ABSTRACT

On the Measurement of Blinding in Randomized, Controlled Trials

Forrest C. Williamson, Ph.D.

Chairpersons: Jane L. Harvill, and James D. Stamey

A key feature to many randomized, controlled trials is that they implement a blind; that is, subjects, experimenters, or both are unaware as to which treatment arm an individual has been assigned. The purpose of the blind is to reduce bias and improve retention. The importance of blinding has been emphasized by groups such as the FDA and CONSORT, but the reporting of blinding is not standard and the quantification of blinding success is rare. Two blinding indexes have been proposed to measure the success of blinding in randomized, controlled trials. The James index relies heavily on respondents saying that they do not know which treatment group an individual is assigned to. The Bang index looks more at the proportion of guesses and whether or not they suggest random or informed guessing. The theory behind the Bang index does not allow respondents to guess among more than two groups, including the control group. We have generalized the Bang index to allow for any number of arms. We find that our index, FBI, is powerful to detect events that cannot be measured using the James index. Also, implementing a Bayesian approach using a flat conjugate prior structure yields similar results to the frequentist approach James and Bang take. A guidance is given to aid in educating regulatory officials and trial investigators on the importance reporting blinding results and providing at least one quantitative measure of blinding success. We suggest that investigators report

indexes from both paradigms (James and Bang) to measure blinding efficacy for all blinded trials.

On the Measurement of Blinding in Randomized, Controlled Trials

by

Forrest C. Williamson, B.A., B.S.MTH., M.S.

A Dissertation

Approved by the Department of Statistical Science

Jack D. Tubbs, Ph.D., Chairperson

Submitted to the Graduate Faculty of
Baylor University in Partial Fulfillment of the
Requirements for the Degree
of
Doctor of Philosophy

Approved by the Dissertation Committee

Jane L. Harvill, Ph.D., Co-Chairperson

James D. Stamey, Ph.D., Co-Chairperson

Dean M. Young, Ph.D.

David Kahle, Ph.D.

Steven G. Driese, Ph.D.

Accepted by the Graduate School
December 2014

J. Larry Lyon, Ph.D., Dean

Page bearing signatures is kept on file in the Graduate School.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I'd like to thank Axio Research, LLC. for introducing me to blinding indexes, and my advisors, Jane L. Harvill and James D. Stamey, for their enthusiasm in taking on this challenge that was new to us all.

DEDICATION

In loving memory of Dr. Tom L. Bratcher,

without whom I never would have

come this far.

CHAPTER ONE

A Generalized Approach to Blinding Indexes

1.1 Why Measure Blinding?

A key feature to many randomized clinical trials is that they are blinded or double blinded. In a blinded trial, the subjects (and researchers in a double blind) do not know which treatment is being administered to an individual. The purpose of the blind is to eliminate bias incurred from the subject's or researcher's knowledge of the treatment being given, and to reduce drop out rate in the placebo or sham groups. In some cases subjects and/or researchers have good intuition as to which treatment is being administered, an intuition usually based on side effects or clinical results. In the Beta Blocker Heart Attack Study trial, Byington et al. Byington et al. (1985) reported that 80 percent of subjects who received propranolol correctly identified their treatment assignment, with an even higher accuracy among clinic personnel. Even though there was no formal unblinding in this circumstance, the belief in one treatment over others reintroduces the bias that we aim to eliminate through blinding. It is necessary to determine how well a study is blinded. The Food and Drug Administration (FDA) suggests that investigators, "administer a questionnaire at study completion to investigate the effectiveness of blinding the subjects and treating and evaluating the physicians" USF (2006). Unfortunately there is no standard for measuring blinding efficacy, nor does the FDA suggest how to approach the issue.

The movement for blinding assessment is relatively new. In trials that implement blinds, it is still uncommon for researchers to measure or report the success of blinding procedures. Hrbjartsson et al. Hrobjartsson et al. (2007) showed that only 31 out of 1599 trials ($< 2\%$) reported tests for the success of blinding. Yet blinding is essential to assure the internal validity of the study findings Park et al. (2008). Particularly, blinding can be a more serious issue in studies with a soft/subjective

endpoint Bang et al. (2010). One reason measuring blinding efficacy is not often implemented is because little is known about how to approach the issue statistically. Even though regulatory agencies encourage investigation in to blinding, limited statistical approaches make it difficult to require this analysis. Recent work has been done to further the idea of measuring blinding. Sections 1.3–1.6 provide an overview of existing approaches and explain why these methods are inadequate, followed by an overview of two blinding indexes. In Sections 1.8–1.10 we conclude this chapter by presenting our new approach to blinding indexes. Examples accompany each of the blinding measures.

1.2 Terminology

Before we go into further detail, we introduce some vocabulary that is used throughout the paper. Definitions come from www.clinicaltrials.gov/ct/info/glossary. To start, a trial *arm* refers to any treatment group in a clinical trial. Arms, trial arms, study arms, treatment arms, treatment groups, and groups are used synonymously in this text. A *blinded* trial is one in which participants are unaware of whether they are in a control arm or experimental arm of the study. The term blinded may also be extended to the investigators. A trial in which both participants and investigators are blinded is referred to as a *double blind* trial.

The control arm of a study often uses a *placebo*, an inactive substance (pill, liquid or powder) with no treatment value. Another possibility is to administer a *sham treatment* to the control group. A sham treatment is one which mimics the investigative treatment but omits a key element of the treatment or procedure. To distinguish between the two, consider a trial in which Arm A receives an experimental capsule to treat chronic headache and Arm B receives a capsule with sugar, known not to treat chronic headache. The capsule given to participants assigned to Arm B is a placebo. Imagine another trial in which Arm A receives an experimental procedure which requires surgery for which an abdominal incision is made, and Arm B receives a procedure in which a similar incision is made on the abdomen but the

experimental part of the surgery is not performed. In this case, Arm B is receiving a sham treatment. It is also possible that the control group receive neither placebo nor sham, but an already adopted standard treatment to which investigators wish to compare the experimental therapy.

Blinding efficacy and *successful blinding* are used to suggest that individuals are unable to identify correct arm allocations. An example of blinding efficacy is when participants or researchers are *randomly guessing*; that is, they are attempting to identify treatment allocation but the accuracy rate is similar to that expected by guessing randomly among the possibilities. For example, in a trial with two arms we would expect a person guessing at random to correctly identify treatment allocations fifty percent of the time. If participants or investigators are able to correctly identify treatment allocations at a rate higher than expected, we say that this group is *unblinded*. Make note that this does not necessarily mean that the group has been formally unblinded, where they are told which arm participants have been assigned to. Rather, it suggests that blinding efficacy/success has not been achieved because participants or investigators are successful at identifying treatment allocations. Finally, *reverse unblinding* (also referred to as *opposite guessing*) is a phenomenon in which participants or researchers are unsuccessful at identifying treatment allocations, beyond what would be expected by random guessing.

Finally, a quick review of study phases. *Preclinical* trials use non-human subjects, such as animals and bacteria. A *Phase I* trial tests a new drug for safety on a small group, usually healthy volunteers. In this stage, side effects are identified and doses are determined. *Phase II* can often be thought of as a trial-run for the big study to come. Researchers test the investigational treatment on subjects who have the condition being treated to determine if there are signs of efficacy, and to further investigate safety in the investigational group. The number of subjects enrolled in a Phase II trial can be small. The larger trial, which is used to show safety, measure efficacy, and report findings for regulatory approval is *Phase III* of the investigational

procedure. Finally, *Phase IV* describes the post-marketed performance of the treatment. Investigators can observe performance in subjects who were not considered for the Phase III trial, and can gather information on long-term efficacy and side effects.

1.3 Assessing Blinding: How it was Done in the Past

Many aspects of blinding assessment are widely debated. What questions do we ask to assess blinding integrity? When do we ask these questions? How do we analyze that data? What constitutes evidence for blinding? What are the implications on study findings? The answers to all of these questions and more lack a consensus among those who study blinding effectiveness. The most common approach to measuring blinding uses a questionnaire that would be administered with an exit survey. In the questionnaire, subjects are asked which treatment they believe they were on during the trial – or in the case of researchers, they are asked what treatment they believe individuals were on during the trial. All possible options are given so that nothing is hidden from the respondent. If the respondent is truly unsure about treatment, "Don't Know" (DK) is an available answer. It should be stated, however, that even the use of DK is somewhat controversial. The main concern is that more people will respond DK if given as an option, as it may be seen as the socially desirable answer (response bias) Bang et al. (2010).

From this exit-survey style data, a contingency table of actual treatment by guessed treatment is assembled. If there is a large proportion of correct guesses, then successful blinding may not have been achieved. We arrive at another debated issue: how to measure blinding success. As we have already stated, regulatory groups such as the FDA and protocolists such as the Consolidated Standards of Reporting Trials (CONSORT) encourage investigation of blinding success, but no specific method has been recommended. Due to the lack of a standard approach, numerous methods have been used to gauge how well a study is blinded Bang et al. (2010). An early means to quantify the success of blinding was taken by Hughes and Krahn Hughes and Krahn (1985) where they looked at the proportion of correct guesses versus the proportion

of incorrect guesses; if there was a significant difference between these proportions then blinding was deemed unsuccessful.

Park Park et al. (2008) summarized a number of other approaches that have been taken. The approaches all attempt to determine whether actual treatment and guessed treatment are associated. Hughes and Krahn Hughes and Krahn (1985) and Margraf et al. Margrat et al. (1991) used a traditional chi-square test for independence. Kolahi et al. Kolohi et al. (2009) used McNemar's test. These methods provide industry desired p-values, but not a numerical measure of the success of blinding. Wisner et al. Wisner et al. (2001) used the kappa (κ) statistic which measures agreement. However in blinding scenarios, disagreement is the more desirable outcome. Therefore the interpretability of κ is not obvious in this context. It is important to note that all of these approaches ignore the DK responses. DK responses are essentially thrown out before the analysis. So not only do these approaches leave out information, they also inflict bias upon themselves by omitting data that could be indicative of a successful blind.

1.4 What is Successful Blinding?

The methods mentioned in Section 1.3 ignore all DK responses, resulting in a source of bias in measuring blinding efficacy. There are two schools of thought on blinding indicators. The first states, "If subjects are properly blinded, they will guess 'randomly' among the available options." For example, it there are two treatment arms then we would expect randomly guessing subjects to guess correctly fifty percent of the time (one chance out of two). Anything other than fifty percent accuracy suggests guessing is not random; furthermore greater than fifty percent accuracy might suggest knowledge of treatment assignment. The second thought on blinding is, "If a person truly does not know their assignment, they have been successfully blinded." This is the source of DK responses. Thus, *blinding success* refers to one or both of these schools of thought being true. The more traditional statistical approaches that have been applied in the past have ignored the second school of thought (incorporat-

ing DK responses), making it easier to conclude that a blinding was unsuccessful. In general, blinding indexes try to harmonize the two components by measuring the randomness of the guesses while also incorporating information from the DK responses.

1.5 The James Blinding Index

James, Bloch, Lee, Kraemer and Fuller James et al. (1996) were the first to introduce a statistic that was built specifically to address the issue of blinding efficacy. James et al. claim that "the [DK] responses, if honestly reported, are the strongest indicator of success of the blinding procedures."

1.5.1 James BI

The James blinding index, BI, is a variant of the κ coefficient which measures agreement. James organizes the data in to a $(k+1) \times k$ frequency table, where k is the number of treatment arms. The actual treatment arms are represented in the columns, and the responses in the rows. An example frequency table using three treatment groups, along with some necessary notation, is given in Table 1.1. In each cell, n_{ij} is the frequency of guesses for treatment arm i, $i = 1, 2, \ldots, k+1$, among subjects assigned to arm j, $j = 1, 2, \ldots, k$. Please note that the notation presented in this section is the notation presented in James et al. (1996) . Notation will change for the Bang Blinding Index, and then will remain unchaged to the end.

Table 1.1: Frequency Table for James BI

Guessed Treatment	Actual Treatment			
	Treatment A	Treatment B	Placebo	Total
Treatment A	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
Treatment B	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
Placebo	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$
Don't Know	n_{01}	n_{02}	n_{03}	$n_{0\cdot}$
Total sample	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	N

James lists a few reasons why the κ statistic should not be used to measure the success of blinding procedures. First, κ measures agreement rather than disagreement.

Also, the bounds of the statistic depend on the number of response categories, which some consider to be an undesirable property for any index. Finally, κ ignores the DK responses. Therefore, James introduces a variation of a κ coefficient that is sensitive to the degree of disagreement rather than agreement, and also incorporates the DK responses.

The traditional κ measures agreement, and therefore assigns a weight of one to correct guesses and a weight of zero to incorrect guesses, with no venue for including DK responses. The measure is then related to what the score would have been had all guesses been random and to perfect agreement. Since correct guesses are "undesirable," the James κ_{D} assigns these a weight of zero, while a DK response receives a weight of one. Since incorrect guesses could be a sign of successful blinding, an intermediate weight is assigned to these guesses. James et al. use one-half as the weight for correctly guessing medicine but wrong dose (or guessing the wrong treatment while on a treatment), and three-fourths as the weight for guessing the wrong medication (or guessing between placebo or active treatment). They suggest that the result is not sensitive to the choice of this intermediate weight. This claim is supported by a simulation study summarized in the manuscript James et al. (1996).

The variant of κ that measures discordance is defined as

$$\kappa_{\mathrm{D}} = \frac{(p_{D_0} - p_{D_e})}{p_{D_e}},$$

where p_{D_0} is the weighted proportion of observed guesses, defined as

$$p_{D_0} = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{w_{ij}p_{ij}}{1 - P_{\mathrm{DK}}}, \quad \text{for } P_{\mathrm{DK}} \neq 1,$$

and p_{D_e} is the weighted proportion of expected guesses, defined as

$$p_{D_e} = \sum_{i=1}^{k}\sum_{j=1}^{k} \frac{w_{ij}p_{i\cdot}(p_{\cdot j} - p_{0j})}{(1 - P_{\mathrm{DK}})^2}.$$

The p_{ij} denote the expected relative frequency, and w_{ij} the weight given to the response in row i column j of the $(k+1) \times k$ frequency table. The James blinding index, BI, is defined as

$$BI = \frac{[1 + P_{\mathrm{DK}} + (1 - P_{\mathrm{DK}})\kappa_{\mathrm{D}}]}{2}.$$

Denote the estimated probabilities by $\hat{p}$. Thus $\hat{p}_{ij} = n_{ij}/N$, $\hat{p}_{0j} = n_{0j}/N$, $\hat{p}_{\cdot j} = n_{\cdot j}/N$, $\hat{p}_{i\cdot} = n_{i\cdot}/N$, and $\hat{P}_{\mathrm{DK}} = n_{0\cdot}/N$.

In this formulation, κ_{D} is similar to κ for the upper $k \times k$ portion of the table and is between negative one and one. The proportions P_{DK} and $(1 - P_{\mathrm{DK}})$ apportion the DK's and the guesses, respectively. Finally, adding one and dividing by two shift and scale the index so that it can take on any value between zero and one. For the James BI index, smaller values (closer to zero) indicate poor blinding and larger values (closer to one) represent successful blinding. For example, if every respondent in the study answers "DK" the index takes on a value of one; however if all responses are correct then BI equals zero. A value of one-half represents "random guessing."

James denotes the estimator as

$$\widehat{BI} = \frac{1 + \hat{P}_{\mathrm{DK}} + (1 - \hat{P}_{\mathrm{DK}})\hat{\kappa}_{\mathrm{D}}}{2},$$

with asymptotic variance

$$\begin{aligned}\mathrm{Var}(\widehat{BI}) &= N^{-1} \left\{ \frac{\sum_{i=1}^{k}\sum_{j=1}^{k} p_{ij}(1-p_{\mathrm{DK}})^2 \left[(1-p_{\mathrm{DK}})w_{ij} - (1+\kappa_{\mathrm{D}}) \sum_{r=1}^{k} \{p_{r\cdot} w_{rj} + (p_{\cdot r} - p_{0r}) w_{ir}\}\right]^2}{4\left[\sum_{i=1}^{k}\sum_{j=1}^{k} p_{i\cdot}(p_{\cdot j} - p_{0j}) w_{ij}\right]^2} \right. \\ &\quad \left. + p_{\mathrm{DK}}(1-p_{\mathrm{DK}}) - (1-p_{\mathrm{DK}})(1+\kappa_{\mathrm{D}}) \left[p_{\mathrm{DK}} + \frac{(1-p_{\mathrm{DK}})(1+\kappa_{\mathrm{D}})}{4}\right] \right\}.\end{aligned}$$

A jackknife procedure was also used to obtain variance estimates for the James index.

1.5.2 Cooperative Study No. 107

The U.S. Department of Veterans Affairs (VA) Cooperative Study No. 107 was a controlled, double-blinded, multicenter study of disulfiram treatment to treat alcoholism Fuller et al. (1986). One of three treatments – disulfiram 1mg, disulfiram 250mg, or riboflavin (placebo) – were administered to 605 men, along with counseling. The study concluded that disulfiram may reduce drinking frequency after relapse, but does not help in sustained abstinence along with counseling. James et al. (1996) use this study as an application to their blinding index, which most heavily depends on the DK responses.

Table 1.2 shows the study coordinator responses for VA Cooperative Study No. 107. It is of interest to examine if the investigators who dealt with patients were able to successfully determine which of those patients received an active treatment versus the placebo. The data is pooled across nine different hospital sites. For this study, $\widehat{BI} = 0.556$. The jack-knife confidence limits are 0.520 to 0.592, which is a near match to the reported asymptotic confidence limits, 0.521 to 0.592. James et al. interpret this as, "a response pattern close to that expected by random guessing, that is, partial but not complete blindness with no [DKs]." Nonetheless, the James blinding index was significantly greater than what would be expected for random guessing, at the 0.05 level.

