}{(25+35+5+35)^2}\\ &= \frac{(60\cdot 30)+(40\cdot 70)}{100^2}\\ &= \frac{1800+2800}{10000}\\ &= \frac{4600}{10000}\\ &= 0.46\end{aligned} \tag{2.16}$$

As we can see, the values for P_e or the expected probabilities for the two examples are not the same. Then Cohen's Kappa for our each example is calculated as:

$$\begin{aligned}K_a &= \frac{0.60-0.54}{1-0.54}\\ &= 0.1303\end{aligned} \tag{2.17}$$

and

$$\begin{aligned} K_b &= \frac{0.60 - 0.46}{1 - 0.46} \\ &= 0.2593 \end{aligned} \tag{2.18}$$

As we can see from Equation (2.17) and Equation (2.18), the resulting kappa statistics are different. This result is due to the fact that while the observed probabilities were the same, the expected probabilities calculated from the tables were not the same. This result demonstrates one of the weaknesses of Cohen's Kappa, namely that tables which demonstrate similar levels of agreement do not necessarily result in the same kappa statistic. This is problematic in that it requires a review of the tables that resulted in the Kappa statistic in order to solidify any conclusions drawn from it.

2.2 Further Development in Assessing Blindness: James et. al.

In 1979 the VA Cooperative Studies program supported a study that tested the efficacy of Disulfiram as a means to end the consumption of alcohol in patients that had been diagnosed with alcoholism or alcohol related maladies James et al. (1996). This study continued until 1983. It was designed in such a way that participants were randomly assigned, with equal probability, to one of three treatment groups. Two of the groups would receive Disulfiram as the treatment and the third would receive a placebo composed of riboflavin in order to be able to check that they were continuing to comply with the study. The two groups that were assigned to receive the Disulfiram were to be given either 250mg or 1mg doses. In this study, the patients were told whether they would be taking the Disulfiram or the placebo, but not which dosage of Disulfiram they had been assigned. The treatment medications were all prepared to be identical, and the patients were instructed not to discuss which treatment group they had been assigned to. At the end of the study the clinicians and program coordinators at each location where the study was being conducted were asked to identify which patients were assigned to what treatment group. One of the other unique aspects of this study was that the clinicians were allowed to state that they didn't know which treatment group a particular patient was assigned to, and these *don't know* responses were considered in the calculation of the degree of success of the blinding. In 1995, Kenneth James and his colleagues submitted a paper detailing this study, known as VA Cooperative Study Number 107 James et al. (1996). The point of interest to us, however, was not whether the treatment was effective. Rather the design of the experiment and the insight with regards to the success of blinding was the point of interest. This included a statistical method that could be used in order to try to measure the degree to which the blinding was successful. This article proposed a *Blinding Index* (BI), which is commonly referred to as the *James Blinding Index* or *James BI*. Where Cohen's Kappa simply gives a rate of agreement with the *don't know* responses not being allowed or being disregarded, James and his colleagues proposed that disagreement, that is, an incorrect guess about treatment allocation, was a more accurate indication of successful blinding, with a *don't know* response being the strongest indication that blinding was successful.

The Blinding Index proposed by James is calculated using the following formula:

$$BI = \frac{[1 + P_{DK} + (1 - P_{DK})K_D]}{2} \tag{2.19}$$

Where K_D is the kappa statistic calculated in a slightly modified fashion to account for the possibility of a *don't know* response, as follows:

$$K_D = \frac{P_{Do} - P_{De}}{P_{De}} \tag{2.20}$$

Now, however, the Kappa statistic is a measure of relative disagreement rather than am measure of agreement. P_{Do} and P_{De} are given weights to adjust the index according to the desirability of each outcome, and are calculated as follows:

$$P_{Do} = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{W_{ij} P_{ij}}{1 - P_{DK}} \tag{2.21}$$

and

$$P_{De} = \sum_{i=1}^{k} \sum_{j=1}^{k} \frac{W_{ij} P_{i.} (P_{.j} - P_{0j})}{(1 - P_{DK})^2} \tag{2.22}$$

where the W_{ij} are the weights for the particular responses and the $P's$ are the expected relative frequencies of the entries in the $(k+1) \times k$ table. The variance of the blinding index is calculated using the following formula:

$$\begin{aligned} Var(\hat{B}I) = \frac{1}{N} \times \Bigg\{ & \left[2 \sum_{i=1}^{k} \sum_{j=1}^{k} p_{i.} (p_{.j} - p_{oj}) w_{ij} \right]^{-2} \\ & \times \left[\sum_{i=1}^{k} \sum_{j=1}^{k} p_{ij} (1 - p_{DK})^2 \left[(1 - p_{DK}) w_{ij} - (1 + \kappa_D) \sum_{r=1}^{k} \{ p_{r.} w_{rj} + (p_{.r} - p_{or}) w_{ir} \} \right]^2 \right] \\ & + p_{DK}(1 - p_{DK}) - (1 - p_{DK})(1 + k_D) \left[p_{DK} + \frac{(1 - p_{DK})(1 + \kappa_D)}{4} \right] \Bigg] \Bigg\} \end{aligned} \tag{2.23}$$

This blinding index was initially developed and used in conjunction with the VA Cooperative Study No. 107, in an effort to improve the statistical methods used to assess the success of blinding in randomized clinical trials. That being the case, the data from that study along with the calculations resulting from the implementation of this blinding index are included in Tables 2.7 and 2.8, with the data in each cell of the table containing the number of responses for the cell, the relative frequency of this response, and the weight assigned to the category corresponding to that particular cell. These weights represent the significance of a particular response as it pertains to the success of blinding. A correct guess, which would be the strongest indication of unblinding, was given a weight of zero. A response that correctly identified that a participant was taking disulfiram but incorrectly identified the correct dosage was seen to be somewhat indicative of successful blinding and was given a weight of 0.5. A response that incorrectly identified whether the participant was taking disulfiram or riboflavin was given a weight of 0.75 since it was a very good indication of successful blinding. A response of *don't know* was seen as the surest sign that blinding was maintained, and was implicitly given a weight of one.

As can be seen in Tables 2.7 and 2.8, the practitioners, specifically the study coordinators and the program therapists, were the subject of the blinding assessment, with the patients having been informed whether they were assigned to the Disulfiram or the Riboflavin, but not told which dosage of Disulfiram they were being given. Note that Table 2.7 is a 4×3 table while Table 2.8 is a 3×2 table. This is because the program therapists were aware that the participants were either receiving disulfiram or riboflavin, but were unaware that there were two dosages of the disulfiram. We also see that the data was presented in an inverted fashion with the rows representing the possible responses, the last row representing the *don't know* responses, the preceding rows representing the three possible treatment assignments, and the columns representing the actual treatment assignments, resulting in a 4×3 data table. This would be compared to the usual method of making the columns represent the supposed or guessed assignments, with the last column representing the *don't know* responses, and the rows representing the actual treatment assignments. In either case, one row and one column would be added for the marginal totals. The data here is represented in the format presented in the original paper detailing the development of the index James et al. (1996).

Table 2.7: Study Coordinator’s Responses

	Actual Assignment			
Guessed Assignment	Disulfiram 1mg.	Disulfiram 250mg.	Riboflavin	Total
Disulfiram 1mg.	41	27	22	90
Proportion	0.08	0.05	0.04	0.17
Weight	0.0	0.5	0.75	
Disulfiram 250mg.	66	72	36	174
Proportion	0.13	0.13	0.07	0.33
Weight	0.5	0.0	0.75	
Riboflavin	30	24	64	118
Proportion	0.05	0.05	0.12	0.22
Weight	0.75	0.75	0.0	
Don’t Know	44	51	52	147
Proportion	0.08	0.1	0.1	0.28
Weight	1.0	1.0	1.0	
Totals	181	174	174	529

Table 2.8: Program Therapist's Responses

	Actual Assignment		
Guessed Assignment	Disulfiram	Riboflavin	Total
Disulfiram	145	34	179
Proportion	0.34	0.08	0.42
Weight	0.0	0.75	
Riboflavin	71	59	130
Proportion	0.17	0.14	0.31
Weight	0.75	0.0	
Don't Know	76	38	114
Proportion	0.18	0.09	0.27
Weight	1.0	1.0	
Totals	292	131	423

We can use Tables 2.7 and 2.8 to calculate the James blinding index. We will begin with the responses given by the Study Coordinators. First we can calculate the κ_D portion of equation 2.19 by first calculating P_{Do} and P_{De}:

$$\begin{aligned}
P_{Do} &= \frac{0.5 \cdot 0.05 + 0.75 \cdot 0.04 + 0.5 \cdot 0.13 +}{(1-0.28)^2} \\
&+ \frac{0.75 \cdot 0.07 + 0.75 \cdot 0.05 + 0.75 \cdot 0.05}{1-0.28} \\
&= \frac{0.2475}{.72} \\
&= 0.34375 \qquad (2.24)
\end{aligned}$$

and

$$\begin{aligned}
P_{De} &= \frac{0.5 \cdot 0.17 \cdot 0.23 + 0.75 \cdot 0.17 \cdot 0.23 + 0.5 \cdot 0.33 \cdot 0.26}{(1-0.28)^2} \\
&+ \frac{0.75 \cdot 0.33 \cdot 0.23 + 0.75 \cdot 0.22 \cdot 0.26 + 0.75 \cdot 0.22 \cdot 0.23}{(1-0.28)^2} \\
&= \frac{0.22955}{0.5184} \\
&= 0.4427 \qquad (2.25)
\end{aligned}$$

Then

$$\begin{aligned}
\kappa_D &= \frac{0.34375 - 0.4427}{0.4427} \\
&= -0.22355 \qquad (2.26)
\end{aligned}$$

And

$$\begin{aligned}BI &= \frac{1+0.28+(1-0.28)(-0.2355)}{2}\\ &= 0.55972\end{aligned} \tag{2.27}$$

Similarly, we can calculate the blinding index for the Program Therapists:

$$\begin{aligned}P_{Do} &= \frac{0.75\cdot 0.08+0.75\cdot 0.17}{0.73}\\ &= \frac{0.1875}{0.73}\\ &= 0.2568\end{aligned} \tag{2.28}$$

And

$$\begin{aligned}P_{De} &= \frac{0.75\cdot 0.42\cdot 0.22+0.75\cdot 0.41\cdot 0.31}{0.73^2}\\ &= \frac{0.1646}{0.5329}\\ &= 0.3089\end{aligned} \tag{2.29}$$

Then

$$\begin{aligned}\kappa_D &= \frac{0.2568-0.3526}{0.3526}\\ &= -0.2717\end{aligned} \tag{2.30}$$

So

$$\begin{aligned}BI &= \frac{1+.27+.73\cdot -0.2717}{2}\\ &= 0.5358\end{aligned} \tag{2.31}$$

The various detractors of the James BI, one of which gave rise to the NewBI which will be discussed in the next section, claimed that giving *don't know* responses such a high weight leads to misleading indices. Most of these agree that while a *don't know* response would ideally indicate true blinding, that there were other reasons a participant might be inclined to give such a response. Some participants might give a *don't know* response for fear of removal from the study if they guessed correctly. Some may give a response of *don't know* in order to not get the practitioners in trouble. If a participant has a good idea which was their assignment but isn't 100% certain they might be inclined to respond *don't know*. While one could suppose that an inflated number of *don't know* guesses might have little effect in the overall scheme of things, if even a small percentage of the participants responded *don't know* when they actually had an opinion as to their assignment, this could have a dramatic effect on the calculation of the index as will be demonstrated by our simulation study.

2.3 Bang et.al.'s New Blinding Index

In April of 2004, Heejung Bang and his colleagues published a paper detailing a new method for assessing the success of blinding Bang et al. (2004). This article reviewed the work done by James and proposed a new method for calculating and assessing the possible unblinding in randomized clinical trials. This new blinding assessment, called NewBI, focused on improving the work of James by reducing the amount by which the result is critically dominated by *don't know* responses and eliminating the ambiguity of interpreting the results of the index. Another advantage to this NewBI would be the ability to assess the possibility of unblinding in the treatment arm and the control arm independently. This was accomplished through the employment of a multinomial distribution approach.

The newBI proposed by Bang is calculated as follows:

$$\widehat{newBI}_i = (2\widehat{r_{i|i}} - 1) \times \left(\frac{n_{i1} + n_{i2}}{n_{i1} + n_{i2} + n_{i3}} \right) \tag{2.32}$$

where

$$\widehat{r_{i|i}} = \frac{n_{ii}}{n_{i1} + n_{i2}} \tag{2.33}$$

and $n_{i1} + n_{i2} \neq 0$. This estimates the proportion of individuals who guess their treatment assignment correctly in the i^{th} treatment arm. More specifically, for the participants assigned to actually receive the treatment, blinding would be assessed by calculating the following:

$$\widehat{newBI}_1 = (2\widehat{r_{1|1}} - 1) \times \left(\frac{n_{11} + n_{12}}{n_{11} + n_{12} + n_{13}} \right) \tag{2.34}$$

which simplifies to:

$$\widehat{newBI}_1 = \frac{n_{11} - n_{12}}{n_{1.}} \tag{2.35}$$

Similarly, when calculating the newBI for the participants who were assigned to the control group, the equation simplifies to the following:

$$\widehat{newBI}_2 = \frac{n_{22} - n_{21}}{n_{2.}} \tag{2.36}$$

This new blinding index will take on values from -1 to 1, with a result of 0 indicating that the responses were not significantly different than a result obtained from random guessing.

The variance for the newBI is calculated using the Formula 2.37:

$$Var(newBI_i) = [P_{1|i}(1-P_{1|i}) + P_{2|i}(1-P_{2|i}) + 2(P_{1|i})(P_{2|i})] \tag{2.37}$$

where $P_{j|i}$ is the conditional probability, which can be estimated by:

$$\widehat{P_{j|i}} = \frac{n_{ij}}{n_{i.}} \tag{2.38}$$

for $(i, j = 1, 2)$.

When we make the appropriate substitutions and simplify, we see that the variance for *newBI*$_1$, which is the variance of the index dealing with the participants assigned to the treatment group, is as follows:

$$Var(newBI_1) = \frac{n_{11}}{n_{1.}} \quad 1 - \frac{n_{11}}{n_{1.}} \quad + \frac{n_{12}}{n_{1.}} \quad 1 - \frac{n_{12}}{n_{1.}} \quad + 2 \quad \frac{n_{11}}{n_{1.}} \qquad \frac{n_{12}}{n_{1.}} \tag{2.39}$$

Similarly, when we calculate the variance for *newBI*$_2$ which represents the index calculated from the responses obtained from the participants who were assigned to the control group, we use the following formula:

$$Var(newBI_2) = \frac{n_{21}}{n_{2.}} \quad 1 - \frac{n_{21}}{n_{2.}} \quad + \frac{n_{22}}{n_{2.}} \quad 1 - \frac{n_{22}}{n_{2.}} \quad + 2 \quad \frac{n_{21}}{n_{2.}} \qquad \frac{n_{22}}{n_{2.}} \tag{2.40}$$

Consider the following example: A study with one thousand participants that has two treatment arms which we will generically call *Treatment* and *Control*, with a two-way classification table as shown in Table 2.9. The newBI would be calculated for the *Treatment* arm as shown in Equation 2.41:

$$\begin{aligned} newBI_t &= \frac{212 - 126}{500} \\ &= 0.172 \end{aligned} \tag{2.41}$$

Table 2.9: NewBI Example

Assignment	Treatment (guess)	Control (guess)	Don't Know	Total
Treatment (actual)	212	126	162	500
Control (actual)	193	159	148	500
Total	405	285	310	1000

And newBI calculated for the *Control* arm would be:

$$\begin{aligned} newBI_c &= \frac{159 - 193}{500} \\ &= -0.068 \end{aligned} \tag{2.42}$$

Calculating the variance for table 2.9 would follow similarly. The variance for the *Treatment* arm would be:

$$\begin{aligned} Var(newBI_t) &= \frac{\frac{212}{500} \cdot (1 - \frac{212}{500}) + \frac{126}{500} \cdot (1 - \frac{126}{500}) + 2 \cdot \frac{212}{500} \cdot \cdot \frac{126}{500}}{500} \\ &= \frac{\frac{212}{500} \cdot \frac{288}{500} + \frac{126}{500} \cdot \frac{374}{500} + 2 \cdot \frac{212}{500} \cdot \frac{126}{500}}{500} \\ &= 0.00129 \end{aligned} \tag{2.43}$$

The variance for the *Control* arm would be:

$$\begin{aligned} Var(newBI_c) &= \frac{\frac{193}{500} \cdot (1 - \frac{193}{500}) + \frac{159}{500} \cdot (1 - \frac{159}{500}) + 2 \cdot \frac{193}{500} \cdot \frac{159}{500}}{500} \\ &= \frac{\frac{193}{500} \cdot \frac{307}{500} + \frac{159}{500} \cdot \frac{341}{500} + 2 \cdot \frac{193}{500} \cdot \frac{159}{500}}{500} \\ &= 0.001399 \end{aligned} \tag{2.44}$$

We can construct a 95% confidence interval for treatment and control arms, which would be as follows:

$$
\begin{aligned}
\text{95\% CI for } newBI_t &= 0.172 \pm 1.96 \cdot \sqrt{0.00129} \\
&= (0.1016, 0.2424) \qquad (2.45)
\end{aligned}
$$

And

$$
\begin{aligned}
\text{95\% CI for } newBI_c &= -0.68 \pm 1.96 \cdot \sqrt{0.001399} \\
&= (-0.7533, -0.6067) \qquad (2.46)
\end{aligned}
$$

We can see from this that the treatment arm shows a failure to maintain blinding since the lower limit is above zero, or that a greater number of participants responded correctly than would be accounted for by chance. We also see that the lower limit for the control arm is below zero, but because the upper limit is also below zero this can indicate that a greater number of participants than would be accounted for by random chance responded incorrectly. Since this indicates a greater number of participants believed that they had been given the treatment, the logical follow-up would be to test for response bias.

Chapter 3

A Simulation Study Comparing Indices

In order to more thoroughly examine the three previously mentioned methods for assessing unblinding, the following R algorithm was programmed. The study simulated 200 study participants. Their responses were simulated in R by drawing them randomly from a uniform distribution, with responses allocated as *responded treatment*, *responded control* or *responded don't know* based upon the random value. The range of responses allocated *don't know* was incremented from 0.0 to 0.95 in increments of 0.05. the remainder of the responses were allocated to *responded treatment* or *responded control*, depending on the actual assignment and whether they guessed the correct or incorrect treatment arm. The range of values assigned to the correct response was calculated in R by use of the following formula:

$$PGC = (1 - PDK) \cdot X \tag{3.1}$$

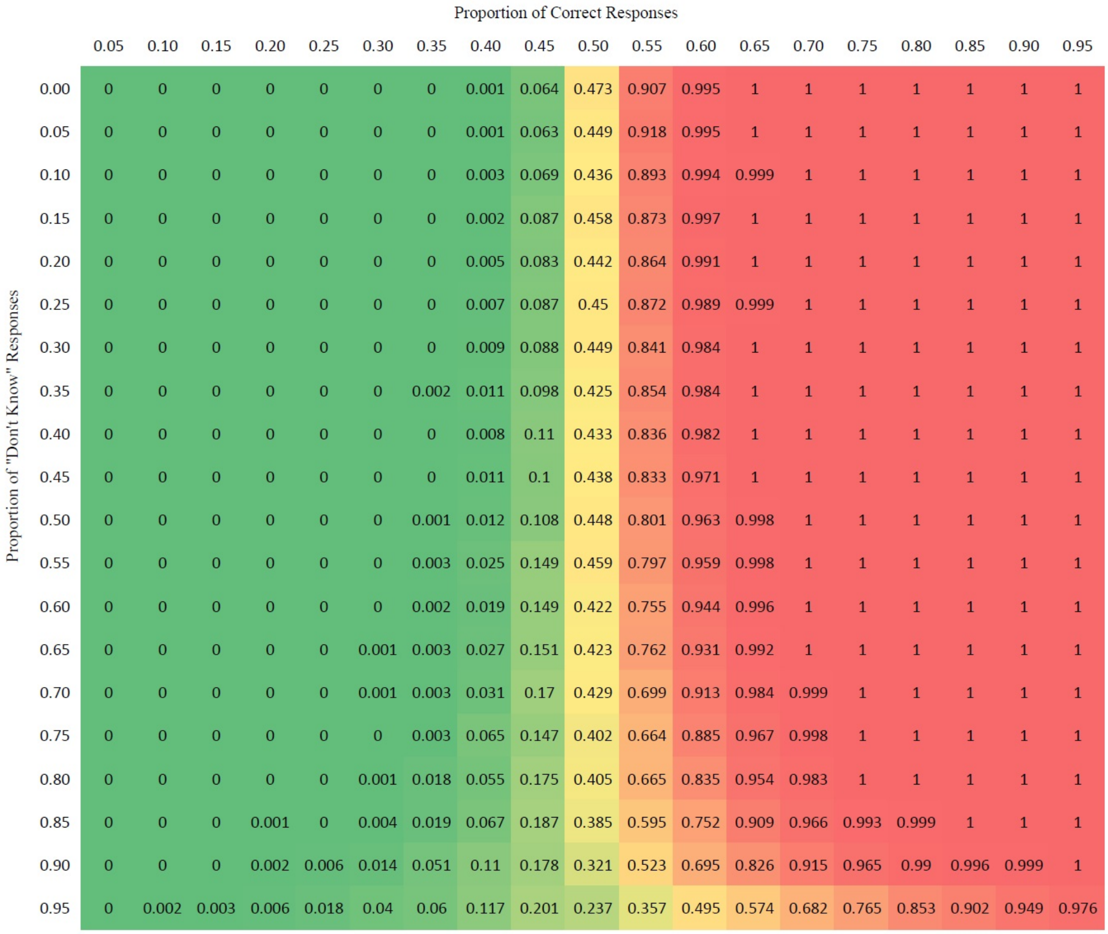
Where PGC is the proportion responding correctly, PDK is the proportion responding don't know and X being incremented from 0.0 to 0.95 in increments of .05. The remaining proportion was allocated to an incorrect response. In every case the following was true, with PGI indicating an incorrect response:

$$PGI = 1 - (PGC + PDK) \tag{3.2}$$

The actual assignment was made by assigning the first hundred participants to the *treatment* group and the second hundred participants to the *control* group. Once the assignments were made, the responses were calculated using a random number generated by R and comparing this random number to the proportions for correct and incorrect as well as don't know responses. The success of blinding was assessed using the three previously discussed indices under the null hypothesis that blinding was maintained, at a level 0f $\alpha = 0.05$. Each index for which the null hypothesis was rejected was a value of one for that trial, while those for which the null hypothesis could not be rejected were given a value of zero. This process of obtaining two hundred random values, determining the response for each participant based upon those random values, and assessing the success or failure of blinding based upon those responses was repeated one thousand times for each of the three hundred ninety combinations possible with regards to the proportion of don't know, correct and incorrect responses. The relative frequency of the failure of blinding was calculated for each of the three hundred ninety combinations, and heat maps were created as a way to visualize the similarities and differences among these indices.