Table 1.2: VA Coop Study No. 107 Study Coordinator Responses

Guessed Treatment	Actual Treatment			
	Disulfiram 1mg	Disulfiram 250mg	Riboflavin	Total
Disulfiram 1mg	41	27	22	90
Disulfiram 250mg	66	72	36	174
Riboflavin	30	24	64	118
Don't Know	44	51	52	147
Total	181	174	147	529

1.6 The Bang Index

Bang et al. Bang et al. (2004) constructed an alternative index to assess the efficacy of blinding in clinical trials. The index can be used for any blinded group; i.e., study subjects, researchers, etc. The basic form of the data is shown in Table 1.3. In the table, $p_{ij} = P(\text{guess } j|\text{assigned treatment } i)$ for $i = 1$ (drug), 2 (placebo), 3 (DK) where DK denotes "Don't know," and N is the total number of participants.

Table 1.3: Number of Subjects by Treatment Assignment and Guess in 2×3 Format

Assignment	Response			
	Drug	Placebo	DK	Total
Drug	$n_{11}(p_{11})$	$n_{12}(p_{12})$	$n_{13}(p_{13})$	N_1
Placebo	$n_{21}(p_{21})$	$n_{22}(p_{22})$	$n_{23}(p_{23})$	N_2
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	N

Motivation for the Bang index arises from that fact that different treatment arms don't necessarily have to reflect the same level of blinding or "unblinding." An aggregate statistic may be misleading because it forces a compromise between the non-homogeneous responses across treatment arms. A big criticism of the James index is that it cannot detect different behaviors between arms. For example, the James index could conclude that a trial is unblinded overall when only one of the arms guessed accurately even though all other arms are randomly guessing. On the other hand, it could also suggest adequate overall blinding even though some of the arms may be inadequately blinded.

1.6.1 Bang BI_i

Bang et al. introduced an arm-specific index in order to detect some of the features mentioned above. Define $r_{ii} = p_{ii}/(p_{i1} + p_{i2})$ to be the proportion of correct guesses among respondents in treatment arm i. We can rewrite this proportion as $r_{ii} = n_{ii}/(n_{i1} + n_{i2})$. In the absence of DK responses, we expect r_{ii} to be around one-half for random guessing. However, in the presence of DK's, the Bang index is

$$BI_i = (2r_{ii} - 1)\frac{(n_{i1} + n_{i2})}{(n_{i1} + n_{i2} + n_{i3})}, \tag{1.1}$$

which represents the proportion of individuals in the ith treatment arm who correctly guess their treatment beyond random balance. Bang et al. impose a trinomial distribution on the counts in the ith arm, with equal probabilities for each treatment arm, then a remainder probability (such that the probabilities sum to one) for DK responses. Note that the index can be negative, which would represent *reverse unblinding*, a scenario in which subjects incorrectly guess the treatment arm more

often than would be expected by random guessing. The statistic can be any number between negative one and one, where for the ith treatment arm

- negative one represents complete reverse unblinding,
- zero represents random guessing, and
- one represents complete unblinding,

and random guessing is evidence for blinding. The Bang index focuses on the balance between correct and incorrect guesses, as opposed to the James index which is more dependent on DK responses. Furthermore, James cannot detect reverse unblinding.

Under the trinomial assumption, we can calculate the variance of BI_i as,

$$Var(BI_i) = \frac{p_{i1}(1-p_{i1}) + p_{i2}(1-p_{i2}) + 2p_{i2}p_{i2}}{N_i}, \tag{1.2}$$

and the *estimated* variance is easily obtained by replacing p_{ij} with the observed proportion.

1.6.2 Cholesterol Reduction in Seniors Program

The Cholesterol Reduction in Seniors Program (CRISP) was a five-center pilot study to assess feasibility of recruitment and efficacy of cholesterol lowering in men and women over the age of 65 years LaRosa et al. (1994). Four hundred thirty one subjects with low-density lipoprotein cholesterol levels between 4.1 and 5.7 mmol/l were randomized into the study. Participants were followed for one year while on a cholesterol-lowering diet plus either placebo or the study drug, Lovastatin. Assessment of blinding is particularly important because whether or not patients found out their treatment assignment may have affected their compliance or attitude toward participation in the study. Not all people participated in the post-trial data collection process.

Table 1.4 shows the results from asking the subjects if they knew what treatment they were assigned to, along with the Bang index and 95% confidence interval. We see that the Lovastatin arm showed significant excess of correct guesses whereas the placebo group displays a pattern consistent with a random distribution of responses

between the two arms. We interpret the index as follows: 21% of participants correctly guessed their treatment beyond random chance in the Lovastatin arm, whereas only 1% did in the placebo arm. Bang admits the index is not developed for purposes of statistical testing, rather the index is to give a measurement for the extent of unblinding.

Table 1.4: CRISP Study Data

Assignment	Response			Total	Bang BI_i
	Lovastatin	Placebo	DK		(95% CI)
Lovastatin	82	25	170	277	0.21 (0.15, 0.26)
Placebo	27	29	83	139	0.01 $(-0.07, 0.10)$
Total	109	54	253	416	

If we were to calculate the James BI on this same data, we would see that $BI = 0.75$ (95% CI: 0.71, 0.78), implying that the CRISP study was well-blinded. Bang has shown us that, when looking at the treatment arms individually, there may be some concern as to the blinding success in the Lovastatin arm. However, because the placebo group was unable to guess their treatment beyond random chance the James index was biased toward a more conservative value.

1.7 James and Bang: What's the Difference and What's Missing?

The James index is very flexible because it can be applied to different types of data structures; most beneficially it can be applied for any number of treatment or control groups. However, because the index generates a single value for the entire study it can be misleading when a well blinded arm and a poorly blinded arm are both present within a trial, as seen in the CRISP trial. These arms conflict with each other, and the final index is a type of average of blinding success across arms. Bang

et al. (2004) developed another index to assess blinding. To address various concerns with the James index, the Bang index is treatment-arm specific. Thus it is capable of detecting dissimilar behavior between blinding arms. Furthermore the Bang index can detect "reverse unblinding" where an incorrect treatment is guessed at a higher rate than the correct treatment, a feature that would have gone undetected if using the James index Bang et al. (2010).

So far, the Bang index has only been developed under the case where there are two arms – specifically an active treatment arm and a placebo or control arm. The authors state that a "generalization to more than two arms is straightforward." However, this isn't necessarily the case. The rest of this chapter is dedicated to the development of an index that is similar to Bang's in theory, but that can accommodate any number of treatment groups, like the James BI.

1.8 A New Blinding Index

For reasons outlined in the previous section we prefer Bang et al.'s approach to measuring blinding. However, perhaps the biggest pitfall of Bang's index is that it is only useful when there are two groups: active treatment and placebo. In some cases, all "control" groups could be assigned to one placebo category, and all active treatments (whether experimental or current standard) into one treatment category. However, the problem with dichotomizing the groups after surveying the subjects is that there needs to be an equal number of controls to treatment groups, to keep with the equal probability "random guessing" assumption of Bang. Alternatively, we could dichotomize before administering the survey, only allowing subjects to respond: placebo, treatment, or DK. However, we lose information in doing this.

One of the advantages the Bang index holds over the James index is that it is treatment-arm specific. We can actually see which arms the sponsor had more trouble (ease) in maintaining the blind. If we merge groups from the beginning, we lose the ability to detect the specific arms in which blinding was most heavily broken, and therefore insight is lost in to what we need to do to better maintain blinding

success. We assert that the only time treatments should be combined in to one possible response is when there are multiple dosage levels of the same experimental treatment.

In what follows, we develop a new blinding index, FBI_i, $i = 1, 2, \ldots, k$, which can be used for any number of treatment arms and maintains Bang's philosophical approach to numerically assessing blinding. Table 1.5 extends notation used by Bang in previous sections.

Table 1.5: $k \times (k+1)$ Frequency Table for the FBI

Assignment	Response							Total
	1	2	$\cdots$	i	$\cdots$	k	DK	
1	n_{11}	n_{12}	$\cdots$	n_{1i}	$\cdots$	n_{1k}	$n_{1,k+1}$	N_1
2	n_{21}	n_{22}	$\cdots$	n_{2i}	$\cdots$	n_{2k}	$n_{2,k+1}$	N_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
i	n_{i1}	n_{i2}	$\cdots$	n_{ii}	$\cdots$	n_{ik}	$n_{i,k+1}$	N_i
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$
k	n_{k1}	n_{k2}	$\cdots$	n_{ki}	$\cdots$	n_{kk}	$n_{k,k+1}$	N_k
Total	$n_{\cdot 1}$	$n_{\cdot 2}$	$\cdots$	$n_{\cdot i}$	$\cdots$	$n_{\cdot k}$	$n_{\cdot k+1}$	N

1.8.1 The FBI_i Statistic

Suppose we have k treatment arms. Then n_{ij} represents the number of subjects in treatment arm i who guessed to be in treatment arm j, $i = 1, \ldots, k$ and $j = 1, \ldots, k, k+1$, where $j = k+1$ represents a DK response (See Table 1.5). Consider again $r_{ii} = p_{ii}/(p_{i1} + p_{i2})$ in the two-arm case. This is the proportion of respondents in arm i who guessed correctly, among those in arm i who did not respond "DK."

The straightforward generalization is

$$r_{ii} = \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij}}.$$

The generalized blinding index can be expressed as

$$FBI_i = \frac{kr_{ii} - 1}{k - 1} \sum_{j=1}^{k} p_{ij}, \tag{1.3}$$

which is the average of the "pairwise" Bang BI_i. We explore the concept of pairwise BI_i indexes in a later section.

1.8.1.1 Bounds. Recall that BI_i is between negative one and one, with a value of zero representing the best measure for blinding success. Positive values of BI_i represent unblinding and negative values suggest reverse unblinding. We consider three extreme cases to explore the bounds of the generalized index, FBI_i.

(1) First suppose all respondents guess correctly (with no DK responses). Then for each arm i, $i = 1, \ldots, k$,

$$\begin{aligned} n_{ii} &= N_i \\ p_{ii} &= n_{ii}/N_i = N_i/N_i = 1 \implies p_{ij} = 0 \forall j \neq i \\ r_{ii} &= \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij}} = \frac{1}{1} = 1 \\ FBI_i &= \frac{kr_{ii} - 1}{k - 1} \sum_{j=1}^{k} p_{ij} = \frac{k \times 1 - 1}{k - 1} \times 1 = 1. \end{aligned}$$

The most extreme degree of unblinding is represented by a FBI_i value of one. For this to happen, all guesses among subjects in arm i must guess correctly, with no DK guesses. This is the only scenario in which the statistic achieves its upper bound. The presence of any incorrect guesses or DK responses draws the FBI_i closer to zero.

(2) Next suppose the opposite where all respondents guess incorrectly (with no DK responses). Now, for $i = 1, \ldots k$,

$$\begin{aligned} n_{ii} &= 0 \\ p_{ii} &= n_{ii}/N_i = 0 \\ r_{ii} &= \frac{0}{\sum_{j=1}^{k} p_{ij}} = \frac{0}{1} = 0 \quad \left(\text{no DK responses} \implies \sum_{j=1}^{k} p_{ij} = 1 \right) \\ FBI_i &= \frac{kr_{ii} - 1}{k-1} \sum_{j=1}^{k} p_{ij} = \frac{k \times 0 - 1}{k-1} \times 1 = \frac{-1}{k-1}. \end{aligned}$$

The most extreme case for "reverse unblinding" is represented by $FBI_i = -(k-1)^{-1}$. This is the only scenario in which the statistic reaches its lower bound. The presence of any correct guesses or DK responses draws the FBI_i closer to zero.

(3) Finally suppose every subject responds DK. Then for all $i = 1, \ldots, k$,

$$\begin{aligned} n_{ii} &= 0 \\ p_{ii} &= 0 \\ r_{ii} &= 0 \qquad \left(\text{all DK responses} \implies \sum_{j=1}^{k} p_{ij} = 0 \right) \\ FBI_i &= \frac{kr_{ii} - 1}{k-1} \sum_{j=1}^{k} p_{ij} = \frac{k \times (-1)}{k-1} \times 0 = 0. \end{aligned}$$

The "best case" scenario when nobody attempts to guess what they are on (presumably because the blind is completely effective) is represented by $FBI_i = 0$. This is *not* the only scenario under which the FBI_i can achieve a value of zero. We show in Section 1.8.4 that the null value of FBI_i is zero; that is,

$$FBI_i = 0 \iff p_{ii} = \frac{1}{k-1} \sum_{j \neq i}^{k} p_{ij} \,. \tag{1.4}$$

Note that (1.4) is satisfied when all subjects respond DK since $p_{i1} = p_{i2} = \cdots = p_{ik} = 0$.

Therefore $FBI_i \in [-(k-1)^{-1}, 1]$, where for any $i = i, \ldots, k$,

- $-(k-1)^{-1}$ represents complete reverse unblinding,
- zero represents successful blinding (either random guessing or no guessing between treatment arms), and
- one represents complete unblinding.

It is important that the index is zero under the two cases where we have either completely random guessing, or where nobody guesses at treatment (all DK responses). At the beginning of the chapter, we discussed two paradigms of successful blinding: random guessing and DK responses. Either of these paradigms, or even a combination of the two, yield the same value of the blinding index. DK responses and random guessing both push the index toward zero, indicating effective blinding. This behavior is not present in the James index, which is in part why we prefer the Bang approach. Unlike the James and Bang approaches, the lower bound of the FBI_i is dependent upon the number of groups. We rely on simulation to see how powerful the statistic is in detecting reverse unblinding, since the presence of more groups pushes the lower bound closer to zero and could possibly yield insignificant results when, in fact there is reverse unblinding.

1.8.1.2 Special Case for $k = 2$. Suppose we have $k = 2$ treatment arms. It is easy to show that the FBI_i simplifies to the Bang blinding index. Assume $k = 2$, then

$$\begin{aligned} FBI_i &= \frac{2r_{ii} - 1}{2 - 1} \sum_{j=1}^{2} p_{ij} \\ &= (2r_{ii} - 1) \sum_{j=1}^{2} p_{ij} \\ &= BI_i \,. \end{aligned}$$

Thus BI_i is a special case of FBI_i when $k = 2$.

1.8.1.3 Estimating FBI_i. The estimated FBI_i index is given by

$$\widehat{FBI_i} = \frac{k\hat{r}_{ii} - 1}{k-1} \sum_{j=1}^{k} \hat{p}_{ij}, \tag{1.5}$$

where

$$\begin{aligned} \hat{r}_{ii} &= \frac{\hat{p}_{ii}}{\sum_{j=1}^{k} \hat{p}_{ij}}, \\ \hat{p}_{ij} &= \frac{n^*_{ij}}{N_i}, \end{aligned}$$

and n^*_{ij} represents the observed value of the random variable n_{ij}, the number of subjects who were in treatment group i but guessed group j. We expand of the form of the estimated index in the next section, which gives another formulation of FBI_i.

1.8.2 Alternative Representation

Since we can easily visualize the data in terms of the frequencies of responses in Table 1.5, it may be more intuitive to rewrite the statistic from (1.3) in terms of these frequencies. Knowing that for all i, j we can write $p_{ij} = n_{ij}/N_i$, we can express r_{ii} as

$$r_{ii} = \frac{p_{ii}}{\sum_{j=1}^{k} p_{ij}} = \frac{(n_{ii}/N_i)}{\left(\sum_{j=1}^{k} n_{ij}\right)/N_i} = \frac{n_{ii}}{\sum_{j=1}^{k} n_{ij}},$$

which represents the proportion of correct guesses among those who did not respond DK. Therefore, we rewrite the FBI_i index as

$$\begin{aligned} FBI_i &= \frac{kr_{ii} - 1}{k-1} \sum_{j=1}^{k} p_{ii} \\ &= \frac{k\left(\frac{n_{ii}}{\sum_{j=1}^{k} n_{ij}}\right) - 1}{k-1} \sum_{j=1}^{k} \frac{n_{ij}}{N_i} \\ &= \frac{k\left(\frac{n_{ii}}{N_i}\right) - \frac{1}{N_i}\sum_{j=1}^{k} n_{ij}}{k-1} \\ &= \frac{kn_{ii} - \sum_{j=1}^{k} n_{ij}}{(k-1)N_i} \\ &= \frac{(k-1)n_{ii} - \sum_{j\neq i}^{k} n_{ij}}{(k-1)N_i}. \end{aligned} \tag{1.6}$$

It is easy to see how we would estimate FBI_i from data simply by substituting the observed frequencies for their corresponding random variables. An alternative form for the estimator given in expression (1.5) is

$$\widehat{FBI}_i = \frac{(k-1)n^*_{ii} - \sum_{j\neq i}^{k} n^*_{ij}}{(k-1)N_i}. \tag{1.7}$$

Recall that BI_i only applied when $k = 2$. Suppose $k > 2$. Then we can compute "pairwise $BI_{i,j}$" which are simply the BI_i statistic comparing the two treatment groups i and $j \neq i$. Thus, there are $k-1$ pairwise $BI_{i,j}$. Then FBI_i becomes the average of the $k-1$ pairwise Bang blinding indexes, $BI_{i,j}$.

$$\begin{aligned} FBI_i &= \frac{(k-1)n_{ii} - \sum_{j\neq i}^{k} n_{ij}}{(k-1)N_i} \\ &= \frac{1}{k-1}\sum_{j\neq i}^{k}\left(\frac{n_{ii} - n_{ij}}{N_i}\right) \\ &= \frac{1}{k-1}\sum_{j\neq i}^{k} BI_{i,j}\,. \end{aligned}$$

We have established a solid link between the Bang blinding index and the generalized FBI_i index, providing additional insight into the equivalence of FBI_i and BI_i when $k = 2$.

1.8.3 Distributional Assumptions and Moments

As in Bang, we still have an arm-specific index. Each index is essentially conditioned on the true treatment assignment, and the FBI_i are calculated independently of one another. We consider a fixed treatment group, i, and work through the details of the index conditioned on being in the ith group. Contained in Table 1.6 is all of the data we need to compute FBI_i. Note Table 1.6 is simply the ith row of Table 1.5.

Table 1.6: Frequency Table for Respondents in the ith Group

Assignment	Response							Total
	1	2	$\cdots$	i	$\cdots$	k	DK	
i	n_{i1}	n_{i2}	$\cdots$	n_{ii}	$\cdots$	n_{ik}	$\$n_{i,k+1}$	N_i

The blinding index for group i is

$$FBI_i = \frac{(k-1)n_{ii} - \sum_{j\neq i}^{k} n_{ij}}{(k-1)N_i}.$$

The total number of respondents N_i in group i is fixed, and is $N_i = \sum_{j=1}^{k+1} n_{ij}$. However, the number of guesses for each response level is random, with some probability assumed for each response level. To model this relationship, we assume a multinomial distribution on the n_{ij}. Specifically,

$$n_{i1}, n_{i2}, \ldots, n_{ik} \sim \text{multinom}\left(N_i, p_{i1}, p_{i2}, \ldots, p_{ik}\right). \tag{1.8}$$

Once we impose this assumption, we know the following about the random variables,

$$\begin{aligned} E\left[n_{ij}\right] &= N_i p_{ij}, & (1.9)\\ Var\left[n_{ij}\right] &= N_i p_{ij}(1-p_{ij}), & (1.10)\\ Cov\left[n_{ij}, n_{il}\right] &= -N_i p_{ij} p_{il}\,. & (1.11) \end{aligned}$$

The FBI_i statistic is a linear combination of the random variables n_{ij}, $j \neq i = 1, \ldots, k$. Thus it is a straightforward exercise to find the expected value and variance of FBI_i.

The expected value of FBI_i is

$$\begin{aligned} E\left[FBI_i\right] &= E\left[\frac{(k-1)n_{ii} - \sum_{j\neq i}^{k} n_{ij}}{(k-1)N_i}\right]\\ &= \frac{1}{(k-1)N_i} E\left[(k-1)n_{ii} - \sum_{j\neq i}^{k} n_{ij}\right]\\ &= \frac{1}{(k-1)N_i}\left\{(k-1)E\left[n_{ii}\right] - \sum_{j\neq i}^{k} E\left[n_{ij}\right]\right\}\\ &= \frac{1}{(k-1)N_i}\left\{(k-1)N_i p_{ii} - N_i \sum_{j\neq i}^{k} p_{ij}\right\} \qquad \text{by (1.9)}\\ &= p_{ii} - \frac{1}{k-1}\sum_{j\neq i}^{k} p_{ij}\,. & (1.12) \end{aligned}$$

To find the variance of FBI_i, note that for any random variables X and any index of sequenced random variables Y_i,

$$\begin{aligned} Var\left[aX - \sum Y_i\right] &= a^2 Var\left[X\right] + Var\left[\sum Y_i\right] + 2Cov\left[aX, -\sum Y_i\right] \\ &= a^2 Var\left[X\right] + \sum Var\left[Y_i\right] + 2\sum_{i<j} Cov\left[Y_i, Y_j\right] \\ &\quad -2a\sum Cov\left[X, Y_i\right]. \end{aligned} \tag{1.13}$$

Using (1.13), we obtain a closed form expression for the variance of FBI_i.

$$\begin{aligned} Var\left[FBI_i\right] &= Var\left[\frac{(k-1)n_{ii} - \sum_{j\neq i}^{k} n_{ij}}{(k-1)N_i}\right] \\ &= \left\{(k-1)N_i\right\}^{-2} Var\left[(k-1)n_{ii} - \sum^{k} n_{ij}\right] \\ &= \left\{(k-1)N_i\right\}^{-2} \Bigg\{(k-1)^2 Var\left[n_{ii}\right] + \sum_{j\neq i}^{k} Var\left[n_{ij}\right] \\ &\quad -2(k-1)\sum_{j\neq i}^{k} Cov\left[n_{ii}, n_{ij}\right] + 2\sum_{\substack{j\neq i \\ j<l}}^{k} Cov\left[n_{ij}, n_{il}\right]\Bigg\} \qquad \text{by (1.13)} \\ &= \left\{(k-1)N_i\right\}^{-2} \Bigg\{(k-1)^2 N_i p_{ii}(1-p_{ii}) + \sum_{j\neq i}^{k} N_i p_{ij}(1-p_{ij}) \\ &\quad +2(k-1)\sum_{j\neq i}^{k} N_i p_{ii} p_{ij} - 2\sum_{\substack{j\neq i \\ j<l}}^{k} N_i p_{ij} p_{il}\Bigg\} \qquad \text{by (1.10), (1.11)} \\ &= \left\{(k-1)^2 N_i\right\}^{-1} \Bigg\{(k-1)^2 p_{ii}(1-p_{ii}) + \sum_{j\neq i}^{k} p_{ij}(1-p_{ij}) \\ &\quad +2(k-1)\sum_{j\neq i}^{k} p_{ii} p_{ij} - 2\sum_{\substack{j\neq i \\ j<l}}^{k} p_{ij} p_{il}\Bigg\}. \end{aligned} \tag{1.14}$$

1.8.4 Behavior Under Random Guessing

Our assumption for good blinding is that subjects who are guessing are guessing randomly; that is, we expect an equal proportion of guesses in each treatment group. Thus, the null hypothesis is

$$H_0 : p_{i1} = p_{i2} = \cdots = p_{ii} = \cdots = p_{ik}. \tag{1.15}$$

We also have the constraint that $\sum_{j=1}^{k+1} p_{ij} = 1$, and hence under the null hypothesis $p_{ij} \leq 1/k$. Note that nothing is assumed about $p_{i,k+1}$, the proportion of DK responses. One consequence of (1.15), which is less restrictive, is that $p_{ii} = \sum_{j \neq i}^{k} p_{ij}/(k-1)$. Therefore by using expression (1.12), the expected value of FBI_i under the null hypothesis of random guessing is zero,

$$\begin{aligned} E\left[FBI_i\right] &= p_{ii} - \frac{1}{k-1}\sum_{j \neq i}^{k} p_{ij}\,. \\ E_0\left[FBI_i\right] &= 0. \end{aligned}$$

Therefore, no matter the number of groups, we are comparing FBI_i against zero.

Similarly, we can reevaluate the variance given in (1.14) under the restriction in (1.15). First, we define new notation. Under the null hypothesis. $p_{i1} = \cdots = p_{ik} \equiv p$. With this notation, we write the variance under the null for FBI_i

$$\begin{aligned} Var\left[FBI_i\right] &= \left\{(k-1)^2 N_i\right\}^{-1}\left\{(k-1)^2 p(1-p) + (k-1)p(1-p)\right. \\ &\quad \left. +2(k-1)^2 p^2 - (k-1)(k-2)p^2\right\} \\ &= \frac{(k-1)p - (k-1)p^2 + p - p^2 + 2(k-1)p^2 - (k-2)p^2}{(k-1)N_i} \\ &= \frac{\left[-(k-1) - 1 + 2(k-1) - (k-2)\right]p^2 + \left[(k-1)+1\right]p}{(k-1)N_i} \\ &= \frac{kp}{(k-1)N_i}. \end{aligned}$$

1.8.5 Theoretical Simplification of Distributional Theory

From expression (1.6) it is apparent that the FBI_i statistic is measuring the difference between the number of correct guesses and the number of incorrect guesses. Additional weight is added to a correct guess. This additional weight is intuitively appealing since for more treatment groups, guessing correctly under a "random guessing" hypothesis is less likely. The weights are chosen so that the expected value under the null hypothesis is zero.

For measuring blinding success of subjects who were assigned to group i, the FBI_i statistic does not distinguish between the types of incorrect guesses. Therefore,

Table 1.6 can be simplified to Table 1.7. Using the notation introduced in Table 1.7, we have

$$FBI_i = \frac{(k-1)n_{i1} - n_{i2}}{(k-1)N_i}.$$

The main advantage to this approach is that, no matter how many groups there are, we can use a trinomial distribution to describe the frequencies. Implied with this advantage is that we have less p_{ij} to estimate ($k > 2$) than we would have otherwise. This will be more helpful when we study these blinding indexes under the Bayesian paradigm. For now, it serves well for computational simplicity.

Table 1.7: Simplification of Frequency Table for Respondents in ith Group

Actual	Response			Total
	Correct (i)	Incorrect	DK	
i	n_{i1} (p_{i1})	n_{i2} (p_{i2})	n_{i3} (p_{i3})	N_i

Formally written, using the format in Table 1.7, we assume

$$n_{i1}, n_{i2} \sim \text{trinom}(N_i, p_{i1}, p_{i2}). \tag{1.16}$$

From here we rewrite the expectation and variance for FBI_i. First, the expected value is

$$\begin{aligned} E\left[FBI_i\right] &= E\left[\frac{(k-1)n_{i1} - n_{i2}}{(k-1)N_i}\right] \\ &= \frac{(k-1)E\left[n_{i1}\right] - E\left[n_{i2}\right]}{(k-1)N_i} \\ &= \frac{(k-1)N_i p_{i1} - N_i p_{i2}}{(k-1)N_i} \\ &= \frac{(k-1)p_{i1} - p_{i2}}{k-1} \\ &= p_{i1} - \frac{1}{k-1}p_{i2}, \end{aligned}$$

which is equivalent to expression (1.12) because here p_{i1} represents the probability of a correct guess and p_{i2} represents the total probability of guessing incorrectly. There is no real computational advantage to the trinomial approach when calculating FBI_i

or the expected value. However, the variance is more simplistic. Using the new notation, the variance is

$$\begin{aligned} Var\left[FBI_i\right] &= Var\left[\frac{(k-1)n_{i1}-n_{i2}}{(k-1)N_i}\right] \\ &= \frac{(k-1)^2 Var\left[n_{i1}\right]+Var\left[n_{i2}\right]-2(k-1)Cov\left[n_{i1},n_{i2}\right]}{(k-1)^2N_i^2} \\ &= \frac{(k-1)^2N_ip_{i1}(1-p_{i1})+N_ip_{i2}(1-p_{i2})+2(k-1)N_ip_{i1}p_{i2}}{(k-1)^2N_i^2} \\ &= \frac{(k-1)^2p_{i1}(1-p_{i1})+p_{i2}(1-p_{i2})+2(k-1)p_{i1}p_{i2}}{(k-1)^2N_i}. \end{aligned}$$

Note that the null hypothesis is now

$$H_0:\, p_{i1}=\frac{1}{k-1}p_{i2},$$

or equivalently

$$H_0:\,(k-1)p_{i1}=p_{i2}\,.$$

Thus, the expected value and variance of FBI_i under the null hypothesis are

$$\begin{aligned} E\left[FBI_i\right] &= p_{i1}-\frac{1}{k-1}p_{i2}=0 \\ Var\left[FBI_i\right] &= \frac{(k-1)^2p_{i1}(1-p_{i1})+p_{i2}(1-p_{i2})+2(k-1)p_{i1}p_{i2}}{(k-1)^2N_i} \\ &= \frac{(k-1)^2p_{i1}(1-p_{i1})+(k-1)p_{i1}\left[1-(k-1)p_{i1}\right]+2(k-1)^2p_{i1}^2}{(k-1)^2N_i} \\ &= \frac{(k-1)p_{i1}(1-p_{i1})+p_{i1}\left[1-(k-1)p_{i1}\right]+2(k-1)p_{i1}^2}{(k-1)N_i} \\ &= \frac{\left[-(k-1)-(k-1)+2(k-1)\right]p_{i1}^2+\left[(k-1)+1\right]p_{i1}}{(k-1)N_i} \\ &= \frac{kp_{i1}}{(k-1)N_i}, \end{aligned}$$

which is exactly what we get from Section 1.8.4.

1.8.6 Expressed Using Shrinkage

We look one last time at FBI_i to gain a deeper understanding of how the two paradigms of successful blinding play in to the final value of the index. When looking at the bounds of the FBI_i statistic in Section 1.8.1.1, we made the claim that the index achieves its maximum (minimum) value only under scenarios where

nobody responds DK. Furthermore the presence of incorrect (correct) guesses or DK responses brings the index closer to zero.

To see why this is the case, rewrite the FBI_i index as

$$FBI_i = \frac{(k-1)n_{ii} - \sum_{j \neq i}^{k} n_{ij}}{(k-1)\sum_{j=1}^{k} n_{ij}} \times \frac{\sum_{j=1}^{k} n_{ik}}{N_i}. \tag{1.17}$$

The first factor in the product in expression (1.17) is the proportion "beyond random guessing" among subjects who did not respond DK. This factor measures the degree of random guessing, with completely random guessing represented numerically by zero. A simplistic approach would be to test this factor against the null hypothesis. However, if we were to only work with this first component we would be ignoring the DK responses, making the generalized approach comparable to approaches that predate James and Bang (see Section 1.3).

The second factor in the product in expression (1.17) yields insight in to precisely how we incorporate information from the DK responses. This term is the proportion of people who guessed a treatment group. Hence, as the proportion of DK responses increases, this term decreases so that the FBI_i statistic is shrunk toward zero. The factorization in expression (1.17) best shows how both paradigms contribute to the index. A high proportion of accurate guessing yields an index value close to one, but even in the presence of high accuracy among subjects who guessed, if there is a large proportion of DK responses then the index will not be as extreme (not as close to one). By a similar means, if there is a relative high degree of reverse unblinding among subjects who guessed the statistic will be negative, but not many subjects guessed a treatment, then there is less evidence of poor blinding and the statistic is made less negative by the shrinkage term. If there really is random guessing, the statistic will be around zero; the addition of DK responses are not as influential but FBI_i will still be brought closer to zero. It is easy to see how random guessing, or DK responses, or both, yield a desirable value for FBI_i .

1.8.7 Simultaneous Inference

For a study with k arms, we compute FBI_i, $i = 1, \ldots, k$, indices. For most clinical trials, the number of arms is not large. When testing the null hypothesis of random guessing, we are performing k simultaneous tests. The same applies to constructing simultaneous confidence sets. Thus there is a need to adjust for multiple testing. Because the tests are independent, we use the familywise error rate as defined below. For a global experiment significance level α, we have

$$\alpha = 1 - (1 - \alpha^*)^k,$$

where α^* is the significance level per comparison, and k is the number of independent comparisons. We in turn use this α^* value to obtain critical values to build confidence intervals around each of the FBI_i statistics. For example, if we have $k = 3$ arms and desire a 5% familywise error rate, then

$$\begin{aligned} \alpha^* &= 1 - \sqrt[k]{1-\alpha} \\ &= 1 - \sqrt[3]{1-0.05} \\ &= 1 - 0.983 \\ &= 0.017. \end{aligned} \tag{1.18}$$

Using normal theory, for two-sided confidence intervals, the critical value would be $cv = 2.388$. An alternative to using normal theory approximations is to use a computational approach such as the jackknife or a bootstrap James et al. (1996).

1.9 Investigation of the FBI_i Statistic via Simulation

There are two approaches to investingating FBI_i in simulation studies. First, we look at how the value of the generalized index changes with response to changes in the blinding scenarios and DK responses. Second, we compare FBI_i to the James BI, interested in how the two differ conditioned on the same simulated data sets. The first simulation does not look at the James index, focusing on the generalized index and its properties as discussed in previous sections.

1.9.1 Nine Settings - Simulation Design

To understand the FBI_i statistic, we simulate blinding survey responses for a single arm of a three-arm trial. It is unnecessary to simulate responses for all three arms when investigating FBI_i alone because the index is arm-specific. The value of FBI_i is dependent on two factors: a blinding scenario (*random*, *unblinded* and *opposite*), and DK response rate.

Concerning the three blinding scenarios, *random* means that all non-DK responses are equally likely, hence the subject is randomly guessing among all possibilities. For simulation purposes we define *unblinded* to mean that of non-DK responses, a subject guesses correctly with probability 0.8; thus *unblinded* means high guessing accuracy. The two incorrect options are chosen with probability 0.1, each. Finally, *opposite* guessing means that of non-DK responses, a subject is correct with probability 0.2. Under the *opposite guessing* blinding scenario, the two incorrect guesses are chosen with probability 0.4, each. Note that these definitions represent fixed probabilities for simulation purposes only, and do not mathematically define what it means to be unblinded or reverse unblinded (opposite guessing).

Table 1.8: Nine Settings

Blinding Scenario	DK Response Rate
random guessing (blinded)	0%
random guessing (blinded)	25%
random guessing (blinded)	70%
unblinded	0%
unblinded	25%
unblinded	70%
opposite guessing (reverse unblinded)	0%
opposite guessing (reverse unblinded)	25%
opposite guessing (reverse unblinded)	70%

The actual percentage of correct/incorrect responses depends on the percent of DK responses. The marginal probabilities defined in the paragraph above are valid for non-DK responses only. We consider three levels of DK response: 0% to represent an "extreme low" case, 25% to represent a moderate amount of DK responses, and 70% to represent a large DK response rate.

There are three blinding scenarios, each of which can be matched with one of three levels of DK response rate. Thus, there are nine distinct settings under consideration.

We have fixed the number of subjects per treatment arm to be $N_i = 200$. Rejection percentages are based on 95% two-sided confidence intervals. We ran 1000 iterations for each scenario. Interval estimates are not shown. However for each setting we provide the mean and standard error of the 1000 simulated values, along with the empirical rejection rate, empirical coverage probability, and the mean and standard error of the empirical interval width.

1.9.2 Nine Settings - Results

This empirical distribution of FBI_i for each of the nine settings is shown in Figure 1.1. This figure is a square matrix of nine frequency histograms, one for each of the nine settings considered in this section. The columns of the matrix represent the different levels of DK responses (0%, 25% and 70%, respectively), and the rows represent the three blinding scenarios (random, unblinded and opposite, respectively).

Below each histogram is the empirical mean (standard error). We see that as the DK rate increases, FBI_i tends toward zero. We have already shown this mathematically in Section 1.8.6. Table 1.9 gives additional descriptive statistics, including rejection percentage, coverage percentage, and mean (standard error) of the interval width. Note that a simultaneous inference correction was made (see Section 1.8.7) to show what the results might look like in a single arm of a three-arm trial.