3.1 Cohen's Kappa

The first index we will examine is Cohen's Kappa statistic. This index performed rather well considering the limitations it faced, not the least of which was that any responses other that treatment or control were disregarded.



Proportion of Correct Responses

Proportion of "Don't Know" Responses	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.00	0	0	0	0	0	0	0	0.001	0.064	0.473	0.907	0.995	1	1	1	1	1	1	1
0.05	0	0	0	0	0	0	0	0.001	0.063	0.449	0.918	0.995	1	1	1	1	1	1	1
0.10	0	0	0	0	0	0	0	0.003	0.069	0.436	0.893	0.994	0.999	1	1	1	1	1	1
0.15	0	0	0	0	0	0	0	0.002	0.087	0.458	0.873	0.997	1	1	1	1	1	1	1
0.20	0	0	0	0	0	0	0	0.005	0.083	0.442	0.864	0.991	1	1	1	1	1	1	1
0.25	0	0	0	0	0	0	0	0.007	0.087	0.45	0.872	0.989	0.999	1	1	1	1	1	1
0.30	0	0	0	0	0	0	0	0.009	0.088	0.449	0.841	0.984	1	1	1	1	1	1	1
0.35	0	0	0	0	0	0	0.002	0.011	0.098	0.425	0.854	0.984	1	1	1	1	1	1	1
0.40	0	0	0	0	0	0	0	0.008	0.11	0.433	0.836	0.982	1	1	1	1	1	1	1
0.45	0	0	0	0	0	0	0	0.011	0.1	0.438	0.833	0.971	1	1	1	1	1	1	1
0.50	0	0	0	0	0	0	0.001	0.012	0.108	0.448	0.801	0.963	0.998	1	1	1	1	1	1
0.55	0	0	0	0	0	0	0.003	0.025	0.149	0.459	0.797	0.959	0.998	1	1	1	1	1	1
0.60	0	0	0	0	0	0	0.002	0.019	0.149	0.422	0.755	0.944	0.996	1	1	1	1	1	1
0.65	0	0	0	0	0	0.001	0.003	0.027	0.151	0.423	0.762	0.931	0.992	1	1	1	1	1	1
0.70	0	0	0	0	0	0.001	0.003	0.031	0.17	0.429	0.699	0.913	0.984	0.999	1	1	1	1	1
0.75	0	0	0	0	0	0	0.003	0.065	0.147	0.402	0.664	0.885	0.967	0.998	1	1	1	1	1
0.80	0	0	0	0	0	0.001	0.018	0.055	0.175	0.405	0.665	0.835	0.954	0.983	1	1	1	1	1
0.85	0	0	0	0.001	0	0.004	0.019	0.067	0.187	0.385	0.595	0.752	0.909	0.966	0.993	0.999	1	1	1
0.90	0	0	0	0.002	0.006	0.014	0.051	0.11	0.178	0.321	0.523	0.695	0.826	0.915	0.965	0.99	0.996	0.999	1
0.95	0	0.002	0.003	0.006	0.018	0.04	0.06	0.117	0.201	0.237	0.357	0.495	0.574	0.682	0.765	0.853	0.902	0.949	0.976

Figure 3.1: Heat Map for Cohen’s Kappa

We see from Figure 3.1 that there begins to be rejection of H_0 using the kappa statistic at 40% correct responses with no don't know responses. By 65% correct responses and above H_0 was rejected in all one thousand iterations. As the proportion of don't know responses increases, the percentages from 45% to 55% remain the transition area from what we could call the *Likely not to reject* H_0 *area* to the *Likely to reject* H_0 *area*. When the proportions of don't know responses exceeds 65% we see a slight widening of that range as well as a gradual major widening of the range in which rejections of H_0 begin to occur. This can be explained by the increasing proportion of don't know responses, which are disregarded when calculating Cohen's Kappa. With fewer responses being used in the actual calculation it is easier for an aberrant set of observations to greatly influence the disposition of the calculated index. This result is far from troubling since we knew from the onset that this is one of the weaknesses with the Kappa statistic. We can also conclude that if 95% of participants responded don't know in an actual study, this fact alone would bear further scrutiny.

Overall Cohen's Kappa performed rather well, despite its obvious limitations when employed for this purpose. This fact is noteworthy, since Cohen proposed this statistic over fifty years ago.

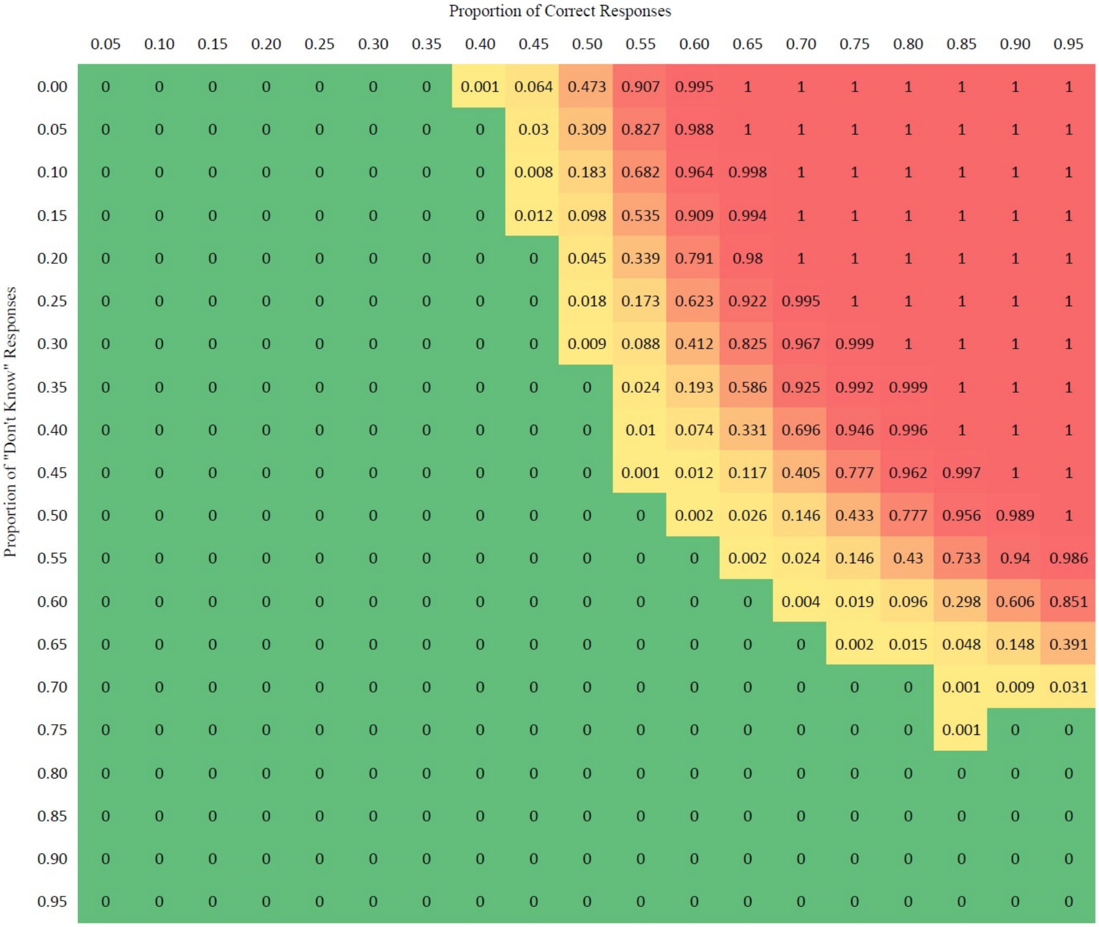
3.2 James' Blinding Index

Figures 3.2 through 3.6 are created using the blinding index proposed by James. There are five versions of this index, calculated with five different sets of parameters for the weight of incorrect and don't know responses. Table 3.1 details the weights employed for this exercise, with the weights employed to produce Figure 3.6 being the same weights suggested by James in his original study:

Table 3.1: Weights for Responses using James' BI

Figure	Correct	Incorrect	Don't Know
3.2	0.0	0.2	0.5
3.3	0.0	0.3	0.6
3.4	0.0	0.4	0.7
3.5	0.0	0.5	0.8
3.6	0.0	0.5	1.0

These heat maps demonstrate that the assertions made by Bang certainly seem to have some merit. That is, it is impossible to obtain a result which will cause the rejection of H_0 at higher proportions of don't know responses. The rejection of H_0 occurred exactly one time with a level of don't know responses at 75%, with none occurring when the proportion of don't know responses exceeded that mark. Changes to the weights made some small difference in the shape and position of the results but not markedly so. In over 20% of the trials there was not a single rejection of H_0, even though at the bottom right corners of the figures we see that the proportion of correct responses was 95%. This is problematic because any time the proportion of correct guesses is above even 80% for those participants who offered a response other than don't know, there should be some question as to the success of blinding, and a proportion of 95% is a rather clear indication that the participants overwhelmingly knew what their assignment was. In defense of James and his colleagues, it could be said that it was only 95% of 30% of the data, but in the overall scheme of things about 28.5% responded correctly, approximately 1.5% responded incorrectly, and about 70% responded don't know. This might lead one to believe that there is a significant question as to whether blinding had failed. However,