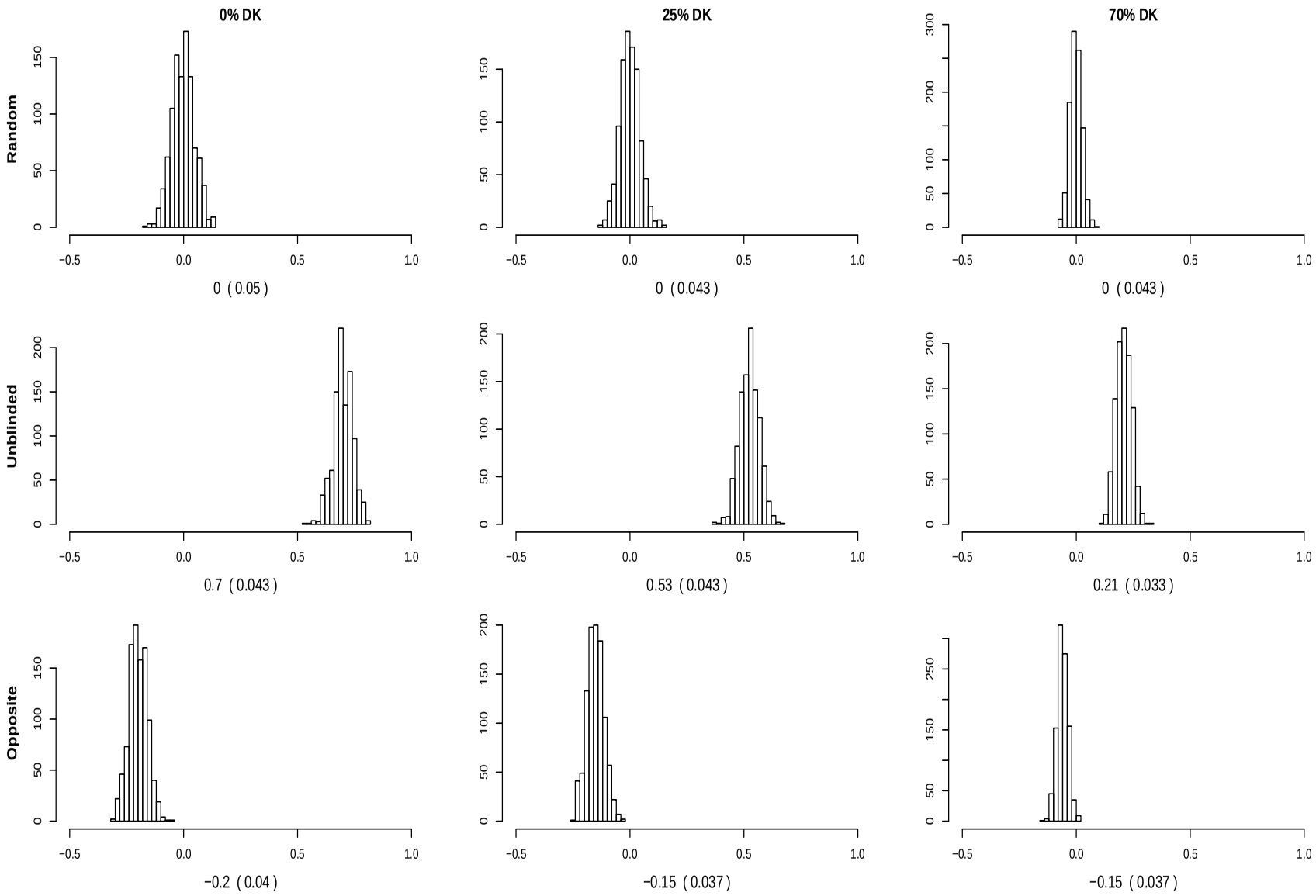

Figure 1.1: Histogram Matrix of FBI_i Distribution for $k = 3$ Arms

Coverage is consistently large, however the rejection percentage rapidly decreases as percent DK responses increases in the opposite guessing (reverse unblinded) group. Otherwise simulations demonstrate what we already know to be true: FBI_i gets closer to zero as DK responses increase (all else constant); the point estimate of FBI_i is positive in the unblinded group, negative in the opposite guessing group, and zero in the randomly guessing group; and the standard error of FBI_i decreases at DK response rate increases, all else constant.

Table 1.9: Simulation Results: Nine Settings

Scenario	DK (%)	FBI			
		FBI (SEE)	Rejection (%)	Coverage	Width (SEE)
Random	0	0.00 (0.051)	2.1	97.9	0.238 (0.006)
Unblinded		0.70 (0.044)	100	97.5	0.202 (0.011)
Opposite		−0.20 (0.042)	98	97.8	0.202 (0.011)
Random	25	0.00 (0.044)	2.2	97.8	0.207 (0.007)
Unblinded		0.052 (0.042)	100	96.1	0.183 (0.010)
Opposite		−0.15 (0.037)	92.2	97.1	0.176 (0.011)
Random	70	0.00 (0.027)	2	98	0.130 (0.010)
Unblinded		0.21 (0.033)	100	96.8	0.143 (0.006)
Opposite		−0.06 (0.024)	53.7	97.8	0.114 (0.010)

1.9.3 Sample Size

We have yet to see how sample size plays a role in the power of FBI_i. In general, small samples should not be of great concern because blinding is often measured at the end of Phase III trials. However, FBI_i can be applied to subgroups (for example, measuring blinding in men versus women or across demographics) which could results in smaller sample sizes. Therefore, we look at power across a full range of sample sizes from as small as ten subjects per arm up to 1000 per arm. Results are summarized in the power plots in Figures 1.2 and 1.3. Note that the null case is included, and

therefore for this case (random guessing) we are not presenting power but rather $1-\alpha^*$, the size. For the other two blinding scenarios, unblinding and opposite guessing, power is reported. We consider $N_i \in \{10, 50, 100, 200, 500, 750, 1000\}$ subjects per arm.

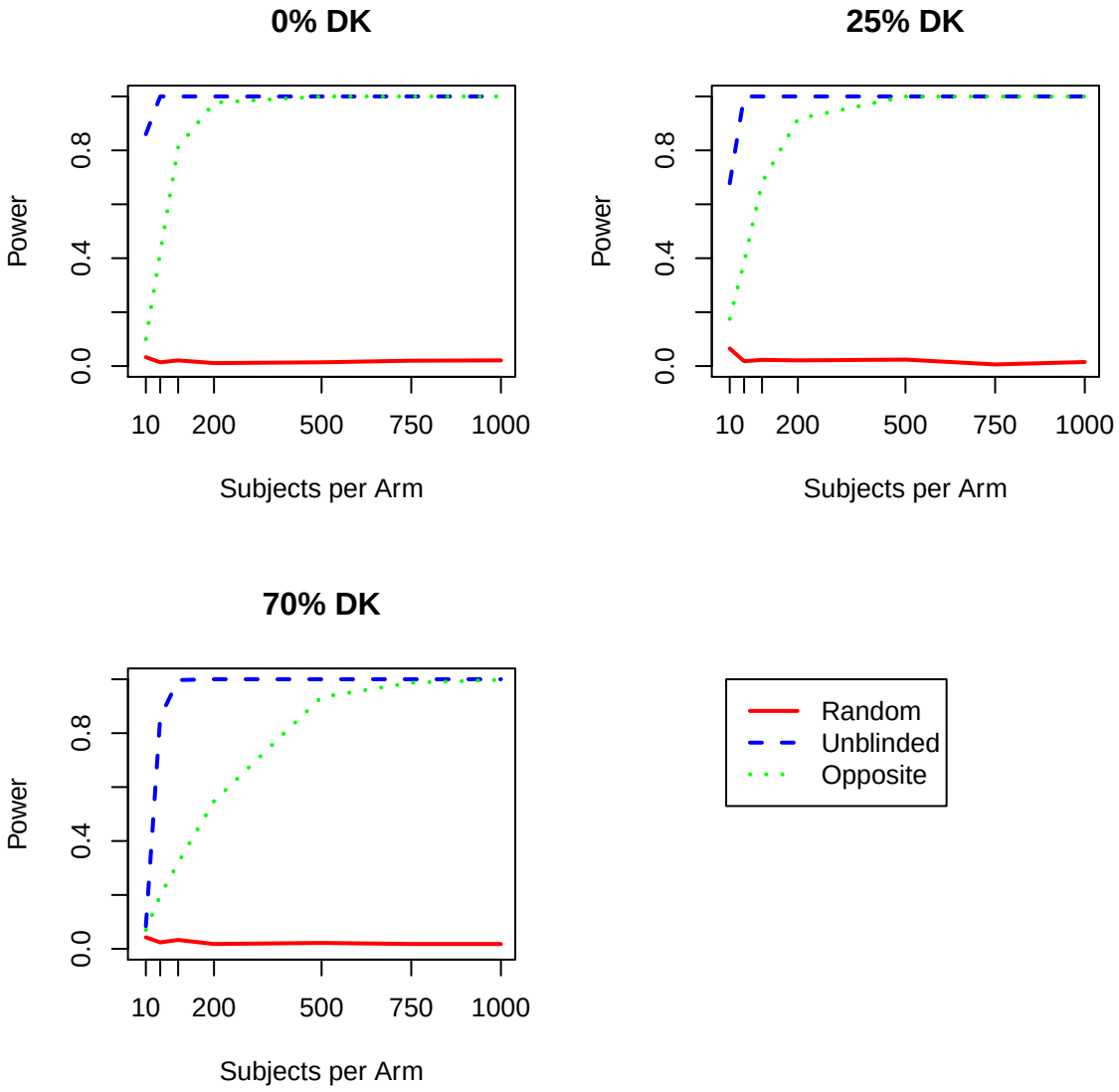

Figure 1.2: Power Study Comparing %DK Responses

Figure 1.2 shows empirical power plotted against sample size across the three DK levels considered from previous simulations (0%, 25% and 70%). We fix the number of treatment arms to be $k = 3$. In all three plots, the randomly guessing group displays large confidence even for small sample sizes. We see that FBI_i is

more powerful to detect unblinding than opposite guessing. For low to moderate DK responses, the power for unblinding is very large with as little as 50 subjects per arm, but with a large amount of DK responses we require about 100 subjects per arm to achieve the same power. Opposite guessing is more sensitive to the proportion of DK responses. A large proportion of DK responses requires upwards of 500 subjects per arm to achieve power close to one.

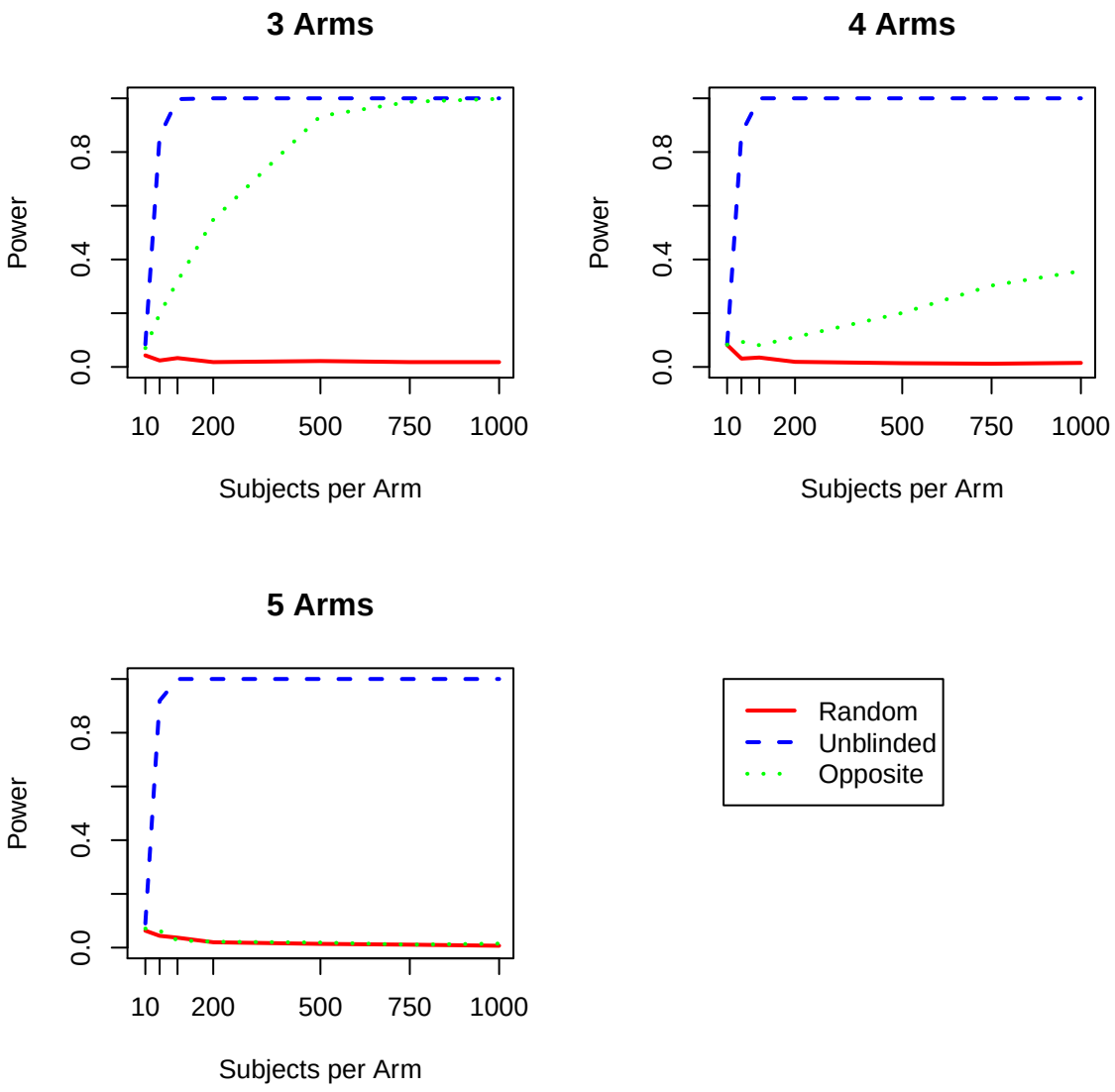

Figure 1.3: Power Study Comparing Number of Arms

Figure 1.3 illustrates an interesting story. We have fixed DK responses at 70%, because it yields the largest separation between groups (Figure 1.2). Now we look at

power plotted against sample size across number of arms. We investigate power for $k = 3$, 4 and 5 arms. Note the first plot of Figure 1.3 is the same as the third plot in Figure 1.2. We observe that for the random and unblinded arms, the relationship between power and sample size is preserved as we increase the number of treatment arms. However we notice a dissimilar trend for the opposite guessing group: as the number of arms increases the power is decreasing (for a fixed N). Even more surprisingly, for $k = 5$ arms, as sample size increases power decreases for the opposite guessing group. This is not a surprise, because the opposite guessing group should have a negative FBI_i index. As the number of groups increases, the index shrinks toward the null value of zero (we see this because the lower bound approaches zero as number of arms increases).

We make an interesting discovery about FBI_i: at some point there become too many treatment groups to detect reverse unblinding. How many treatment groups is too many? This is actually not a matter of percent DK responses but of how inaccurately this group guesses opposite to their true assignment. We used the same definitions for *unblinded* and *opposite guessing* as we did in previous simulations, unblinded subjects guess correctly 80% of the time they guess, and opposite guessing subjects guess correctly only 20% of the time they guess. To explain the loss in power, note that we have 20% guessing correctly in the opposite guessing arm, if they guess (not DK). Therefore the other 80% of non-DK guesses are divided among the remaining arms. So if we have $k = 5$ arms, the unblinded group is guessing 20% of the time in *each* of the five assignments. This is exactly the same as the *random guessing* group for $k = 5$ arms. Because there are so many other options, the difference between random guessing and opposite guessing becomes smaller and smaller. Note this is true for any amount of DK responses. For example, if we redefine *opposite* to mean subjects guess correctly only 10% of the time, then the distribution of FBI_i for *opposite guessing* group and the *randomly guessing* group are identical at $k = 10$ arms.

1.9.4 More Treatment Arms

Under all of the simulation work thus far, we have limited ourselves to three treatment arms. We continue to look at the nine settings proposed in Table 1.8.

Recall the histogram matrix from Section 1.9.2. We have made similar graphics for $k = 4$ and $k = 5$ arms (see Figures 1.4 and 1.5, respectively). To stay consistent with previous simulation work, we keep the definitions for random guessing, unblinded and opposite guessing; we assign 200 subjects per group; and results shown are based on 1000 simulations. One thing that is consistent across the graphics is that the FBI_i statistic always has a mean value of zero in the random group, as should be the case. For all nine cases, as the number of groups increases the standard error of FBI_i decreases. Because we fixed the percentage of correct non-DK responses in the unblinded group at 80%, and the percentage of incorrect non-DK responses in the opposite guessing group at 20%, it happens that as the number of groups increases the FBI_i statistics for the unblinded and opposite guessing groups increases as well. FBI_i is increasing in the unblinded group because we have a fixed 80% correct among non-DK responses, and in the calculation of the index the weight given to correct guesses is $k - 1$ and therefore increases with increasing k. The reverse unblinded (opposite guessing) group also shows an increase in FBI_i with increasing k because the lower bound of the statistic approaches zero as k increases (ref. Section 1.8.1.1). Note that for $k = 5$ the opposite guessing group and the randomly guessing group show the same distributions for FBI_i. An explanation for this is given in Section 1.9.3.