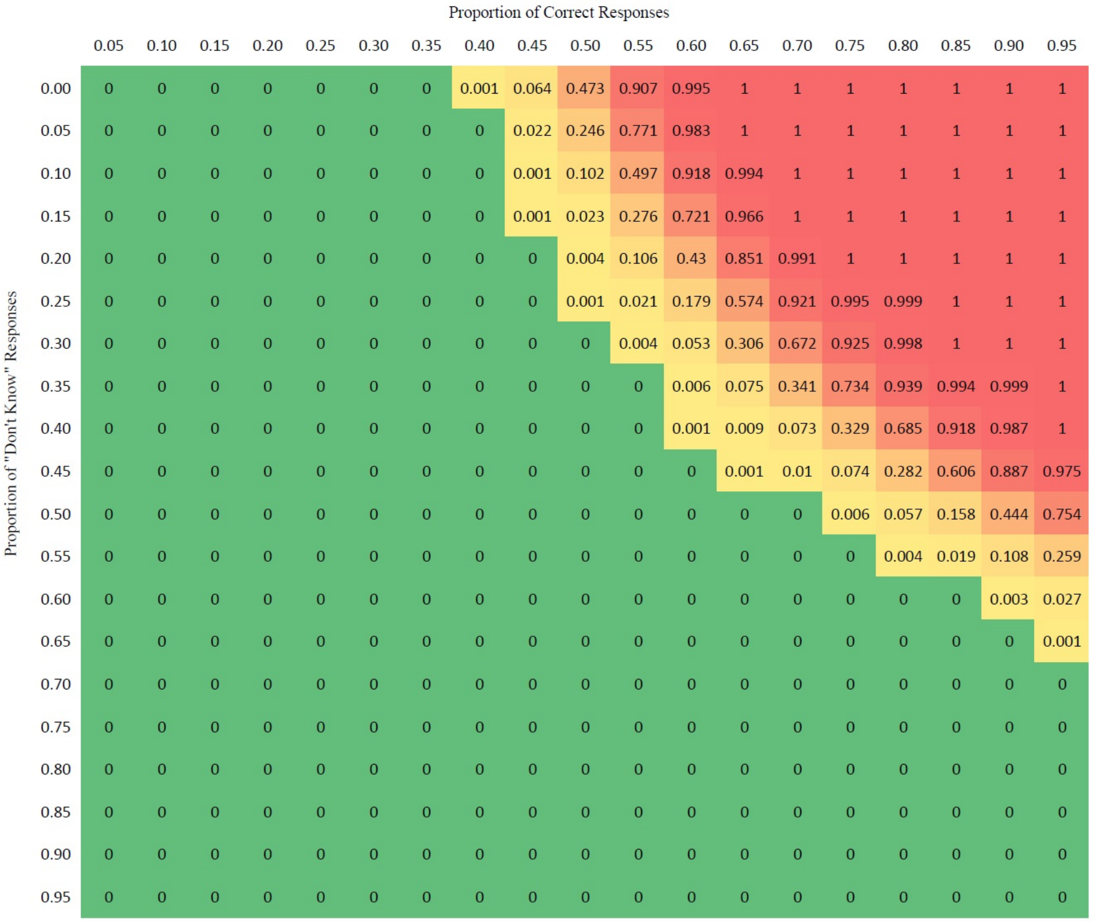


Proportion of Correct Responses

Proportion of "Don't Know" Responses	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.00	0	0	0	0	0	0	0	0.001	0.064	0.473	0.907	0.995	1	1	1	1	1	1	1
0.05	0	0	0	0	0	0	0	0	0.03	0.309	0.827	0.988	1	1	1	1	1	1	1
0.10	0	0	0	0	0	0	0	0	0.008	0.183	0.682	0.964	0.998	1	1	1	1	1	1
0.15	0	0	0	0	0	0	0	0	0.012	0.098	0.535	0.909	0.994	1	1	1	1	1	1
0.20	0	0	0	0	0	0	0	0	0	0.045	0.339	0.791	0.98	1	1	1	1	1	1
0.25	0	0	0	0	0	0	0	0	0	0.018	0.173	0.623	0.922	0.995	1	1	1	1	1
0.30	0	0	0	0	0	0	0	0	0	0.009	0.088	0.412	0.825	0.967	0.999	1	1	1	1
0.35	0	0	0	0	0	0	0	0	0	0	0.024	0.193	0.586	0.925	0.992	0.999	1	1	1
0.40	0	0	0	0	0	0	0	0	0	0	0.01	0.074	0.331	0.696	0.946	0.996	1	1	1
0.45	0	0	0	0	0	0	0	0	0	0	0.001	0.012	0.117	0.405	0.777	0.962	0.997	1	1
0.50	0	0	0	0	0	0	0	0	0	0	0	0.002	0.026	0.146	0.433	0.777	0.956	0.989	1
0.55	0	0	0	0	0	0	0	0	0	0	0	0	0.002	0.024	0.146	0.43	0.733	0.94	0.986
0.60	0	0	0	0	0	0	0	0	0	0	0	0	0	0.004	0.019	0.096	0.298	0.606	0.851
0.65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.002	0.015	0.048	0.148	0.391
0.70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.001	0.009	0.031
0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.001	0	0
0.80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

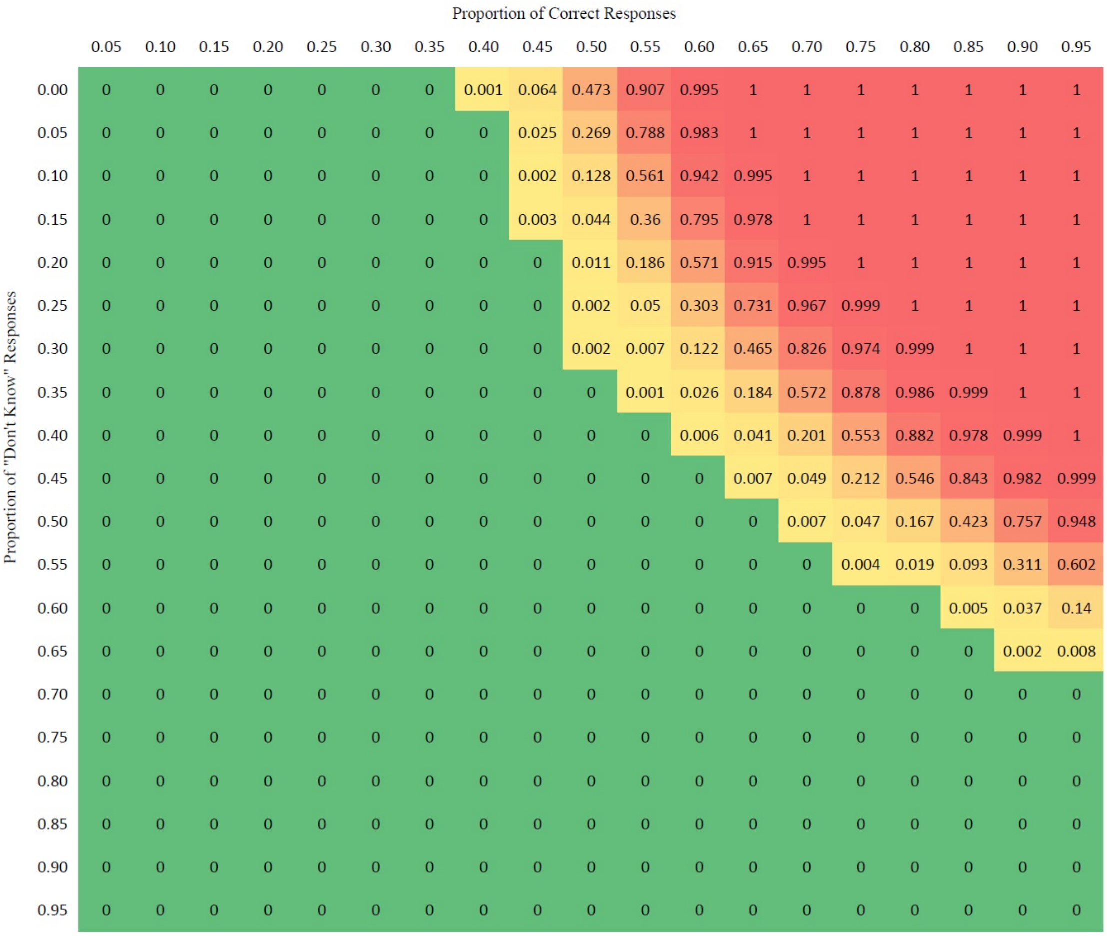
Figure 3.2: Heat Map for James' BI with Weights 0/0.2/0.5

the index makes no An index to assess the success of blinding is only useful if it actually indicates that unblinding has occurred when unblinding has obviously occurred. At the very least we must consider that the blinding index proposed by James is very conservative in rejecting H_0.



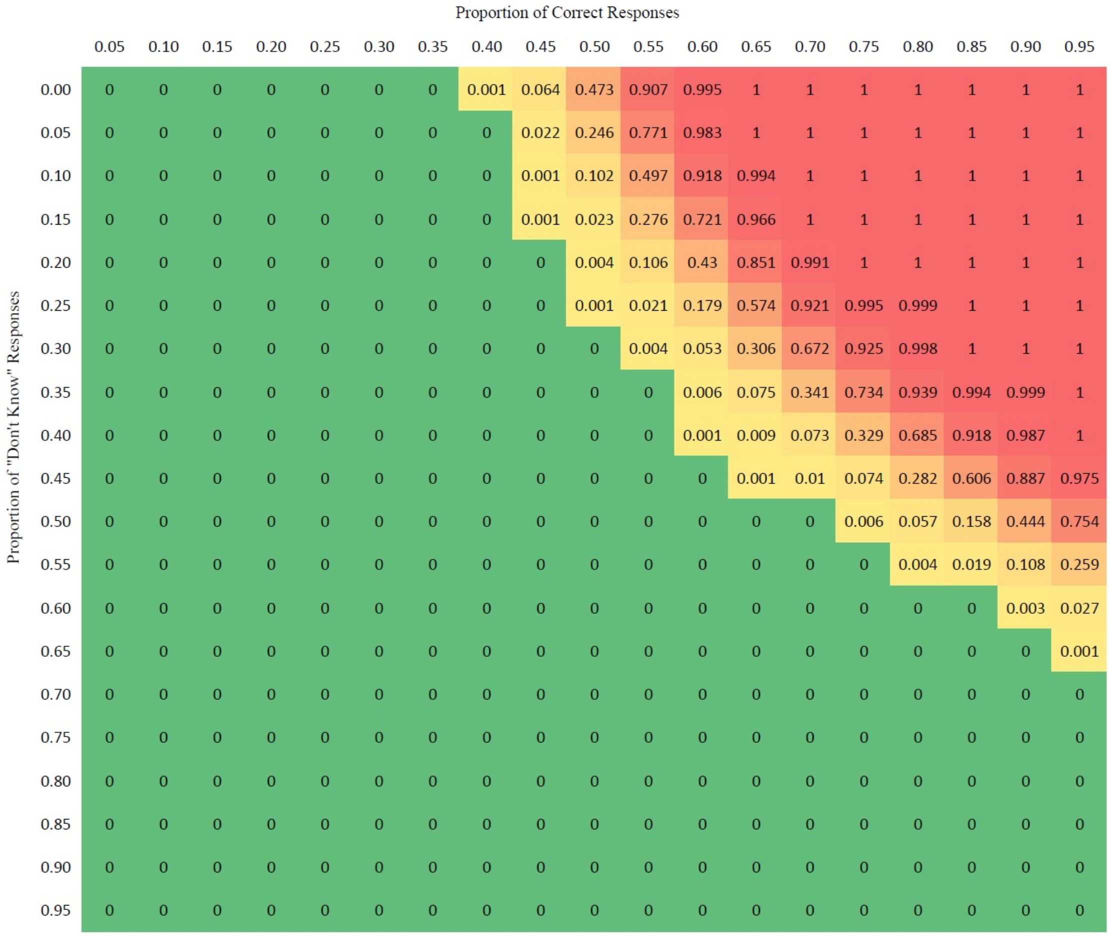
	Proportion of Correct Responses																		
Proportion of "Don't Know" Responses	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.00	0	0	0	0	0	0	0	0.001	0.064	0.473	0.907	0.995	1	1	1	1	1	1	1
0.05	0	0	0	0	0	0	0	0	0.022	0.246	0.771	0.983	1	1	1	1	1	1	1
0.10	0	0	0	0	0	0	0	0	0.001	0.102	0.497	0.918	0.994	1	1	1	1	1	1
0.15	0	0	0	0	0	0	0	0	0.001	0.023	0.276	0.721	0.966	1	1	1	1	1	1
0.20	0	0	0	0	0	0	0	0	0	0.004	0.106	0.43	0.851	0.991	1	1	1	1	1
0.25	0	0	0	0	0	0	0	0	0	0.001	0.021	0.179	0.574	0.921	0.995	0.999	1	1	1
0.30	0	0	0	0	0	0	0	0	0	0	0.004	0.053	0.306	0.672	0.925	0.998	1	1	1
0.35	0	0	0	0	0	0	0	0	0	0	0	0.006	0.075	0.341	0.734	0.939	0.994	0.999	1
0.40	0	0	0	0	0	0	0	0	0	0	0	0.001	0.009	0.073	0.329	0.685	0.918	0.987	1
0.45	0	0	0	0	0	0	0	0	0	0	0	0	0.001	0.01	0.074	0.282	0.606	0.887	0.975
0.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.006	0.057	0.158	0.444	0.754
0.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.004	0.019	0.108	0.259
0.60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.003	0.027
0.65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.001
0.70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.3: Heat Map for James' BI with Weights 0/0.3/0.6