1.10 Simulation Study Comparing FBI_i to BI for Various Blinding Scenarios

We designed a simulation study to compare FBI_i to James's BI. We do not compare FBI_i to Bang's BI_i, because for the data to be conformable would require $k = 2$ arms. We proved for this case, the indexes are identical (ref. Section 1.8.1.2). Therefore, we repeat a similar process as Bang et al. to compare FBI_i and Jame's

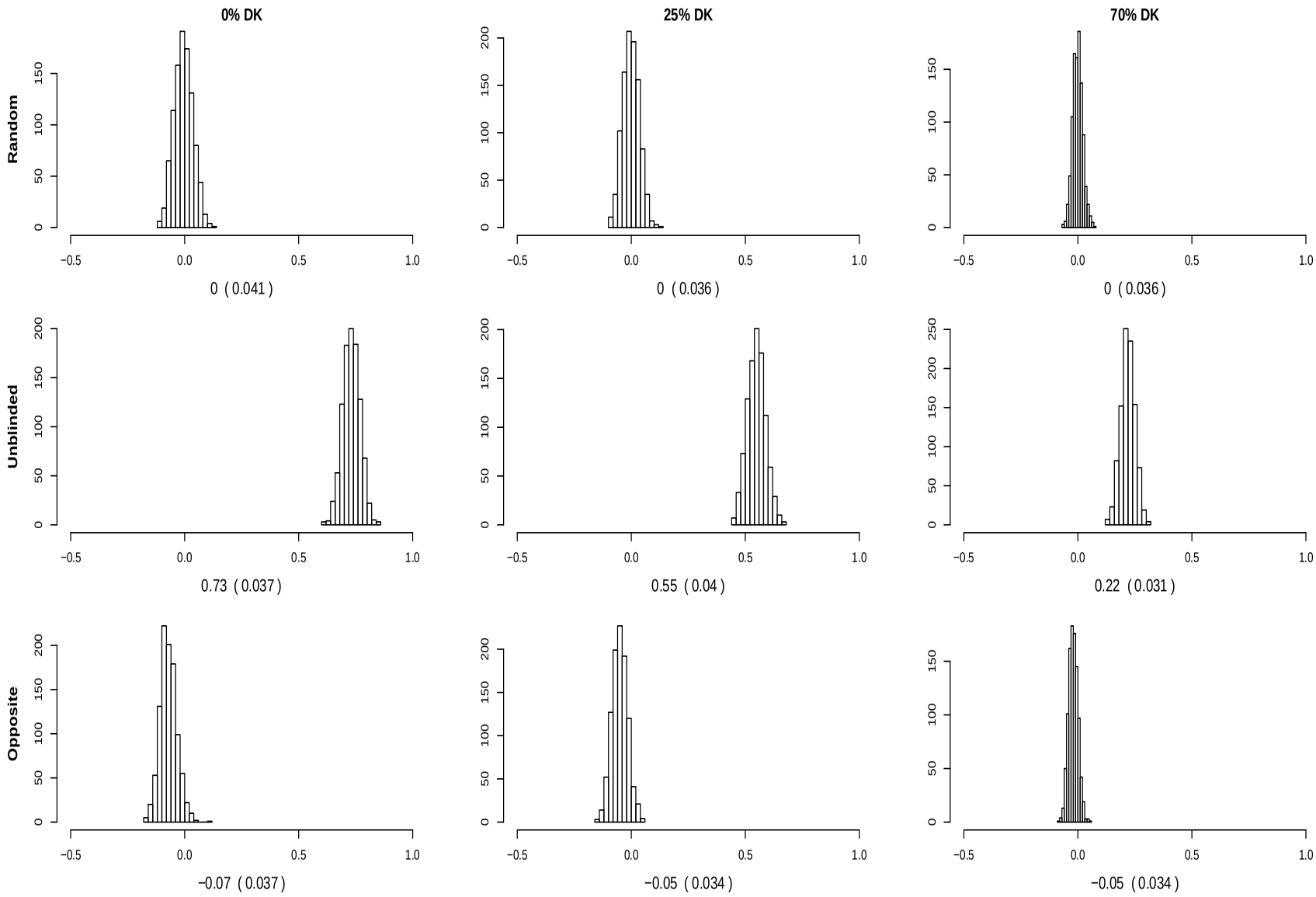

Figure 1.4: Histogram Matrix of FBI_i Distribution for $k = 4$ Arms

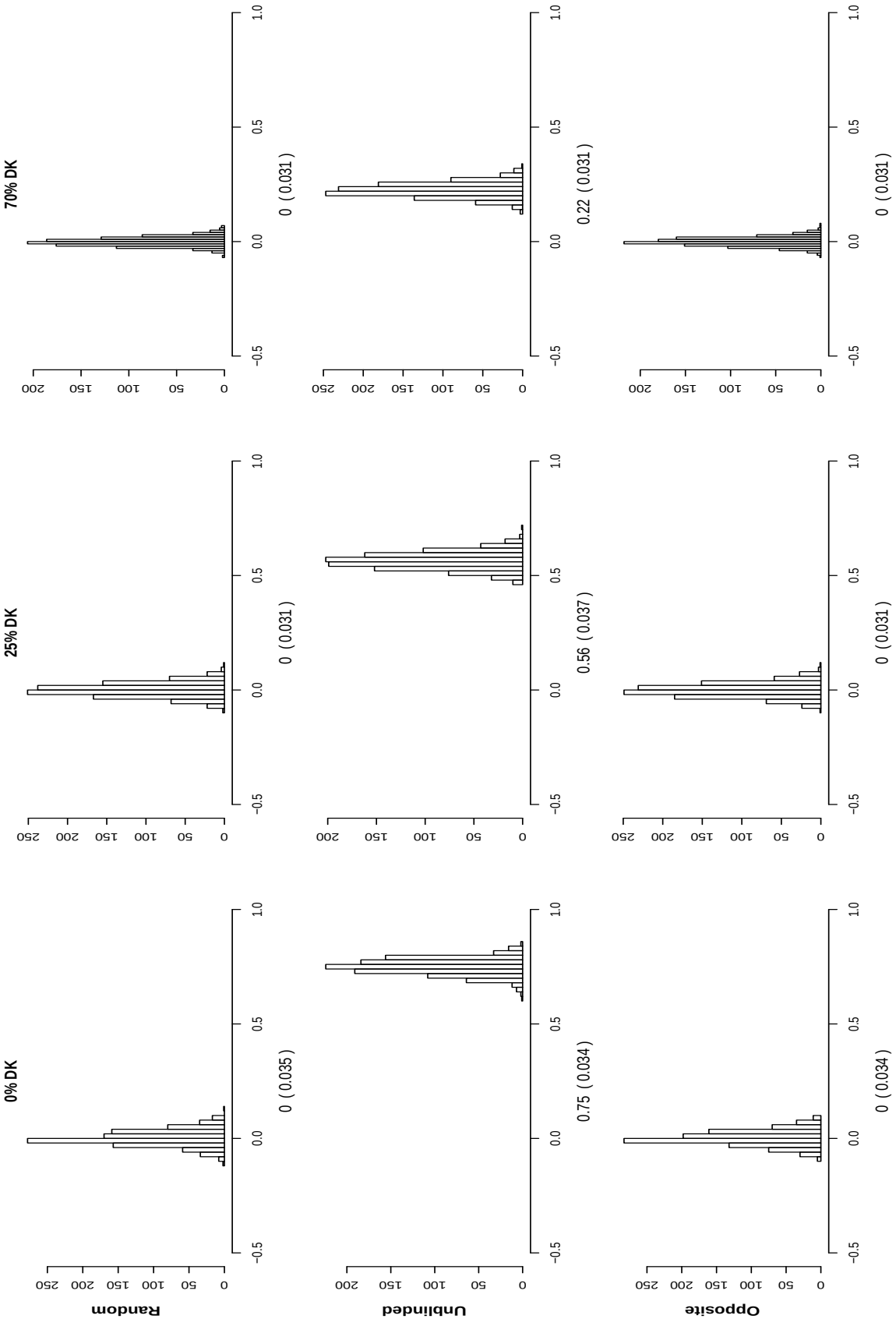

Figure 1.5: Histogram Matrix of FBI_i Distribution for $k = 5$ Arms

BI with special consideration given to $k > 2$ arms. We use the same combinations of blinding scenarios and DK response rate as described in Section 1.9.

1.10.1 Ten Cases of Blinding

Because the James index is aggregate for all treatment arms, there are more than just the nine distinct settings from Table 1.8. Table 1.10 describes ten cases for $k = 3$ arms. Each case is a unique combination of blinding scenarios and DK response rates for a three-arm trial.

We have fixed the number of subjects per treatment arm to be $N_i = 200$. Rejection percentages are based on 95% confidence intervals (one-sided for BI, and two-sided simultaneous intervals with a multiple comparison correction for FBI_i). We ran 1000 iterations for each scenario. We used the jackknife procedure described in James (1996) to obtain interval estimates, which in turn are used to compute rejection rates. Interval estimates are not reported. However for each occurrence of BI and FBI_i we give the empirical mean and standard error from the 1000 simulated values, along with the empirical rejection rate and empirical coverage probability (global coverage, the probability of all three indices capturing the true value in each case, is reported for FBI_i). Because the James index uses one-sided intervals we do not compute interval width. For information on interval width of the FBI_i indices, refer to Table 1.9.

Note that not all distinct scenarios are present in Table 1.10 for the James BI. However each of the nine distinct possibilities for FBI_i are represented at least once (ref. Section 1.9.2). We address other variations of the randomization assignments in Section 1.10.2.2. For the purpose of calculating BI, we let assignment A represent

Table 1.10: Simulation Design

Case	Assignment	Response (%)			
		DK	A	B	C
1	A	0	33.3	33.3	33.3
A: random	B		33.3	33.3	33.3
B: random	C		33.3	33.3	33.3
C: random	A	25	25	25	25
	B		25	25	25
	C		25	25	25
	A	70	10	10	10
	B		10	10	10
	C		10	10	10
2	A	0	33.3	33.3	33.3
A: random	B		33.3	33.3	33.3
B: random	C		10	10	80
C: unblinded	A	25	25	25	25
	B		25	25	25
	C		7.5	7.5	60
	A	70	10	10	10
	B		10	10	10
	C		3	3	24
3	A	0	33.3	33.3	33.3
A: random	B		33.3	33.3	33.3
B: random	C		40	40	20
C: opposite	A	25	25	25	25
	B		25	25	25
	C		30	30	15
	A	70	10	10	10
	B		10	10	10
	C		12	12	6
4	A	0	33.3	33.3	33.3
A: random	B		10	80	10
B: unblinded	C		10	10	80
C: unblinded	A	25	25	25	25
	B		7.5	60	7.5
	C		7.5	7.5	60
	A	70	10	10	10
	B		3	24	3
	C		3	3	24
5	A	0	33.3	33.3	33.3
A: random	B		40	20	40
B: opposite	C		40	40	20
C: opposite	A	25	25	25	25
	B		30	15	30
	C		30	30	15
	A	70	10	10	10
	B		12	6	12
	C		12	12	6

Table 1.10: Simulation Design (cont'd)

Case	Assignment	Response (%)			
		DK	A	B	C
6	A	0	33.3	33.3	33.3
A: random	B		10	80	10
B: unblinded	C		40	40	20
C: opposite	A	25	25	25	25
	B		7.5	60	7.5
	C		30	30	15
	A	70	10	10	10
	B		3	24	3
	C		12	12	6
7	A	0	80	10	10
A: unblinded	B		10	80	10
B: unblinded	C		10	10	80
C: unblinded	A	25	60	7.5	7.5
	B		7.5	60	7.5
	C		7.5	7.5	60
	A	70	24	3	3
	B		3	24	3
	C		3	3	24
8	A	0	80	10	10
A: unblinded	B		10	80	10
B: unblinded	C		40	40	20
C: opposite	A	25	60	7.5	7.5
	B		7.5	60	7.5
	C		30	30	15
	A	70	24	3	3
	B		3	24	3
	C		12	12	6
9	A	0	80	10	10
A: unblinded	B		40	20	40
B: opposite	C		40	40	20
C: opposite	A	25	60	7.5	7.5
	B		30	15	30
	C		30	30	15
	A	70	24	3	3
	B		12	6	12
	C		12	12	6
10	A	0	20	40	40
A: opposite	B		40	20	40
B: opposite	C		40	40	20
C: opposite	A	25	15	30	30
	B		30	15	30
	C		30	30	15
	A	70	6	12	12
	B		12	6	12
	C		12	12	6

the placebo group and use the weighting scheme described in Section 1.5.1. Results are presented in Table 1.11.

Looking at the results in Table 1.11, we notice a few trends. First, for each of the cases, as the proportion of DK responses increases the James index also increases. This is not unexpected, we have already said that James's index is dominated by the amount of DK responses. Thus, no matter the blinding situation, if more people answer DK, then there is more evidence for blinding (see specific case descriptions below for information on how extreme this relationship is). Next, as the percent of DK responses increases we see that the FBI_i index shrinks toward zero, as it should. For arms guessing at random, the value of FBI_i is already zero, so the only change is in the standard error which gets smaller as the DK response rate increases. Additionally, we see that James BI is not powerful in detecting unblinding in the presence of a large percent of DK responses. Each case is considered individually in the proceeding text.

Case 1: Case 1 represents the null hypothesis of "random guessing" for all three arms under each of the three DK response rates. The null value for BI is one-half (ignoring DK responses) and is zero in each arm for FBI_i. When DK responses are present, we see that the FBI_i is not affected and remains zero. However, we know that BI increases with the proportion of DK responses. The empirical rejection rate of BI is 4.7%, which is close to the advertised significance level of 5%. For the FBI_i indexes, we see they are all centered at zero. As the percentage of DK responses increases, the standard errors decrease. From expression (1.18) we know the empirical rejection percentages for each of the FBI_i should be approximately 1.7%.

Case 2: Case 2 is similar to Case 1, but with arm C unblinded. The James index detects the unblinding in the absence of DK responses. This suggests the study as a whole is not adequately blinded when really the problem is coming from only one of the three arms. However, under moderate and high DK

response rates the unblinded arm does not carry enough weight to suggest poor overall blinding. The FBI_i index for the two blinded arms is the same is at was in the previous case whereas we see a positive value of FBI_3, the unblinded group. The FBI_i index shrinks toward zero as the DK response rate increases, but the empirical rejection rate is not compromised.

Case 3: Rather than having one unblinded arm as in Case 2, Case 3 introduces an opposite guessing (reverse unblinded) arm in its place, with the other two arms still randomly guessing. James BI does not detect opposite guessing, but rather uses it as evidence in support of blinding. We see this in that the James blinding index values are consistently larger in this case than in Case 1. The small (zero) empirical rejection rate for no DK responses also demonstrated how the James index confuses reverse unblinding with blinding. We see that the FBI_i index has large power to detect reverse unblinding (opposite guessing) for no to moderate DK responses. It is more difficult for FBI_i to detect reverse unblinding when there is a large DK response rate because the FBI_i index is approaching the null value, zero.

Case 4: Case 4 consists of one blinded (random guessing) arm and two unblinded arms. FBI_i clearly shows which arm follows which blinding scenario. However, even though two of the three arms are unblinded, the power of James drops quickly as DK responses increase. For a large proportion of DK responses, James has zero power to detect unblinding. This serves as a reminder that the James index was not made to measure the degree of unblinding. The FBI_i index clearly rejects that groups B and C are blinded. But how does this work when 70% of respondents don't know? Along with random guessing, DK responses are also supposed to be evidence of blinding. At this point it seems the Bang paradigm is dominated by equality of guessing among non-DK groups. However the interpretation of FBI_i holds an advantage over James's BI . For 0% and 70% DK responses we reject random guessing in

arms B and C using FBI_i. The difference between these is that with 0% DK responses subjects are guessing 70% beyond random chance ($\widehat{FBI_i} = 0.70$), and at 70% DK responses subjects are guessing 21% beyond random chance ($\widehat{FBI_i} = 0.21$). There is a *degree of severity* of unblinding. So, despite the large proportion of DK responses, FBI_i still detects the imbalance of guesses in these arms. James's index cannot do this. Later we discuss the advantage of using the indexes together, but for now note that this is a good example of when it would be beneficial to consider both indexes for the same data set.

Case 5: Case 5 has one randomly guessing arm and two opposite guessing arms. For James, opposite guessing is evidence of blinding. Therefore, there is not much difference from Case 3 for the James index. The FBI_i, on the other hand, show a clear difference between these two blinding scenarios.

Case 6: One of each blinding scenario (random, unblinded and opposite guessing) is present in this case. The FBI_i approach has a clear advantage here, as it can detect all three scenarios. With James we conclude either "well blinded" or "not well blinded." Unsurprisingly, James's conclusion falls in line with the amount of DK responses because the hodgepodge of blinding schemes allows BI to bounce between conclusions easily.

Case 7: From here on out there are no more randomly guessing arms. Case 7 in particular has all unblinded arms. For zero and moderate DK responses, James picks up on the unblinded nature of the groups. However, even though all arms are unblinded, for large DK responses James concludes that the study as a whole is adequately blinded. On the other extreme, FBI_i always rejects the null hypothesis of random guessing. Mathematically this is correct, but again we have the issue where it seems the DKs are being ignored. If we look not only at the hypothesis test but also at the value of the FBI_i index itself, we see the difference in FBI_i for the different levels of DK responses. We

emphasize again the importance of the point estimate of FBI_i, not only the statistical significance.

Table 1.11: Simulation Results

Case	DK (%)	James' BI		FBI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
1	0	0.50 (0.015)	4.7	A	0.00 (0.052)	1.9
				B	0.00 (0.050)	1.3
				C	0.00 (0.050)	2
	25	0.62 (0.015)	0	A	0.00 (0.045)	2.6
				B	0.00 (0.043)	2.1
				C	0.00 (0.043)	1.8
	70	0.85 (0.012)	0	A	0.00 (0.027)	1.6
				B	0.00 (0.027)	1
				C	0.00 (0.027)	2.1
2	0	0.40 (0.015)	100	A	0.00 (0.050)	1.7
				B	0.00 (0.050)	1.8
				C	0.70 (0.044)	100
	25	0.55 (0.017)	0.4	A	0.00 (0.044)	2.3
				B	0.00 (0.043)	1.7
				C	0.52 (0.043)	100
	70	0.82 (0.014)	0	A	0.00 (0.027)	1.8
				B	0.00 (0.027)	1.9
				C	0.21 (0.032)	100
3	0	0.54 (0.014)	0	A	0.00 (0.051)	1.8
				B	0.00 (0.050)	1.6
				C	-0.20 (0.041)	98.8
	25	0.66 (0.014)	0	A	0.00 (0.042)	1.6
				B	0.00 (0.042)	1
				C	-0.15 (0.037)	93.9
	70	0.86 (0.011)	0	A	0.00 (0.027)	2
				B	0.00 (0.027)	1.4
				C	-0.06 (0.025)	53
4	0	0.28 (0.013)	100	A	0.00 (0.048)	1.9
				B	0.70 (0.043)	100
				C	0.70 (0.044)	100
	25	0.46 (0.018)	77.4	A	0.00 (0.042)	1.7
				B	0.53 (0.043)	100
				C	0.53 (0.044)	100
	70	0.78 (0.016)	0	A	0.00 (0.028)	1.8
				B	0.21 (0.032)	100
				C	0.21 (0.031)	100
5	0	0.57 (0.013)	0	A	0.00 (0.050)	1.5
				B	-0.20 (0.042)	98
				C	-0.20 (0.042)	97.5
	25	0.68 (0.014)	0	A	0.00 (0.043)	1.1
				B	-0.15 (0.038)	92.5
				C	-0.15 (0.038)	93.1
	70	0.87 (0.011)	0	A	0.00 (0.027)	2
				B	-0.06 (0.024)	55.3
				C	-0.06 (0.024)	52.1

Table 1.11: Simulation Results (cont'd)

Case	DK (%)	James' BI		FBI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
6	0	0.41 (0.014)	100	A	0.00 (0.051)	2.1
				B	0.70 (0.044)	100
				C	-0.20 (0.042)	98
	25	0.56 (0.016)	0	A	0.00 (0.044)	2.2
				B	0.52 (0.042)	100
				C	-0.15 (0.037)	92.2
	70	0.82 (0.014)	0	A	0.00 (0.027)	2
				B	0.21 (0.033)	100
				C	-0.06 (0.024)	53.7
7	0	0.15 (0.012)	100	A	0.70 (0.042)	100
				B	0.70 (0.043)	100
				C	0.70 (0.042)	100
	25	0.36 (0.018)	100	A	0.52 (0.042)	100
				B	0.53 (0.044)	100
				C	0.52 (0.043)	100
	70	0.74 (0.016)	0	A	0.21 (0.032)	100
				B	0.21 (0.032)	100
				C	0.21 (0.032)	100
8	0	0.29 (0.012)	100	A	0.70 (0.042)	100
				B	0.70 (0.042)	100
				C	-0.20 (0.041)	98.4
	25	0.46 (0.017)	67.7	A	0.53 (0.043)	100
				B	0.53 (0.044)	100
				C	-0.15 (0.038)	92.2
	70	0.80 (0.015)	0	A	0.21 (0.033)	100
				B	0.21 (0.034)	100
				C	-0.06 (0.024)	51.6
9	0	0.44 (0.013)	99.7	A	0.70 (0.044)	100
				B	-0.20 (0.043)	97.6
				C	-0.20 (0.042)	97.4
	25	0.60 (0.014)	0	A	0.52 (0.042)	100
				B	-0.15 (0.037)	93.5
				C	-0.15 (0.036)	94.2
	70	0.84 (0.012)	0	A	0.21 (0.033)	100
				B	-0.06 (0.025)	55
				C	-0.06 (0.024)	53.9
10	0	0.60 (0.013)	0	A	-0.20 (0.043)	97.6
				B	-0.20 (0.043)	97.7
				C	-0.20 (0.042)	97.6
	25	0.70 (0.013)	0	A	-0.15 (0.037)	93.4
				B	-0.15 (0.037)	91.7
				C	-0.15 (0.036)	93.1
	70	0.88 (0.010)	0	A	-0.06 (0.024)	54.3
				B	-0.06 (0.024)	56.3
				C	-0.06 (0.025)	56.5

Case 8: Case 8 consists of two unblinded arms and one reverse unblinded arm. Reverse unblinding falls under the category of blinding support for James, so this is very similar to Case 7, but now the James BI doesn't hold out as

long to support blinding. With increasing DK responses the value of BI approaches one faster than in the previous case.

Case 9: Case 9 consists of two reverse unblinded arms and one unblinded arm. For moderate and large DK response rates, the James BI suggests blinding is achieved, although none of the arms are unblinded. Further evidence that BI is dominated by the DK response rate.

Case 10: In this last case, all three arms are reverse unblinded. James has no power to detect this event and uses it as support for blinding. The FBI_i indexes behave as we have already seen for other opposite guessing arms in the previous cases.

In considering each case individually, we have seen some examples where BI and FBI_i agree and others where they disagree. Most importantly we are beginning to see that there is an advantage to considering both of the indexes together in order to obtain a better picture of blinding success. We talk more about the practical use of FBI_i in Chapter 3, a guidance for blinding indexes. The ideas in Chapter 3 are similar to those expressed by Bang Bang et al. (2010) and Park Park et al. (2008).

1.10.2 Limitations to the Bang Paradigm

For FBI_i, it does not matter which of the treatment groups (A, B or C) represents the placebo. Take for example Case 2 from Table 1.10. We have assignments A and B guessing randomly and assignment C unblinded. For the FBI_i statistic, this is the same as any other permutation of two randomly guessing arms and one unblinded arm because we measure blinding independently in each arm. This is not the case for the James index, it makes a difference if the unblinded arm is the placebo arm or not.

The James index does not treat all incorrect guesses equally (recall the weight structure in Section 1.3.1). For a subject in a non-placebo group, the weight of their guess depends on the guess itself – guessing placebo is weighted more than guessing another treatment. However, FBI_i assigns equal weight to each incorrect

guess within an arm. There are some concerns that arise when using the weight structure, for example

(1) What happens when the blinding scenarios are permuted across trial arms from the simulations in Table 1.10?

(2) What happens when the incorrect guesses are not split evenly among incorrect responses?

Mathematically the James BI changes with concerns (1) and (2), but the FBI_i does not. Bang did not have to address this issue because there are not different types of incorrect guesses for $k = 2$ arms. However it could be of concern when working with the more generalized FBI_i.

1.10.2.1 Does a Weighted Index Make Sense? Can we create a *weighted* index that accounts for the concerns above? First, we are able to write the FBI_i index in a weighted form,

$$FBI_i = \frac{w_{ii}n_{ii} + \sum_{j\neq i}^{k+1} w_{ij}n_{ij}}{(k-1)N_i},$$

where the weights, w_{ij}, are as follows,

$$\begin{cases} w_{ii} = k-1 \\ w_{ij} = -1 \qquad \text{for } j \in \{1,\ldots,k\}\backslash\{i\} \\ w_{i,k+1} = 0 \end{cases}$$

With the weight structure above, we get back the FBI_i index. However, we could try to change the weighting scheme to create a new index for which the general form is

$$WFBI_i = \frac{w_{ii}n_{ii} + \sum_{j\neq i}^{k+1} w_{ij}n_{ij}}{(k-1)N_i}$$

for any general weight structure. Consider a trial with three arms. We could calculate $WFBI_i$ for any choice of weight matrix $\mathbf{W} = [w_{ij}]$, $i = 1,\ldots k$, $j = 1,\ldots,k+1$. Consider two possible choices for $\mathbf{W}$, given in Table 1.12.

Minor challenges arose when trying to determine the weights for $WFBI_i$ that parallel the weight scheme used by James. Firstl, the James null value is 0.5 whereas

for FBI_i it is zero. For James BI, values close to one represent successful blinding whereas it is the opposite for FBI_i. These basic differences must be accounted for in obtaining a comparable weighting scheme. Taking into account that positive values of FBI_i indicate *unblinding*, the largest weight is given to the unblinding scenarios.

Table 1.12: Weight Choices for $WFBI_i$

(a) FBI_i Weights

Assignment	Guess			
	Placebo	Treatment B	Treatment C	DK
Placebo	2	−1	−1	0
Treatment B	−1	2	−1	0
Treatment C	−1	−1	2	0

(b) James BI Weight Equivalence

Assignment	Guess			
	Placebo	Treatment B	Treatment C	DK
Placebo	w_{ii}	−0.5	−0.5	0
Treatment B	−0.5	w_{ii}	−0.75	0
Treatment C	−0.5	−0.75	w_{ii}	0

One approach would be to place weight only on incorrect responses and not change anything about correct responses. Consider the null value of $WFBI_i$ under the weighting scheme in Table 1.12b, where $w_{ii} = k - 1 = 2$ (same as FBI_i weight). For a placebo group,

$$\begin{aligned} WFBI_{1\,(H_0)} &= \frac{2p_{11} - 0.5p_{12} - 0.5p_{13}}{2} \\ &= \frac{1}{2} \qquad \text{if } H_0 : p_{11} = p_{12} = p_{13}. \end{aligned}$$

For a non-placebo group (say group B, WLG),

$$\begin{aligned} WFBI_{2\,(H_0)} &= \frac{2p_{22} - 0.5p_{21} - 0.75p_{23}}{2} \\ &= \frac{3}{8} \qquad \text{if } H_0 : p_{21} = p_{22} = p_{23}. \end{aligned}$$

There are three problems here. First, the null values of p_{ij} are not preserved so that we are no longer comparing $WFBI_i$ against zero for "random guessing." Second, the null value of the index is now dependent on the number of groups *and* which arm is being measured. The third problem, not addressed here, is that the lower bound of the index is not constant for fixed k. The lower bound of FBI_i depends on the number of arms, k, and therefore is constant for all FBI_i from the same study. For $WFBI_i$ it is easy to see that when there are no correct guesses, the lower bound depends on both k and $\mathbf{W}$.

Perhaps some of the above concerns are amenable. When developing FBI_i, the weight of correct guesses was chosen so that the index has a value of zero under the null hypothesis. Is it possible to choose weights such that the expacted balue of $FBI_i is zero under the null hypothesis? In the example above, let \mathrm{w}_{11} = 1$ and $w_{22} = w_{33} = 1.25$. The indexes are recentered at zero. An obvious result is that in this choice of weights, we are stating that a correct guess is not equal for all groups. In this example, a correct guess is more valuable if you're not on the placebo than if you are, as it is given more weight. Philosophically this does not make sense in the context of this problem. But even if we were to accept this, the bounds of the index are still not constant between arms, for fixed k.

A final concern with choosing weights to center the index is that the interpretation of the FBI_i is no longer valid. If we were to introduce uneven weights among incorrect guesses, the index would no longer be measuring the *proportion guessing correctly beyond random balance*, because we would be giving equally probable events unequal weights. For all of the reasons addressed in this section, we conclude that a weighted index does not naturally follow under the Bang paradigm to measuring blinding. We would have to alter our approach to measuring unblinding to be able

to add weights to the responses. Bang does propose an alternate BI_i statistic which uses a weighting scheme for responses, but the weights are self-reported confidence of guesses and do not discriminate against different types of incorrect answers. Also, because Bang only considers $k = 2$ arms, the weights are not a complication as they are in the generalized case.

1.10.2.2 Are We Really Missing Out? We have learned that a weighted index for our approach is not feasible. How much information is being lost by not adding the weight feature? We have designed a simulation to consider the two scenarios listed above: permutation of groups guessing incorrectly, and uneven distribution of incorrect responses. We apply each of these changes to one of the cases we considered in the simulation from Section 1.10.1.

Recall Case 2 from Table 1.10. In this scenario, we have two arms randomly guessing and one arm unblinded. This scenario is reproduced in Table 1.13, along with variations of this scenario which address the two concerns mentioned above. Recall that for purposes of computing the James index, arm A represents the placebo group. The cases investigated in this simulation are as follows:

- 2a – Case 2 from Table 1.10;
- 2b – rather than the unblinded group being a treatment group, we make it so that the unblinded group is the placebo group (no change in the allocation of incorrect guesses);
- 2c – same unblinding scenario as 2a, but with all incorrect guesses being the other treatment group; and
- 2d – same unblinding scenario as 2a, but with all incorrect guesses being the placebo group.

Table 1.13: Special Examination Scenarios

Case	Assignment	Response (%)			
		DK	A	B	C
2a	A	0	33.3	33.3	33.3
A: random	B		33.3	33.3	33.3
B: random	C		10	10	80
C: unblinded with	A	25	25	25	25
incorrect guesses	B		25	25	25
split evenly	C		7.5	7.5	60
between A and B	A	70	10	10	10
	B		10	10	10
	C		3	3	24
2b	A	0	80	10	10
A: unblinded with	B		33.3	33.3	33.3
incorrect guesses	C		33.3	33.3	33.3
split evenly	A	25	60	7.5	7.5
between B and C	B		25	25	25
B: random	C		25	25	25
C: random	A	70	24	3	3
	B		10	10	10
	C		10	10	10
2c	A	0	33.3	33.3	33.3
A: random	B		33.3	33.3	33.3
B: random	C		0	20	80
C: unblinded with	A	25	25	25	25
all incorrect	B		25	25	25
guesses as B	C		0	15	60
(extreme case)	A	70	10	10	10
	B		10	10	10
	C		0	6	24
2d	A	0	33.3	33.3	33.3
A: random	B		33.3	33.3	33.3
B: random	C		20	0	80
C: unblinded with	A	25	25	25	25
all incorrect	B		25	25	25
guesses as A	C		15	0	60
(extreme case)	A	70	10	10	10
	B		10	10	10
	C		6	0	24

To determine the effect the weighs have to detect differences in blinding schemes, Cases 2b–2d are compared to Case 2a which is simply Case 2 from previous simulation work. The value of FBI_i should not change between these four cases, but due to its weighting scheme BI may change. Case 2b is the case where we permute which arm is unblinded. Note that having arm B as the unblinded arm is equivalent to

case 2b where we flip arms B and C. Because B and C are both treatment arms (not placebo), there would be no difference in the James index. Cases 2c and 2d are extreme examples of for when incorrect guesses are not split equally among the incorrect responses. Case 2c forces all incorrect guesses to be for the other treatment (which counts more towards unblinding when the guesser is in a treatment arm), and Case 2d forces all incorrect guesses to be for the placebo arm (which counts less towards unblinding when the guesser is in a treatment arm). Case 2a is a balance of these two, having half guess the placebo and half guess the other treatment. The extent to which the James index changes from Case 2a under each of the other three cases (2b–2d) gives an indication of how much more insight the James BI is providing over FBI_i.

As with before, 200 subjects are allocated to each treatment arm and the results are based on 1000 replications. The results are summarized in Table 1.14, and the values for Case 2a in Table 1.14 are identical to the results of Case 2 in Table 1.11. To compare the FBI_i index with the James BI for Case 2a, read the description in the previous section. The FBI_i index does not change between cases here, so we do not re-summarize.

Case 2a: In brief, The James BI for Case 2a increases as the percent DK responses increases, which is always true for the James BI (all else constant). For no DK responses, we find that the one group being unblinded is sufficient for the study overall to be deemed unsuccessful in preserving the blind via James BI. In fact, it is strong enough that H_0 is rejected 100% of the time. However, for 25% and 70% DK responses we never rejected H_0. This again supports that the James BI is dominated by DK responses.

Case 2b: We have made the placebo arm unblinded and both treatment arms blinded (randomly guessing). The value of the James index decreases only slightly under this scenario – 0.022, 0.016 and 0.006 for 0%, 25% and 70% DK responses, respectively. These changes did not yield any difference in rejection percent-

age. We still reject 100% of the time when there are no DK responses, and 0% of the time for 25% and 70% DK responses. Thus, there is no difference in the conclusions.

Case 2c: Arm C is again the unblinded group, with the others randomly guessing. The difference here is that all incorrect guesses among respondents assigned to arm C are for the other treatment group, B. According to James, this is more evidence for unblinding than incorrectly guessing the placebo group, therefore we expect a larger value of BI than in Case 2a. Indeed this is the case. The value of the James index increases by 0.008, 0.005 and 0.003 for 0%, 25% and 70% DK responses, respectively. Again, there are no differences in the conclusion we would make about either of these scenarios as it relates to blinding.

Case 2d: This extreme case is the complement to Case 2c – now all incorrect guesses among respondents assigned to C are for the placebo group, A. James claims this is not as strong of evidence for unblinding, and therefore we expect the James index values to decrease. This does happen for all three values: decreases of 0.007, 0.005 and 0.002 for 0%, 25% and 70% DK responses, respectively. As with the other cases, there is no change in rejection percentage. All conclusions based on the hypothesis test are the same.

Although the value of the BI index does change under the different scenarios above, none of the scenarios lead to a change in conclusion for any of the 1000 replications. Also, because the James index does not have a meaningful interpretation, the small changes in BI are meaningless. Any change in FBI_i represents a change in accuracy beyond random balance (recall the interpretation of the index). James does not offer such an interpretation, and therefore $BI = 0.813$ versus $BI = 0.817$ has no practical difference, because there are several ways in which BI increases and thus the index value alone does not tell us why it increases.

Table 1.14: Special Examination Results

Case	DK (%)	James' BI		FBI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
2a	0	0.396 (0.015)	100	A	0.00 (0.050)	1.7
				B	0.00 (0.050)	1.8
				C	0.70 (0.044)	100
	25	0.547 (0.017)	0	A	0.00 (0.044)	2.3
				B	0.00 (0.043)	1.7
				C	0.52 (0.043)	100
	70	0.819 (0.014)	0	A	0.00 (0.027)	1.8
				B	0.00 (0.027)	1.9
				C	0.21 (0.032)	100
2b	0	0.374 (0.014)	100	A	0.70 (0.041)	100
				B	0.00 (0.048)	1.6
				C	0.00 (0.048)	1.3
	25	0.531 (0.016)	0	A	0.53 (0.040)	100
				B	0.00 (0.042)	1.1
				C	0.00 (0.045)	2.4
	70	0.813 (0.014)	0	A	0.21 (0.032)	100
				B	0.00 (0.026)	1.1
				C	0.00 (0.027)	1.6
2c	0	0.403 (0.014)	100	A	0.00 (0.051)	2.3
				B	0.00 (0.050)	1.6
				C	0.70 (0.041)	100
	25	0.552 (0.017)	0	A	0.00 (0.045)	2.8
				B	0.00 (0.043)	2.3
				C	0.52 (0.043)	100
	70	0.822 (0.014)	0	A	0.00 (0.026)	2.1
				B	0.00 (0.026)	1.4
				C	0.21 (0.033)	100
2d	0	0.389 (0.014)	100	A	0.00 (0.051)	2.3
				B	0.00 (0.050)	1.6
				C	0.70 (0.041)	100
	25	0.542 (0.016)	0	A	0.00 (0.045)	2.8
				B	0.00 (0.043)	2.3
				C	0.52 (0.043)	100
	70	0.817 (0.014)	0	A	0.00 (0.026)	2.1
				B	0.00 (0.026)	1.4
				C	0.21 (0.033)	100

Therefore, although the BI can measure special situations that are mathematically equal under FBI_i, the conclusion does not differ. James shows that even if we change the choice of weights, the value of BI does not change considerably. This is further evidence of the importance of DK responses for BI. We do not feel as if FBI_i is at a disadvantage to James by not detecting these differences. James is

not powerful to detect these differences, therefore these special situations are not a criteria on which to judge the competing indexes.