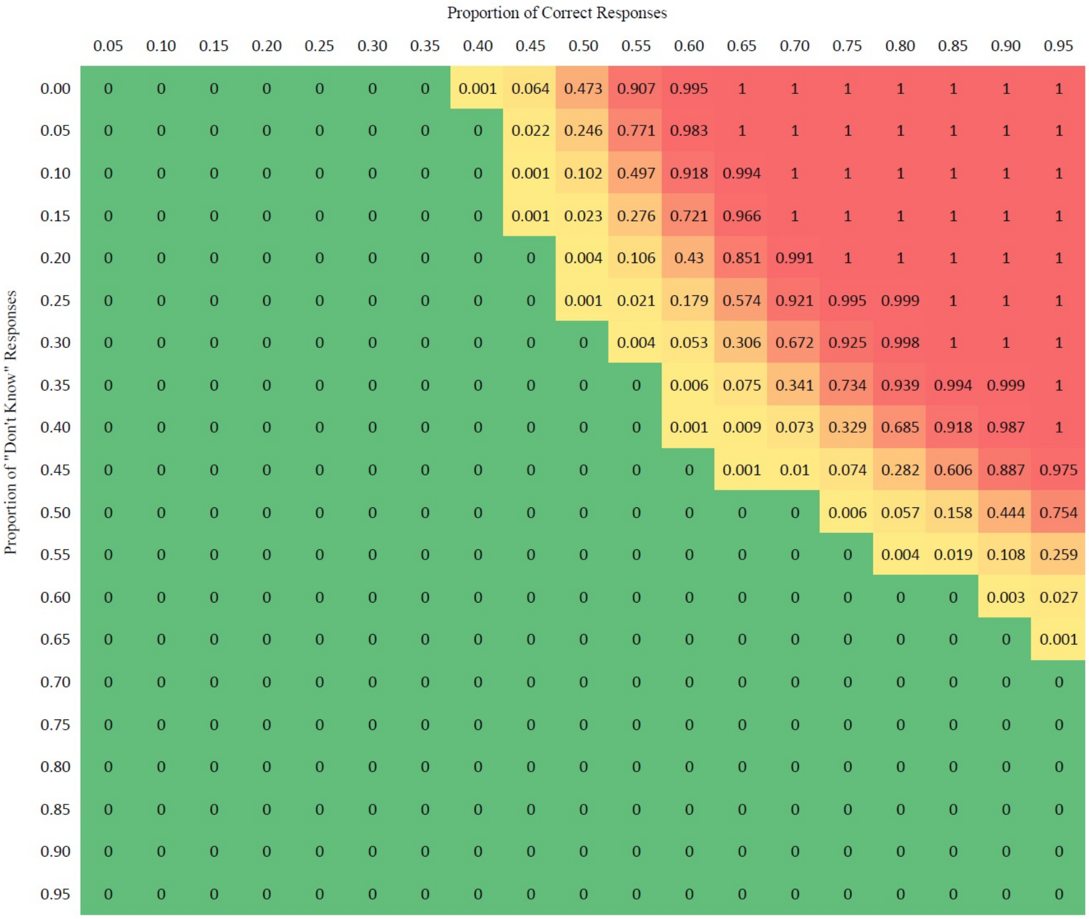
Proportion of "Don't Know" Responses	Proportion of Correct Responses																		
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.00	0	0	0	0	0	0	0	0.001	0.064	0.473	0.907	0.995	1	1	1	1	1	1	1
0.05	0	0	0	0	0	0	0	0	0.025	0.269	0.788	0.983	1	1	1	1	1	1	1
0.10	0	0	0	0	0	0	0	0	0.002	0.128	0.561	0.942	0.995	1	1	1	1	1	1
0.15	0	0	0	0	0	0	0	0	0.003	0.044	0.36	0.795	0.978	1	1	1	1	1	1
0.20	0	0	0	0	0	0	0	0	0	0.011	0.186	0.571	0.915	0.995	1	1	1	1	1
0.25	0	0	0	0	0	0	0	0	0	0.002	0.05	0.303	0.731	0.967	0.999	1	1	1	1
0.30	0	0	0	0	0	0	0	0	0	0.002	0.007	0.122	0.465	0.826	0.974	0.999	1	1	1
0.35	0	0	0	0	0	0	0	0	0	0	0.001	0.026	0.184	0.572	0.878	0.986	0.999	1	1
0.40	0	0	0	0	0	0	0	0	0	0	0	0.006	0.041	0.201	0.553	0.882	0.978	0.999	1
0.45	0	0	0	0	0	0	0	0	0	0	0	0	0.007	0.049	0.212	0.546	0.843	0.982	0.999
0.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0.007	0.047	0.167	0.423	0.757	0.948
0.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.004	0.019	0.093	0.311	0.602
0.60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.005	0.037	0.14
0.65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.002	0.008
0.70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.4: Heat Map for James' BI with Weights 0/0.4/0.7



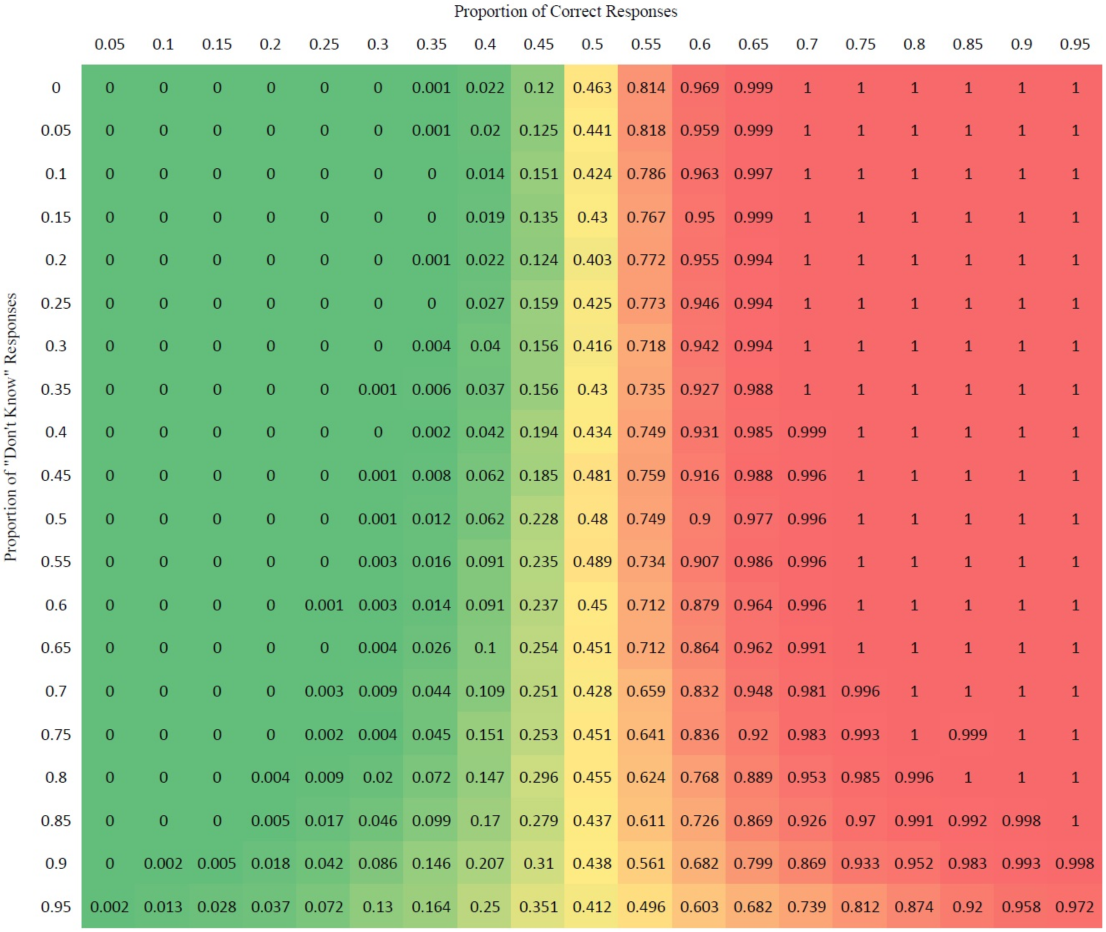
	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.00	0	0	0	0	0	0	0	0.001	0.064	0.473	0.907	0.995	1	1	1	1	1	1	1
0.05	0	0	0	0	0	0	0	0	0.022	0.246	0.771	0.983	1	1	1	1	1	1	1
0.10	0	0	0	0	0	0	0	0	0.001	0.102	0.497	0.918	0.994	1	1	1	1	1	1
0.15	0	0	0	0	0	0	0	0	0.001	0.023	0.276	0.721	0.966	1	1	1	1	1	1
0.20	0	0	0	0	0	0	0	0	0	0.004	0.106	0.43	0.851	0.991	1	1	1	1	1
0.25	0	0	0	0	0	0	0	0	0	0.001	0.021	0.179	0.574	0.921	0.995	0.999	1	1	1
0.30	0	0	0	0	0	0	0	0	0	0	0.004	0.053	0.306	0.672	0.925	0.998	1	1	1
0.35	0	0	0	0	0	0	0	0	0	0	0	0.006	0.075	0.341	0.734	0.939	0.994	0.999	1
0.40	0	0	0	0	0	0	0	0	0	0	0	0.001	0.009	0.073	0.329	0.685	0.918	0.987	1
0.45	0	0	0	0	0	0	0	0	0	0	0	0	0.001	0.01	0.074	0.282	0.606	0.887	0.975
0.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.006	0.057	0.158	0.444	0.754
0.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.004	0.019	0.108	0.259
0.60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.003	0.027
0.65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.001
0.70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.5: Heat Map for James' BI with Weights 0/0.5/0.8