1.11 Application

We bring back the James application from Section 1.5.2 of disulfiram to treat alcoholism. The data is re-summarized in Table 1.15 in the $k \times (k+1)$ format structure we use for the BI_i and FBI_i indexes. Recall that for this data, the James blinding index had a value of 0.556 with 95% jackknife confidence limits 0.520 and 0.592. James concluded this is "a response pattern close to that expected by random guessing." As we have already mentioned, the value of the index itself has no interpretation. So even though the value is close to the null value, one-half, we can only judge it by it being significantly greater than one-half and thus must conclude there is a degree of blinding (in this case, blinding by random guessing).

Table 1.15: VA Coop Study No. 107 Study Coordinator Responses

Actual Treatment	Guessed Treatment				
	Riboflavin	Disulfiram 1mg	Disulfiram 250mg	DK	Total
Riboflavin	64	22	36	52	174
Disulfiram 1mg	30	41	66	44	181
Disulfiram 250mg	24	27	72	51	174
Total	118	90	174	147	529

To demonstrate how the FBI_i index can be used in conjunction with the James index, we apply the FBI_i to the same set of data. Table 1.16 gives FBI_i point estimates and simultaneous 95% confidence intervals for the data in Table 1.15. For subjects assigned to the placebo group, study coordinators were able to guess assignment allocation accurately 20% beyond random guessing ($\widehat{FBI}_i = 0.20$). We see that this result is statistically significant at the 5% global level. Study coordinators

were not successful, however, at identifying subjects in the 1mg disulfiram group ($\widehat{FBI_i} = -0.04$, not significant). Finally, subjects assigned to the 250mg disulfiram group were correctly identified by study coordinators 27% beyond random chance ($\widehat{FBI_i} = 0.27$). This result is also statistically significant.

Recall that the James index is dominated by DK responses. For the VA Cooperative Study No. 107, approximately 28% of respondents answered "DK". Thus it is not surprising that the index was significantly greater than one-half, indicating successful blinding beyond random guessing. However, when we look at the results by treatment arm, the *only* arm that was not identified beyond random chance by coordinators was the 1mg disulfiram group. Study coordinators were able to identify the correct treatment for patients assigned to the riboflavin (placebo) and 250mg disulfiram groups beyond random chance. However, does this necessarily indicate poor blinding of study coordinators for these two groups? We already know about the influence of the DK responses on the James BI. But we have also shown via simulation that, even with a large amount of DK respondents, the FBI_i index is powerful for detecting non-random guessing. Looking back to the motivation for this index, we pointed out that both DK responses and a lack of random guessing should contribute to the final determination of blinding success.

Table 1.16: VA Coop Study No. 107 measured using FBI_i

Treatment Arm	FBI_i	Confidence Interval
Riboflavin (placebo)	0.20	$(0.086, 0.315)$
Disulfiram (1mg)	-0.04	$(-0.145, 0.068)$
Disulfiram (250mg)	0.27	$(0.153, 0.381)$

Consider only the placebo group. Responders were accurate 20% beyond random guessing, as indicated by the point estimate of the FBI_i statistic. According to the FBI_i index, this is a degree of unblinding. But note that for this treatment group nearly 30% of responses were "DK". This should carry some weight. If we were to

compute FBI_i ignoring DK responses, the riboflavin arm would have had a blinding index of $\widehat{FBI_i} = 0.29$. This is the true proportion of guessing beyond random chance, without the shrinkage term that accounts for DK responses. This is a good reminder of what Bang has stated about the BI_i index from the beginning that the index was not developed for the purpose of hypothesis testing, but rather as a metric. We agree with Bang to this extent: a significant value for FBI_i (BI_i) does not necessarily indicate poor blinding, but rather that the guessers were more accurate than chance should allow; and it is necessary to look at the value of the index itself rather than the significance alone.

We find that the enormous advantage the generalized approach holds over James's approach is that we can talk about the degree to which a treatment arm is unblinded. We go in to more discussion on how to use and interpret the statistic in Chapter 3, which is a review intended to serve as a regulatory guidance on blinding indexes. But first, we investigate the question: does a Bayesian approach to blinding have anything to offer?

CHAPTER TWO

Blinding Indexes Under the Bayes Paradigm

The Bayes approach has not been considered in estimating blinding success. We take the opportunity to go back over work presented in Chapter One, this time using a Bayesian approach. First we look at the prior structure used for all results in the chapter. Then we discuss the results of two simulation studies – the first a replica of simulations presented in Bang Bang et al. (2010) for $k = 2$ arms, and the second a repeat of simulations presented in Section 1.10.1. Next we reanalyze the CRISP study introduced in Section 1.6.2 from the Bayesian perspective, and finally end with a discussion and ideas for future investigation.

2.1 Prior Structure

Refer to the notation introduced in Section 1.8. Recall that the random variables $n_{i1}, n_{i2}, \ldots, n_{i,k+1}$ represent the frequency of guesses for each possible outcome among respondents assigned to the ith arm. The total number of subjects allocated to (or who guessed from) the ith arm is denoted by N_i. That is, $N_i = n_{i1} + n_{i2} + \cdots + n_{i,k+1}$. Let $\mathbf{n}_i$ denote the vector of random variables, $(n_{i1}, \ldots, n_{i,k+1})'$. Rather than thinking in terms of frequencies, we can consider the *probability* of guessing in the jth arm when assigned to arm i. This is denoted by p_{ij}, $j = 1, \ldots, k+1$. Let $\mathbf{p}_i \equiv (p_{i1}, \ldots, p_{i,k+1})'$. In Section 1.8.3 parametric forms were applied to derive moments for the FBI_i statistic. Recall that the set of random variables $\mathbf{n}$ have a multinomial distribution

$$\mathbf{n}_i | N_i, \mathbf{p}_i \sim \mathcal{M}_{k+1}(N_i; \mathbf{p}_i), \tag{2.1}$$

with probability mass function

$$p(\mathbf{n}_i | \mathbf{p}_i) = \frac{N_i!}{\prod_{j=1}^{k+1} n_{ij}!} \prod_{j=1}^{k+1} p_{ij}^{n_{ij}}. \tag{2.2}$$

$E(n_{ij}) = N_i p_{ij}$, $Var(n_{ij}) = N_i p_{ij}(1-p_{ij})$, and $Cov(n_{ij}, n_{il}) = -N_i p_{ij} p_{il}$. To model prior information about probability vector $\mathbf{p}_i$, we use a Dirichlet distribution with hyperparameter vector $\boldsymbol{\alpha}_i = (\alpha_{i1}, \alpha_{i2}, \ldots, \alpha_{i,k+1})'$. We write,

$$\mathbf{p}_i \sim \mathcal{D}_{k+1}(\boldsymbol{\alpha}_i), \tag{2.3}$$

with probability density function

$$\pi(\mathbf{p}_i|\boldsymbol{\alpha}_i) = \frac{\Gamma(\alpha_{i0})}{\prod_{j=1}^{k+1}\Gamma(\alpha_{ij})} \prod_{j=1}^{k+1} p_{ij}^{\alpha_{ij}-1}. \tag{2.4}$$

$E(p_{ij}) = \alpha_{ij}/\alpha_{i0}$, $Var(p_{ij}) = [\alpha_{ij}(\alpha_{i0}-\alpha_{ij})]/[\alpha_{i0}^2(\alpha_{i0}+1)]$, and $Cov(p_{ij}, p_{il}) = -(\alpha_{ij}\alpha_{il})/[\alpha_{i0}^2(\alpha_{i0}+1)]$, where $\alpha_{i0} \equiv \sum_{j=1}^{k+1}\alpha_{ij}$.

With likelihood (2.2) and prior (2.4), the posterior $p(\mathbf{p}_i|\mathbf{n}_i)$ is

$$\begin{aligned}
p(\mathbf{p}_i|\mathbf{n}_i) &\propto p(\mathbf{n}_i|\mathbf{p}_i)\pi(\mathbf{p}_i) \\
&= \frac{N_i!}{\prod_{j=1}^{k+1} n_{ij}!} \prod_{j=1}^{k+1} p_{ij}^{n_{ij}} \times \frac{\Gamma(\alpha_{i0})}{\prod_{j=1}^{k+1}\Gamma(\alpha_{ij})} \prod_{j=1}^{k+1} p_{ij}^{\alpha_{ij}-1} \\
&= \frac{\Gamma(N_i+1)\Gamma(\alpha_{i0})}{\prod_{j=1}^{k+1}\Gamma(n_{ij}+1)\Gamma(\alpha_{ij})} \prod_{j=1}^{k+1} p_{ij}^{n_{ij}+\alpha_{ij}-1} \\
&\propto \frac{\Gamma(N_i+\alpha_{i0})}{\prod_{j=1}^{k+1}\Gamma(n_{ij}+\alpha_{ij})} \prod_{j=1}^{k+1} p_{ij}^{n_{ij}+\alpha_{ij}-1},
\end{aligned} \tag{2.5}$$

which is the probability density function of a $\mathcal{D}_{k+1}(\mathbf{n}_i + \boldsymbol{\alpha}_i)$. Thus, the Dirichlet distribution is conjugate for the multinomial distribution. The following are true about the posterior distribution of $\mathbf{p}_i$,

$$E_p(p_{ij}) = \frac{n_{ij}+\alpha_{ij}}{N_i+\alpha_{i0}} \tag{2.6}$$

$$Var_p(p_{ij}) = \frac{(n_{ij}+\alpha_{ij})[(N_i+\alpha_{i0})-(n_{ij}+\alpha_{ij})]}{(N_i+\alpha_{i0})^2(N_i+\alpha_{i0}+1)} \tag{2.7}$$

$$Cov_p(p_{ij}, p_{il}) = \frac{-(n_{ij}+\alpha_{ij})(n_{il}+\alpha_{il})}{(N_i+\alpha_{i0})^2(N_i+\alpha_{i0}+1)}. \tag{2.8}$$

For the simulation studies and application that follow, we have chosen to use a non-hierarchical, diffuse prior such that $\boldsymbol{\alpha}_i = \mathbf{1}'$ for each $i \in 1, \ldots, k$. Under the frequentist analysis, each arm has an index calculated independently from the other arms. We mimic this under the Bayes paradigm by using independent priors for each

of the treatment arms (i.e., each treatment arms is given its own Dirichlet prior). In the discussion section (2.4) we allude to other options for the prior structure.

2.2 Simulation Studies

Because the blinding indexes are a function of $\mathbf{p}_i$, rather than solving for the closed-form distribution of BI or FBI_i, we have chosen to use `WinBUGS` to compute the posterior estimates. To run the full simulations presented in this section, `R2WinBUGS` was used.

For all simulations, we use one chain with a burn-in period of $2,000$ iterations, using the subsequent $10,000$ iterations for estimation, and setting the thinning option to one (no thinning). In total, each `WinBUGS` run is comprised of $12,000$ iterations. To stay consistent with simulation work presented in Chapter One, we use $1,000$ *replications*. Thus, `WinBUGS` reports posterior information after $12,000$ iterations, and that posterior information is stored in `R` which then calls `WinBUGS` to run another $12,000$ iterations. We do this $1,000$ times to obtain the Bayes estimators reported below. Note the difference in the use of the word "iteration" between this chapter and the previous chapter (what was called an iteration in Chapter One is now called a replication).

For all simulations, the prior structure discussed in Section 2.1 is used. Descriptive plots are provided in Appendix D. These include: dynamic trace, time series, kernel density, and autocorrelation function plots. One of each plot is provided for one random `WinBUGS` run for each of the three blinding scenarios (blinded, unblinded, reverse unblinded) at each of three DK response rates (low, medium, high). In general, the autocorrelation function figures show that problems with mixing were not an issue (using a thinning value larger than one is unnecessary). The dynamic trace and time-series plots do not show large fluctuations between successive iterations, suggesting that the Markov Chains have converged. The kernel-smoothed histograms are symmetrical, so the posterior mean is an adequate summary of the center of the

distribution. In brief, we have checked that convergence is not an issue and that our choice of simulation arguments, such as thinning and length of chain, are adequate.

For simulation results in this chapter, a multiple comparison correction is not performed to compute interval estimates. That idea was part of the novel approach in the previous work. We go back to the convention of not using a correction factor for the BI_i and therefore also for FBI_i.

2.2.1 Two Arms

Because this is the first time these indexes have been computed using Bayes, it is necessary to go back to the simulation study in Bang Bang et al. (2010) to investigate performance with two treatment arms. Due to slight differences in simulation design, Bang's simulation scenarios have been rerun. Refer to Appendix E.1 for Bang's published results and details on the difference between the original simulations and the ones rerun here.

Bang considers six cases, each at three rates for DK responses. Bang uses 0% for a low DK response rate, 30% for a moderate DK response rate, and 70% for a high DK response rate. For Case 1, both arms are randomly guessing, for Case 2 both are reverse unblinded, and for Case 3 both are unblinded. Cases 4, 5 and 6 allow for different permutations of the three blinding scenarios between treatment arms. As before, we assume Arm A represents the placebo arm when computing BI. Each arm has a sample size of 250. Design of the six cases is given in Table 2.1. Results are summarized in Table 2.2.1. As before, there are three blinding scenarios each investigated at three levels of DK response. Thus, there are nine distinct settings for which we compute FBI_i (BI_i). Histograms of the empirical posterior means for each of the nine settings are given as a histogram matrix in Figure 2.1.

Note that Bang does not use a correction factor for simultaneous inference. All results in this section use the correction factor proposed in Section 1.8.7. To see Bang's original findings, as well as a reproduction of our Frequentist and Bayesian findings without using a correction factor, refer to Appendix E: Supplemental Material.

Table 2.1: Simulation Setting from Bang et al. (2004)

Case	Assignment	Response (%)		
		DK	A	B
1	A	0	50	50
A: random	B		50	50
B: random	A	30	35	35
	B		35	35
	A	70	15	15
	B		15	15
2	A	0	10	90
A: opposite	B		90	10
B: opposite	A	30	10	60
	B		60	10
	A	70	10	20
	B		20	10
3	A	0	90	10
A: unblinded	B		10	90
B: unblinded	A	30	60	10
	B		10	60
	A	70	20	10
	B		10	20
4	A	0	10	90
A: opposite	B		10	90
B: unblinded	A	30	10	60
	B		10	60
	A	70	10	20
	B		10	20
5	A	0	50	50
A: random	B		90	10
B: opposite	A	30	35	35
	B		60	10
	A	70	15	15
	B		20	10
6	A	0	50	50
A: random	B		90	10
B: unblinded	A	30	35	35
	B		60	10
	A	70	15	15
	B		20	10

Frequentist and Bayesian methods yield similar results. This alone is an interesting find, as we did not know how the indexes would perform using the Bayesian paradigm. We move on to the case of a three-arm trial with Bayes methods.

Table 2.2: Bayes Simulation Results for $k = 2$ Arms

Case	DK (%)	James' BI		FBI/BI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
1	0	0.50 (0.022)	4.4	A	0.00 (0.063)	5.6
				B	0.00 (0.064)	5.9
	30	0.65 (0.022)	0	A	0.00 (0.054)	5.4
				B	0.00 (0.051)	5.3
	70	0.85 (0.016)	0	A	0.00 (0.034)	5.3
				B	0.00 (0.036)	5.3
2	0	0.90 (0.013)	0	A	-0.79 (0.037)	100
				B	-0.79 (0.037)	100
	30	0.90 (0.013)	0	A	-0.49 (0.041)	100
				B	-0.50 (0.043)	100
	70	0.90 (0.014)	0	A	-0.10 (0.034)	82.1
				B	-0.10 (0.033)	83.2
3	0	0.11 (0.013)	100	A	0.79 (0.037)	100
				B	0.79 (0.037)	100
	30	0.40 (0.022)	99.9	A	0.49 (0.043)	100
				B	0.50 (0.043)	100
	70	0.80 (0.017)		A	0.10 (0.033)	82.6
				B	0.10 (0.034)	83.5
4	0	0.50 (0.013)	3	A	-0.79 (0.037)	100
				B	-0.79 (0.037)	100
	30	0.65 (0.016)	0	A	-0.49 (0.043)	100
				B	-0.49 (0.042)	100
	70	0.85 (0.015)	0	A	-0.10 (0.033)	83.8
				B	-0.10 (0.033)	83.9
5	0	0.70 (0.018)	0	A	0.00 (0.061)	3.9
				B	-0.79 (0.037)	100
	30	0.77 (0.017)	0	A	0.00 (0.052)	5.2
				B	-0.49 (0.041)	100
	70	0.87 (0.014)	0	A	0.00 (0.034)	5.1
				B	-0.10 (0.035)	82.3
6	0	0.30 (0.018)	100	A	0.00 (0.061)	3.6
				B	0.79 (0.037)	100
	30	0.53 (0.020)	0.1	A	0.00 (0.050)	4.2
				B	0.49 (0.041)	100
	70	0.80 (0.017)	0	A	0.00 (0.034)	3.7
				B	0.10 (0.034)	83.9