	Proportion of Correct Responses																		
Proportion of "Don't Know" Responses	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.00	0	0	0	0	0	0	0	0.001	0.064	0.473	0.907	0.995	1	1	1	1	1	1	1
0.05	0	0	0	0	0	0	0	0	0.022	0.246	0.771	0.983	1	1	1	1	1	1	1
0.10	0	0	0	0	0	0	0	0	0.001	0.102	0.497	0.918	0.994	1	1	1	1	1	1
0.15	0	0	0	0	0	0	0	0	0.001	0.023	0.276	0.721	0.966	1	1	1	1	1	1
0.20	0	0	0	0	0	0	0	0	0	0.004	0.106	0.43	0.851	0.991	1	1	1	1	1
0.25	0	0	0	0	0	0	0	0	0	0.001	0.021	0.179	0.574	0.921	0.995	0.999	1	1	1
0.30	0	0	0	0	0	0	0	0	0	0	0.004	0.053	0.306	0.672	0.925	0.998	1	1	1
0.35	0	0	0	0	0	0	0	0	0	0	0	0.006	0.075	0.341	0.734	0.939	0.994	0.999	1
0.40	0	0	0	0	0	0	0	0	0	0	0	0.001	0.009	0.073	0.329	0.685	0.918	0.987	1
0.45	0	0	0	0	0	0	0	0	0	0	0	0	0.001	0.01	0.074	0.282	0.606	0.887	0.975
0.50	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.006	0.057	0.158	0.444	0.754
0.55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.004	0.019	0.108	0.259
0.60	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.003	0.027
0.65	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0.001
0.70	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.75	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.80	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.85	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.90	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0.95	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Figure 3.6: Heat Map for James' BI with Weights 0/0.5/1



Proportion of Correct Responses

Proportion of "Don't Know" Responses

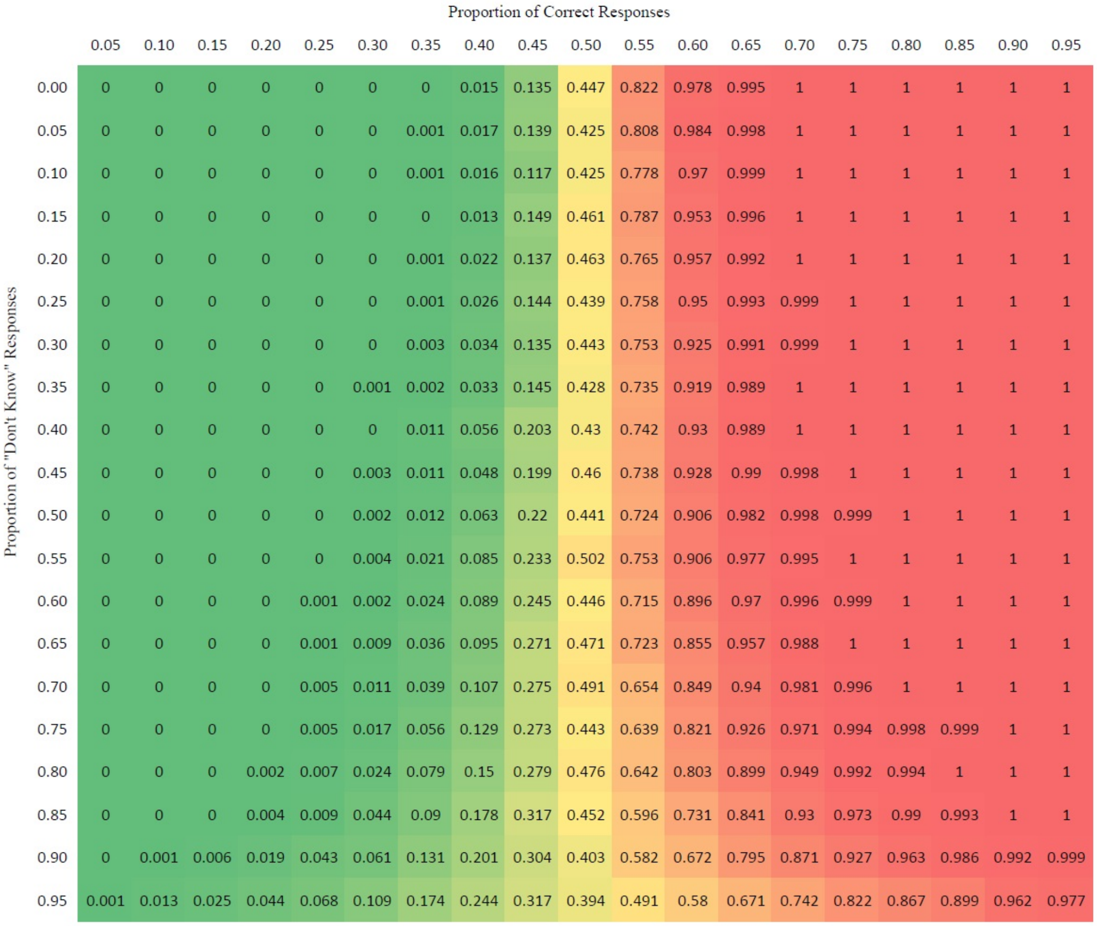
	0.05	0.1	0.15	0.2	0.25	0.3	0.35	0.4	0.45	0.5	0.55	0.6	0.65	0.7	0.75	0.8	0.85	0.9	0.95
0	0	0	0	0	0	0	0.001	0.022	0.12	0.463	0.814	0.969	0.999	1	1	1	1	1	1
0.05	0	0	0	0	0	0	0.001	0.02	0.125	0.441	0.818	0.959	0.999	1	1	1	1	1	1
0.1	0	0	0	0	0	0	0	0.014	0.151	0.424	0.786	0.963	0.997	1	1	1	1	1	1
0.15	0	0	0	0	0	0	0	0.019	0.135	0.43	0.767	0.95	0.999	1	1	1	1	1	1
0.2	0	0	0	0	0	0	0.001	0.022	0.124	0.403	0.772	0.955	0.994	1	1	1	1	1	1
0.25	0	0	0	0	0	0	0	0.027	0.159	0.425	0.773	0.946	0.994	1	1	1	1	1	1
0.3	0	0	0	0	0	0	0.004	0.04	0.156	0.416	0.718	0.942	0.994	1	1	1	1	1	1
0.35	0	0	0	0	0	0.001	0.006	0.037	0.156	0.43	0.735	0.927	0.988	1	1	1	1	1	1
0.4	0	0	0	0	0	0	0.002	0.042	0.194	0.434	0.749	0.931	0.985	0.999	1	1	1	1	1
0.45	0	0	0	0	0	0.001	0.008	0.062	0.185	0.481	0.759	0.916	0.988	0.996	1	1	1	1	1
0.5	0	0	0	0	0	0.001	0.012	0.062	0.228	0.48	0.749	0.9	0.977	0.996	1	1	1	1	1
0.55	0	0	0	0	0	0.003	0.016	0.091	0.235	0.489	0.734	0.907	0.986	0.996	1	1	1	1	1
0.6	0	0	0	0	0.001	0.003	0.014	0.091	0.237	0.45	0.712	0.879	0.964	0.996	1	1	1	1	1
0.65	0	0	0	0	0	0.004	0.026	0.1	0.254	0.451	0.712	0.864	0.962	0.991	1	1	1	1	1
0.7	0	0	0	0	0.003	0.009	0.044	0.109	0.251	0.428	0.659	0.832	0.948	0.981	0.996	1	1	1	1
0.75	0	0	0	0	0.002	0.004	0.045	0.151	0.253	0.451	0.641	0.836	0.92	0.983	0.993	1	0.999	1	1
0.8	0	0	0	0.004	0.009	0.02	0.072	0.147	0.296	0.455	0.624	0.768	0.889	0.953	0.985	0.996	1	1	1
0.85	0	0	0	0.005	0.017	0.046	0.099	0.17	0.279	0.437	0.611	0.726	0.869	0.926	0.97	0.991	0.992	0.998	1
0.9	0	0.002	0.005	0.018	0.042	0.086	0.146	0.207	0.31	0.438	0.561	0.682	0.799	0.869	0.933	0.952	0.983	0.993	0.998
0.95	0.002	0.013	0.028	0.037	0.072	0.13	0.164	0.25	0.351	0.412	0.496	0.603	0.682	0.739	0.812	0.874	0.92	0.958	0.972

Figure 3.7: Heat Map for Bang's NewBI (Control)

3.3 Bang's NewBI

The heat maps generated by the index proposed by Bang et.al. produced very similar results to that of 3.1, as seen in Figures 3.8 and 3.7. They differ however in that the region that we called the *transition region* for the kappa statistic is wider when using the NewBI for both the heat map generated by the data for the treatment arm as well as the heat map generated by the data from the control arm.

Figures 3.8 and 3.7 also show the same widening that was observed in figure 3.1



	Proportion of Correct Responses																		
Proportion of "Don't Know" Responses	0.05	0.10	0.15	0.20	0.25	0.30	0.35	0.40	0.45	0.50	0.55	0.60	0.65	0.70	0.75	0.80	0.85	0.90	0.95
0.00	0	0	0	0	0	0	0	0.015	0.135	0.447	0.822	0.978	0.995	1	1	1	1	1	1
0.05	0	0	0	0	0	0	0.001	0.017	0.139	0.425	0.808	0.984	0.998	1	1	1	1	1	1
0.10	0	0	0	0	0	0	0.001	0.016	0.117	0.425	0.778	0.97	0.999	1	1	1	1	1	1
0.15	0	0	0	0	0	0	0	0.013	0.149	0.461	0.787	0.953	0.996	1	1	1	1	1	1
0.20	0	0	0	0	0	0	0.001	0.022	0.137	0.463	0.765	0.957	0.992	1	1	1	1	1	1
0.25	0	0	0	0	0	0	0.001	0.026	0.144	0.439	0.758	0.95	0.993	0.999	1	1	1	1	1
0.30	0	0	0	0	0	0	0.003	0.034	0.135	0.443	0.753	0.925	0.991	0.999	1	1	1	1	1
0.35	0	0	0	0	0	0.001	0.002	0.033	0.145	0.428	0.735	0.919	0.989	1	1	1	1	1	1
0.40	0	0	0	0	0	0	0.011	0.056	0.203	0.43	0.742	0.93	0.989	1	1	1	1	1	1
0.45	0	0	0	0	0	0.003	0.011	0.048	0.199	0.46	0.738	0.928	0.99	0.998	1	1	1	1	1
0.50	0	0	0	0	0	0.002	0.012	0.063	0.22	0.441	0.724	0.906	0.982	0.998	0.999	1	1	1	1
0.55	0	0	0	0	0	0.004	0.021	0.085	0.233	0.502	0.753	0.906	0.977	0.995	1	1	1	1	1
0.60	0	0	0	0	0.001	0.002	0.024	0.089	0.245	0.446	0.715	0.896	0.97	0.996	0.999	1	1	1	1
0.65	0	0	0	0	0.001	0.009	0.036	0.095	0.271	0.471	0.723	0.855	0.957	0.988	1	1	1	1	1
0.70	0	0	0	0	0.005	0.011	0.039	0.107	0.275	0.491	0.654	0.849	0.94	0.981	0.996	1	1	1	1
0.75	0	0	0	0	0.005	0.017	0.056	0.129	0.273	0.443	0.639	0.821	0.926	0.971	0.994	0.998	0.999	1	1
0.80	0	0	0	0.002	0.007	0.024	0.079	0.15	0.279	0.476	0.642	0.803	0.899	0.949	0.992	0.994	1	1	1
0.85	0	0	0	0.004	0.009	0.044	0.09	0.178	0.317	0.452	0.596	0.731	0.841	0.93	0.973	0.99	0.993	1	1
0.90	0	0.001	0.006	0.019	0.043	0.061	0.131	0.201	0.304	0.403	0.582	0.672	0.795	0.871	0.927	0.963	0.986	0.992	0.999
0.95	0.001	0.013	0.025	0.044	0.068	0.109	0.174	0.244	0.317	0.394	0.491	0.58	0.671	0.742	0.822	0.867	0.899	0.962	0.977

Figure 3.8: Heat Map for Bang’s NewBI (Treatment)

calculated using Cohen’s Kappa. The widening here occurs also as a function of the increasing proportion of don’t know responses in how the index can be affected by aberrant sets of observations.

Based on the simulation study and the comparison of the various indices it is the opinion of the author that the most appropriate index to employ for the further study of ways that the assessment of blinding can be improved is the NewBI developed by Bang. While it is very similar in characteristics to Cohen’s Kappa, the ability to asses the unblinding in the treatment arm and the control arm separately is a potentially valuable tool in checking for possible response bias. The gradual fashion with which the index transitions from“Likely to not reject H_0” to “Likely to reject H_0” is also seemingly more in line with what we would intuitively expect than the results produced by Cohen’s Kappa. For this reason, the NewBI will be employed in the final goal for this paper, which is the development of statistical methodology to determine the possible source behind unblinding.

Chapter 4

Determining the Causes of Unblinding

4.1 Assumed causes of unblinding

When blinding fails, this is called unblinding. This happens when, for whatever reason, the patient becomes aware of the treatment arm they were assigned to, or the medical practitioner administering the treatment or assessing the results of the study becomes aware of the assignment for a particular patient. This can happen for a variety of reasons. Generally these reasons can be classified into two categories: expected and problematic. If a participant is assigned to receive the actual drug or treatment in a study and not the placebo, an obvious physiological result due to the efficacy of the treatment can be the source of unblinding. This might occur, for example, if the participant were involved in a study investigating a medication designed to counter erectile dysfunction. Positive results would likely indicate the participant was assigned to receive the actual treatment. This type of unblinding, while it may introduce some amount of bias, is expected and unavoidable. The second type of unblinding is considered problematic because it is a result of a failure to adopt appropriate protocols, or a result of the practitioners failing to adequately adhere to the protocols in place. This type of unblinding is certainly avoidable and could lead to the introduction of bias. The whole reason for the employment of protocols to institute blinding is to avoid such possible bias, since it brings the validity of the entire trial into question. There are many who believe that it is impossible to discern whether unblinding occurs as a result of the treatment effects or as a result of failed protocols. The following method can be used, however, to determine not only if unblinding occurs, but also if it occurs as a result of failed protocols.

4.2 A Statistical Method for Detecting a Possible Failure in Protocols

Clinical trials are often conducted at multiple centers, in which case the trial is called a multicenter trial. This gives us an opportunity to investigate the amount of unblinding at each location and compare it to the unblinding experienced at other locations to see if there is a statistically significantly elevated amount of unblinding at any of the locations where the trial was executed. Since the treatments would be identical across the various sites, any significant change in the amount of unblinding could be attributed to the failure of protocols. This is the cause of unblinding that bears further investigation, since the protocols are in place to help insure freedom from bias as well as an absence of outright fraud.

The method which was developed involves calculating the blinding index for each location separately. The NewBI is being employed for this purpose, which will also give results on the possible differences between the treatment arms at each location. A simulation was run in order to demonstrate the method. This simulation will have treatment arms that we will generically call the drug arm and the control arm. One

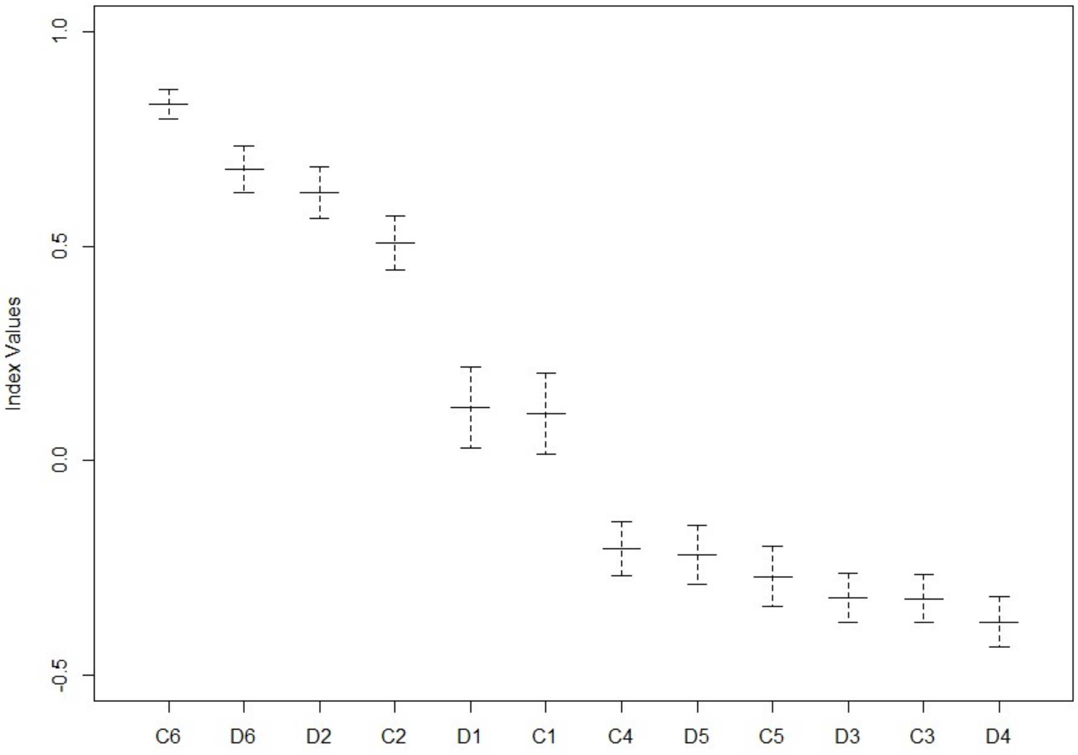


Figure 4.1: Plot of Blinding Measurements Between Centers

thousand participants were randomly assigned to either the drug or control arm, with each arm containing five hundred participants. Each participant was randomly assigned to one of six centers. Each center was randomly assigned proportions for correct, incorrect and don't know responses. The possible proportions of don't know responses were 0.2, 0.4 or 0.6. The possible proportions of correct responses among those who did not respond don't know were 0.2, 0.4, 0.6, 0.8 and 1.0. The remaining responses were assigned to an incorrect response.

The blinding index was calculated for each treatment arm separately, and a 95% confidence interval was created and plotted in Figure 4.1. The actual treatment, or drug was labeled *D* while the control was labeled *C*.

The results are listed in descending order to help identify areas of concern. Figure 4.1 shows that the twelve results are situated in four distinct groups. These preliminary results are enough to direct us as to possible problems. This would include the fact that C6 seems to be unblinded as well as being significantly different than D6. We also see that centers 1, 2 and 6 all seem to exhibit positive unblinding, which indicates more participants responded correctly than can be accounted for by random chance. We also see that C4 and D4 may be demonstrating significantly different results. In order to fully justify any conclusions a more rigorous approach is required. Such an approach would be to compare each of the center / treatment combinations pairwise with each other center / treatment combination to see if there is a statistically significant difference. The comparisons rely on the assumption that the indices are normally distributed, such that:

$$BI_n \sim N(\mu_n, \sigma_n) \tag{4.1}$$

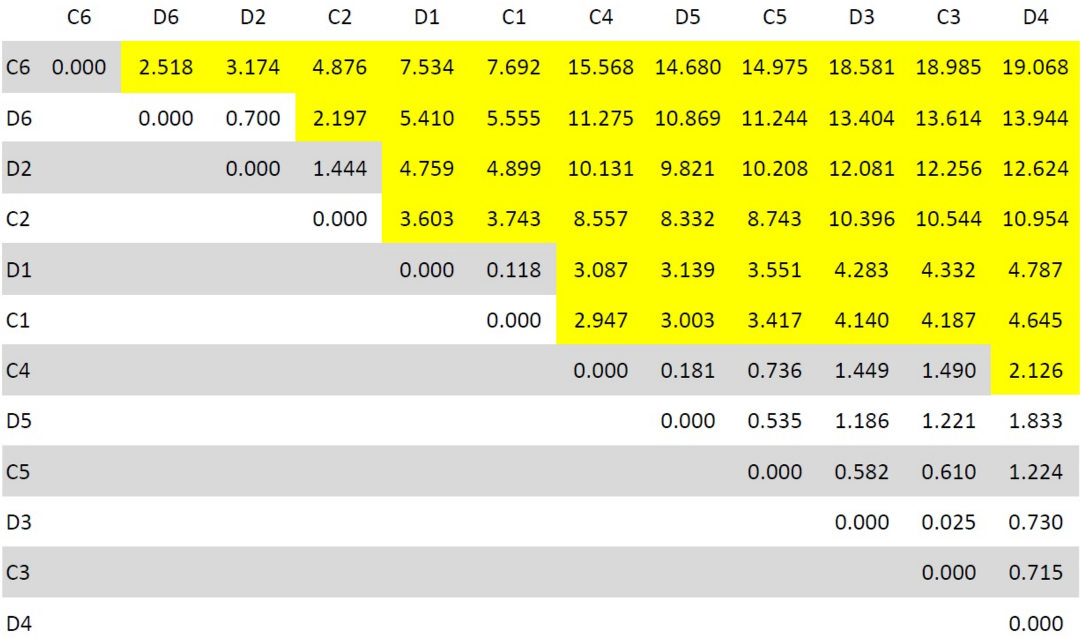
This assumption was tested and confirmed in R using the Shapiro-Wilk normality test. Under this assumption we can then create a comparison pairwise between the index calculated at two centers based upon the following null hypothesis:

$$H_0 : \mu_n = \mu_m \text{ for } B_n \text{ and } B_m \tag{4.2}$$

and

$$H_a : \mu_n = \mu_m \tag{4.3}$$

With this in mind we can calculate a Z score for the difference between B_n and B_m, where B_n and B_m represent the index calculated for two of the centers, using the following formula:



	C6	D6	D2	C2	D1	C1	C4	D5	C5	D3	C3	D4
C6	0.000	2.518	3.174	4.876	7.534	7.692	15.568	14.680	14.975	18.581	18.985	19.068
D6		0.000	0.700	2.197	5.410	5.555	11.275	10.869	11.244	13.404	13.614	13.944
D2			0.000	1.444	4.759	4.899	10.131	9.821	10.208	12.081	12.256	12.624
C2				0.000	3.603	3.743	8.557	8.332	8.743	10.396	10.544	10.954
D1					0.000	0.118	3.087	3.139	3.551	4.283	4.332	4.787
C1						0.000	2.947	3.003	3.417	4.140	4.187	4.645
C4							0.000	0.181	0.736	1.449	1.490	2.126
D5								0.000	0.535	1.186	1.221	1.833
C5									0.000	0.582	0.610	1.224
D3										0.000	0.025	0.730
C3											0.000	0.715
D4												0.000

Figure 4.2: Z Scores Comparing Location Results

$$Z = \frac{BI_n - BI_m}{\sqrt{\sigma_n^2 + \sigma_m^2}} \tag{4.4}$$

With $\alpha = 0.05$, we fail to reject H_0 for Z values between -1.96 and 1.96. In other words, under the null hypothesis:

$$-1.96 \leq Z \leq 1.96 \tag{4.5}$$

We can safely reject the null hypothesis for Z values outside of this range, meaning there is a statistically significant difference between the blinding indices BI_n and BI_m.

Table 4.2 shows the Z scores for pairwise comparisons. We see from the Z scores that some of our initial observations are shown to have statistically significant foundations. One item that would be of concern in an actual study is that the assessment of blinding for C6 is significantly different than that of D6, meaning in this case the participants in the control arm had a significantly higher rate of unblinding than those in the drug arm. Similarly, the index is significantly higher in C4 than in D4. This might indicate that the practitioners administering the study medications or placebos in locations four and six may not have strictly followed the appropriate protocols. This could be a sign that unblinding occurred as a result the efficacy of the medication if the unblinding was higher in the drug arm. The higher incidence of unblinding in the control arm might indicate participants speaking with each other and discussing outcomes.The lack of a physiological response to the placebo might be compared to the response resulting from the actual treatment, resulting in unblinding.

One of the reasons the NewBI was chosen for this final simulation was the ability to assess the treatment arms separately. Figure 4.3 shows us the blinding index assessed for the Drug arm. The blue line represents the overall blinding index for the drug arm and the red lines indicate the 95% confidence interval for that index. We see from this figure that centers six and two exhibit a level of unblinding which bears further investigation.

As before we will calculate the Z scores for each center, this time comparing them to the overall index for the drug arm. Table 4.1 shows those scores. Equation 4.6 is used to calculate the Z scores, with BI_d being the overall index calculated for the drug arm and σ_d^2 the variance for the drug arm. We can observe from Table 4.1 that centers six and two have statistically significant positive unblinding with Z scores of 7.274 and 6.350 respectively.

$$Z = \frac{BI_n - BI_d}{\sqrt{\sigma_n^2 + \sigma_d^2}} \tag{4.6}$$

We can now examine the results for the indices calculated for the control arm. Figure 4.4 give us a preliminary idea regarding what we should expect. We once again calculate the Z scores for these and observe from Table 4.2 that centers six and two exhibit positive

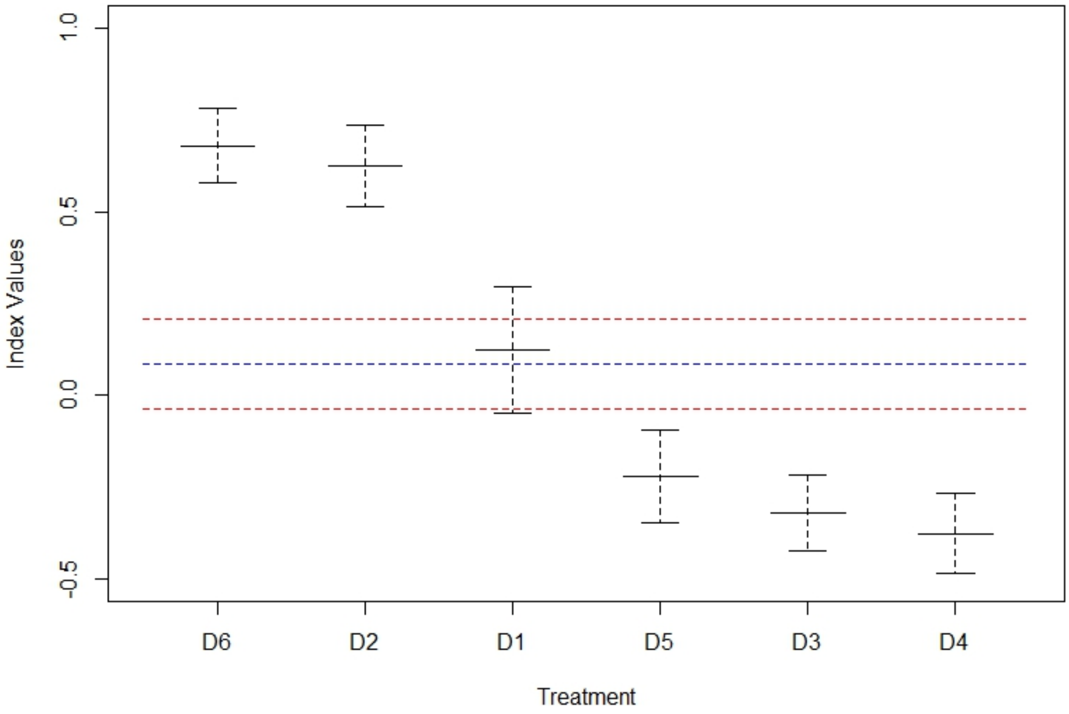


Figure 4.3: Indices for the Drug arm

Table 4.1: Z Scores for the Drug arm

D6	D2	D1	D5	D3	D4
7.274	6.350	0.353	-3.389	-4.898	-5.508

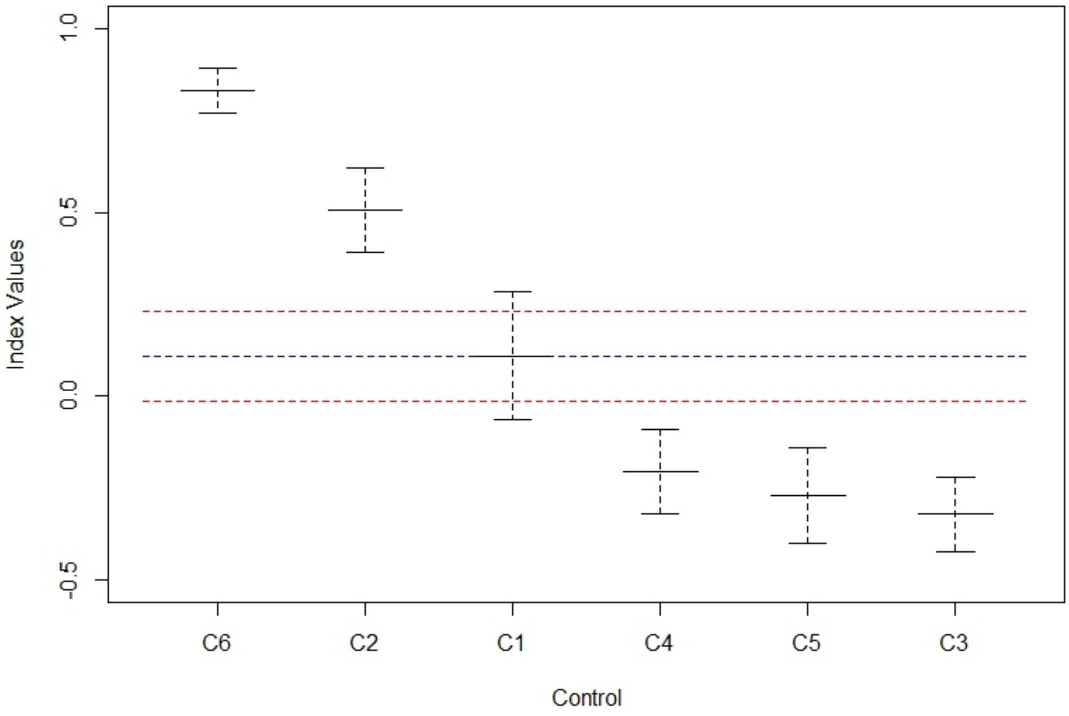


Figure 4.4: Indices for the Control arm

unblinding which is significantly greater than the index calculated on the entire control arm. By examining the indices in this way we can see if there is significant unblinding without worrying whether the unblinding has occurred as a result of the efficacy of the treatment.

$$Z = \frac{BI_n - BI_c}{\sqrt{\sigma_n^2 + \sigma_c^2}} \tag{4.7}$$

This simulation demonstrates that it is possible to observe unblinding that is likely to have occurred as a result in a failure to follow protocols. Unblinding which was caused by an easily discernible physiological effect would not account for such results since

Table 4.2: Z Scores for the Control arm

C6	C2	C1	C4	C5	C3
10.447	4.671	0.003	-3.672	-4.167	-5.322

the same medication or placebo would be used study-wide. Examining trials in this manner could improve the assessment of the success of blinding as well as increasing the overall accountability required of practitioners. The validity of a trial can be further demonstrated when possibilities for bias are eliminated, and this method can verify that those steps were successful.

Bibliography

Bang, H., Ni, L., and Davis, C. E. (2004). Assessment of Blinding in Clinical Trials. *Controlled Clinical Trials*, 25(2):143–156.

Freidman, L. M., Furberg, C. D., and DeMets, D. (2010). *Fundamentals of Clinical Trials*. Springer-Verlag New York.

James, K. E., Bloch, D. A., Lee, K. K., Kraemer, H. C., and Fuller, R. K. (1996). An Index for Assessing Blindness in a Multi-Centre Clinical Trial: Disulfiram for Alcohol Cessation - a VA Cooperative Study. *Statistics in Medicine*, 15:1421–1434.