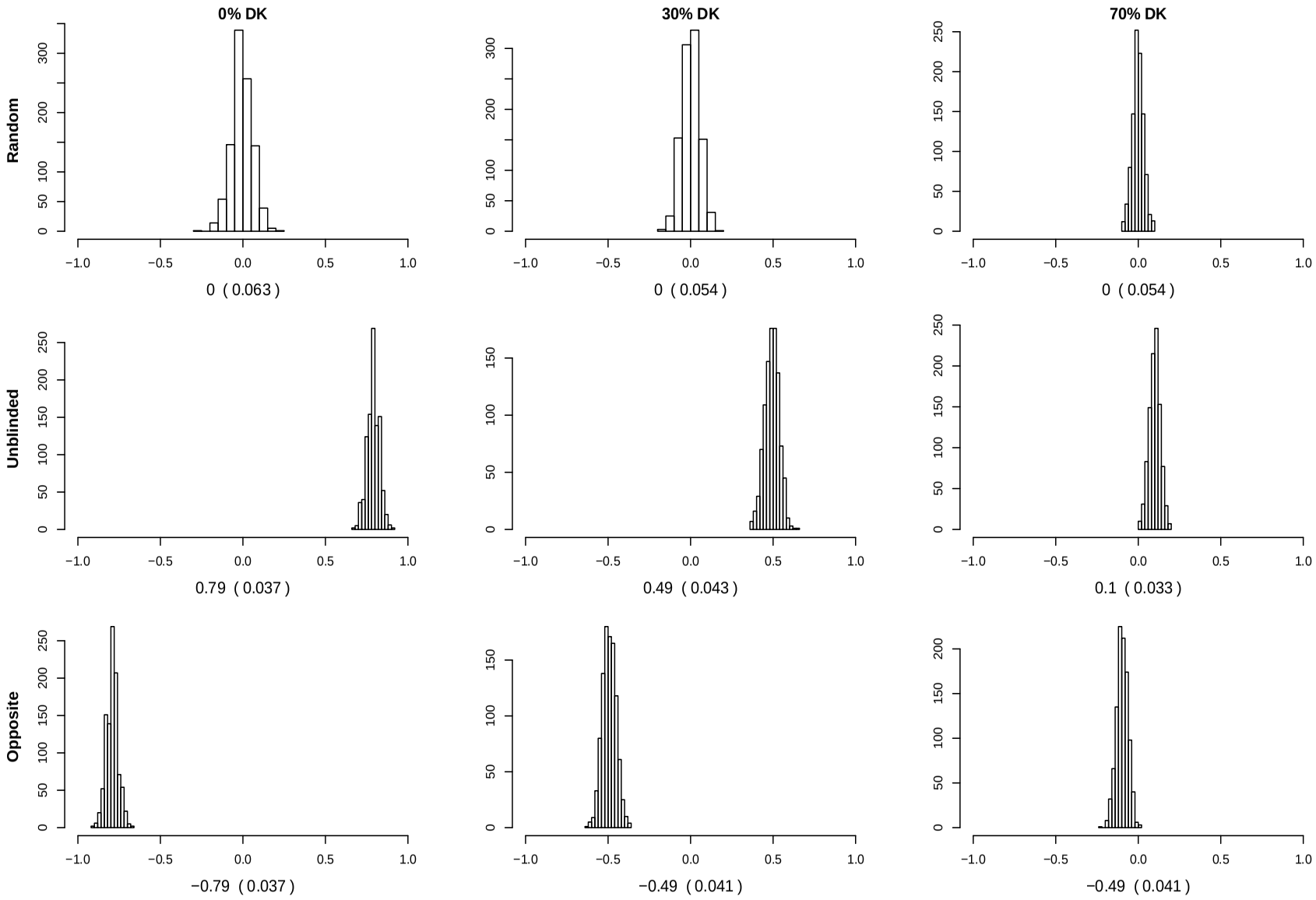

Figure 2.1: Histogram Matrix of FBI_i Posterior Means for $k = 2$ Arms

2.2.2 Three Arms

Simulation work from Sections 1.9 and 1.10 is reproduced here using the Bayes approach defined in Section 2.1. Figures 2.4 and 2.2 show the empirical posterior distribution and empirical coverage, respectively, for the nine settings introduced in Section 1.9. Figure 2.2 displays both frequentist and Bayesian summaries for easy comparison. Figure 2.3 is a matrix of box plot summaries for the empirical interval width of FBI_i under the nine settings, compared to the empirical width using the Frequentist approach in the previous chapter (ref. Table 1.9).

The empirical distributions of the nine FBI_i values under Frequentist and Bayesian approaches are very similar (ref. Figure 1.1). Empirical coverage under the Bayes approach is at least as good as coverage under the Frequentist approach from the previous chapter. Interval width is also similar between the Frequentist and Bayesian paradigms. It appears that using the prior structure from expression (2.3) does not heavily influence the FBI_i statistic.

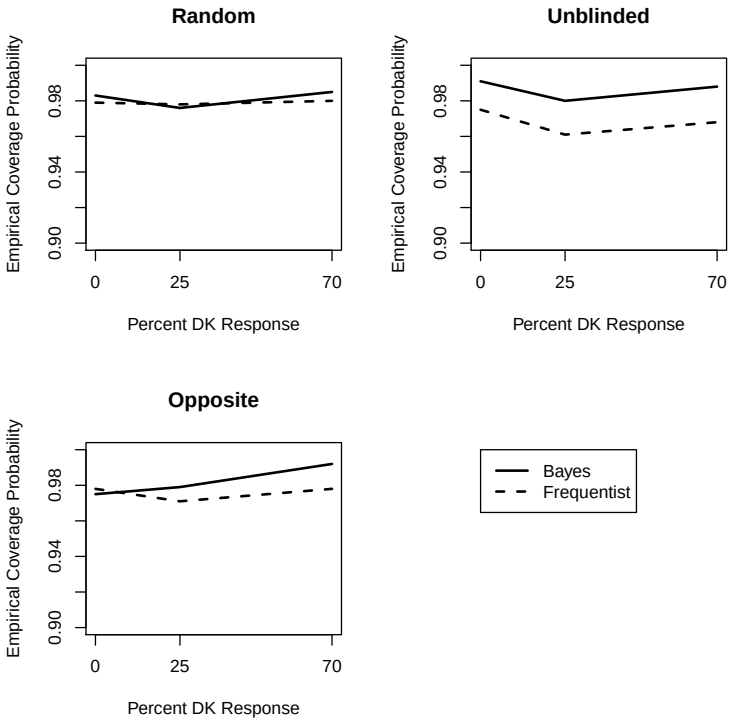

Figure 2.2: Empirical Coverage

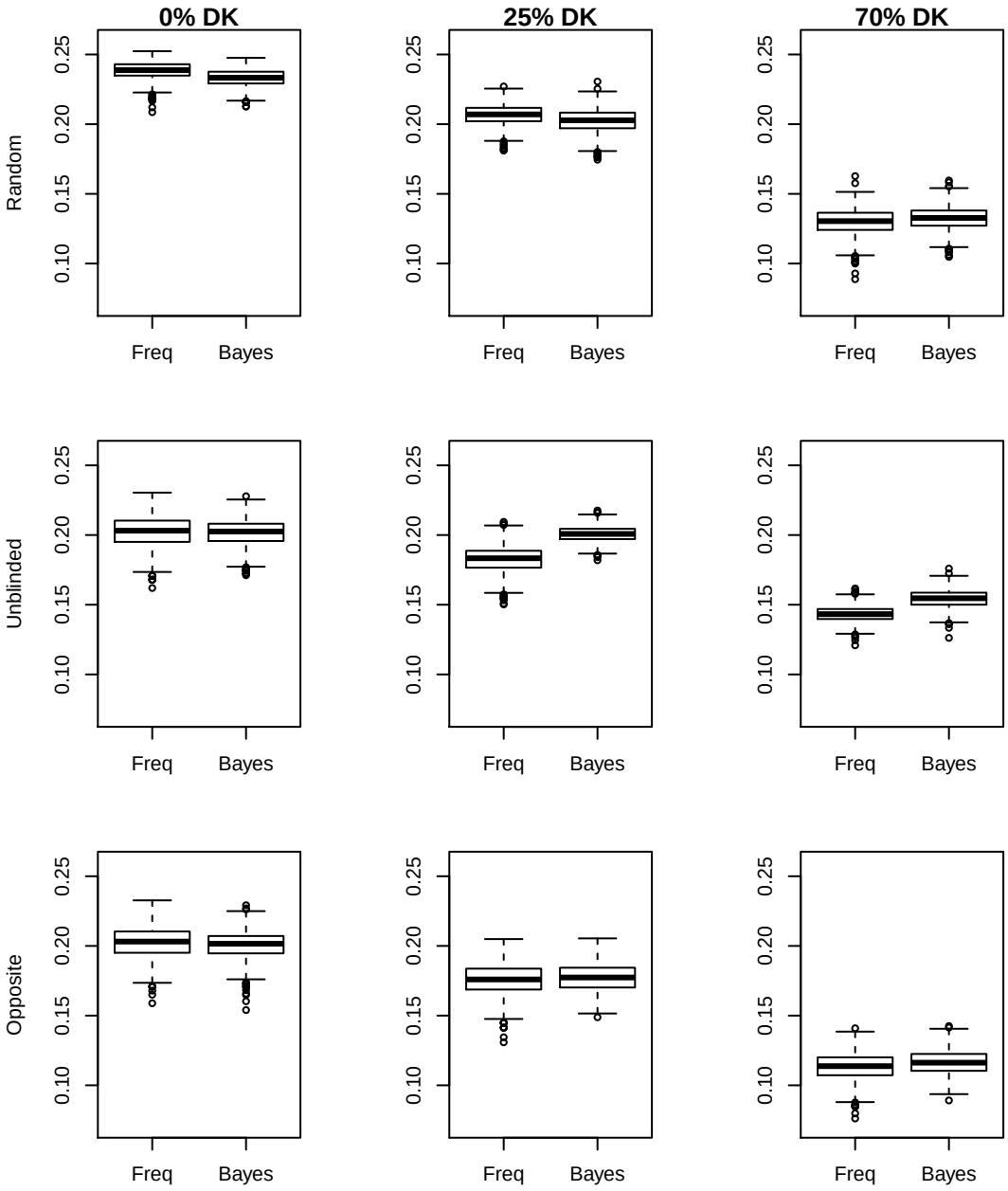

Figure 2.3: Empirical Interval Width

Table 2.3 is a summary of the ten cases (ref. Section 1.10.1) using the Bayesian approach to estimate the posterior distributions of BI and FBI_i. Expression (2.3) assumes independent responses between arms, thus we retain this assumption when computing James's BI under the Bayes paradigm. Previously, it was discovered that that the empirical posteriors of FBI_i are similar to their frequentist analogs (ref.

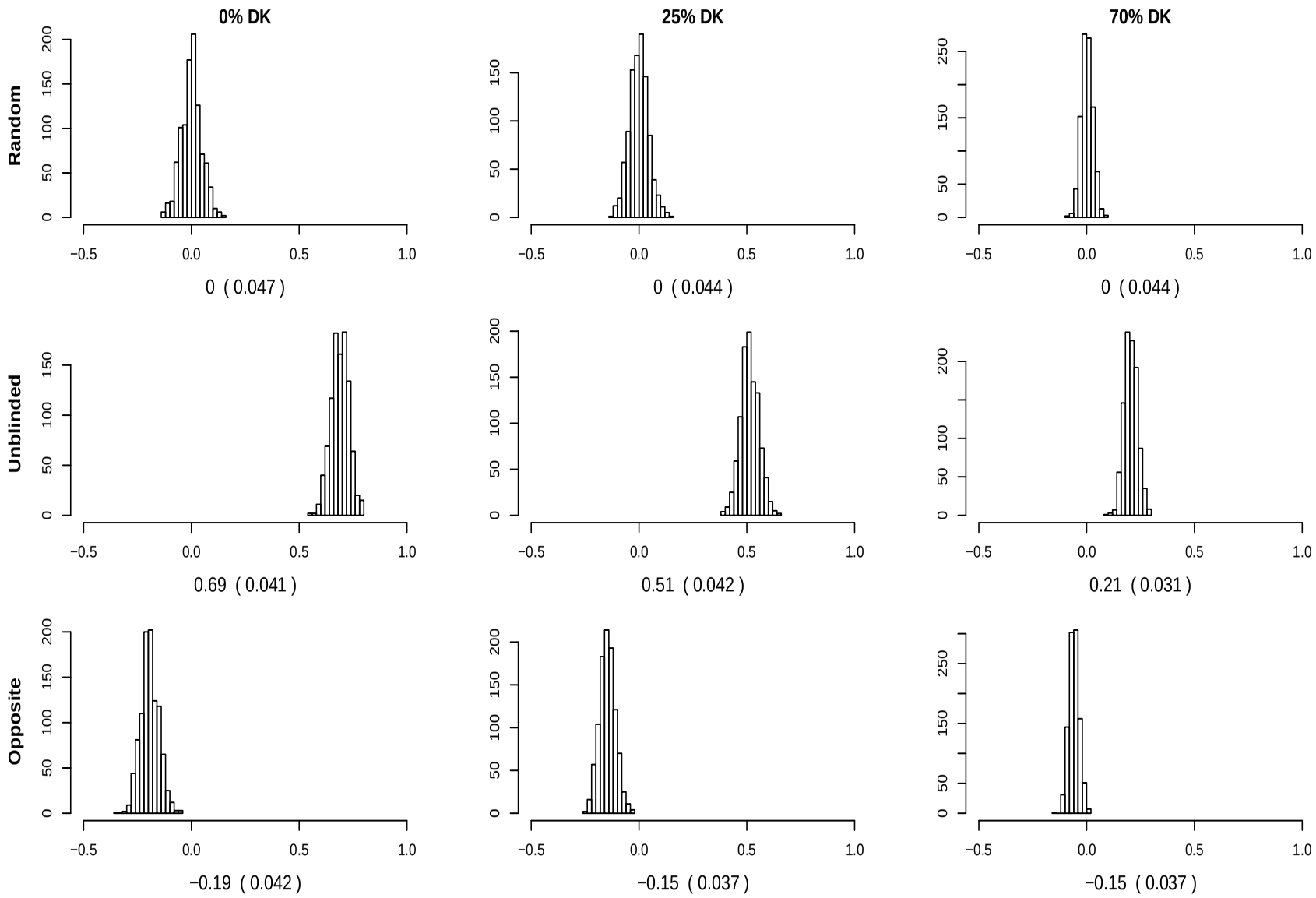

Figure 2.4: Histogram Matrix of FBI_i Posterior Means for $k = 3$ Arms

Figure 2.4), thus the values in Table 2.3 are similar to those in Table 1.11 for FBI_i. We also find that the James BI is similar under frequentist and Bayesian approaches.

Table 2.3: Bayes Simulation Results for $k = 3$ Arms

Case	DK (%)	James' BI		FBI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
1	0	0.50 (0.015)	3.3	A	0.00 (0.050)	2
				B	0.00 (0.049)	2
				C	0.00 (0.050)	1.8
	25	0.62 (0.015)	0	A	0.00 (0.042)	1.4
				B	0.00 (0.043)	1.6
				C	0.00 (0.044)	2
	70	0.85 (0.012)	0	A	0.00 (0.027)	1.5
				B	0.00 (0.028)	2.1
				C	0.00 (0.027)	2.1
2	0	0.36 (0.013)	100	A	0.00 (0.047)	1.2
				B	0.00 (0.051)	2.1
				C	0.68 (0.043)	100
	25	0.52 (0.016)	0.3	A	0.00 (0.044)	1.4
				B	0.00 (0.041)	1.5
				C	0.51 (0.042)	100
	70	0.80 (0.015)	0	A	0.00 (0.027)	1.6
				B	0.00 (0.027)	1.7
				C	0.21 (0.032)	100
3	0	0.54 (0.014)	0	A	0.00 (0.049)	1.9
				B	0.00 (0.049)	1.5
				C	-0.20 (0.042)	97.4
	25	0.66 (0.014)	0	A	0.00 (0.043)	2.1
				B	0.00 (0.041)	1.6
				C	-0.15 (0.037)	89.3
	70	0.84 (0.011)	0	A	0.00 (0.027)	1.6
				B	0.00 (0.027)	2.2
				C	-0.06 (0.024)	47.7
4	0	0.26 (0.013)	100	A	0.00 (0.049)	1.2
				B	0.69 (0.042)	100
				C	0.69 (0.042)	100
	25	0.44 (0.043)	94.7	A	0.00 (0.043)	1.7
				B	0.52 (0.040)	100
				C	0.51 (0.041)	100
	70	0.77 (0.015)	0	A	0.00 (0.026)	1.2
				B	0.21 (0.030)	100
				C	0.21 (0.030)	100
5	0	0.57 (0.013)	0	A	0.00 (0.050)	2.3
				B	-0.20 (0.042)	97.5
				C	-0.20 (0.042)	96.8
	25	0.68 (0.014)	0	A	0.00 (0.041)	1.5
				B	-0.15 (0.037)	91.2
				C	-0.15 (0.037)	91.4
	70	0.87 (0.01)	0	A	0.00 (0.027)	1.4
				B	-0.06 (0.024)	46.4
				C	-0.06 (0.024)	47.9

Table 2.3: Bayes Simulation Results for $k = 3$ Arms (cont'd)

Case	DK (%)	James' BI		FBI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
6	0	0.44 (0.013)	99.7	A	0.00 (0.047)	1.7
				B	0.69 (0.041)	100
				C	-0.19 (0.042)	96.4
	25	0.58 (0.016)	0	A	0.00 (0.044)	2.4
				B	0.51 (0.042)	100
				C	-0.15 (0.037)	89.1
	70	0.83 (0.013)	0	A	0.00 (0.027)	1.5
				B	0.21 (0.031)	100
				C	-0.06 (0.023)	47.1
7	0	0.16 (0.013)	100	A	0.68 (0.042)	100
				B	0.69 (0.042)	100
				C	0.69 (0.042)	100
	25	0.37 (0.018)	100	A	0.51 (0.043)	100
				B	0.51 (0.041)	100
				C	0.51 (0.041)	100
	70	0.74 (0.017)	0	A	0.21 (0.032)	100
				B	0.21 (0.033)	100
				C	0.21 (0.031)	100
8	0	0.34 (0.013)	100	A	0.68 (0.042)	100
				B	0.69 (0.040)	100
				C	-0.20 (0.042)	96.6
	25	0.50 (0.016)	3.3	A	0.52 (0.041)	100
				B	0.52 (0.042)	100
				C	-0.15 (0.037)	90.8
	70	0.80 (0.014)	0	A	0.21 (0.031)	100
				B	0.21 (0.031)	100
				C	-0.06 (0.024)	47.6
9	0	0.47 (0.013)	74	A	0.69 (0.041)	100
				B	-0.20 (0.041)	97.8
				C	-0.20 (0.042)	97.4
	25	0.60 (0.014)	0	A	0.51 (0.042)	100
				B	-0.15 (0.036)	91.1
				C	-0.15 (0.037)	89.7
	70	0.84 (0.012)	0	A	0.21 (0.032)	100
				B	-0.06 (0.024)	47.6
				C	-0.06 (0.024)	44.1
10	0	0.60 (0.012)	0	A	-0.20 (0.042)	96.5
				B	-0.20 (0.040)	97.7
				C	-0.20 (0.041)	97.7
	25	0.70 (0.012)	0	A	-0.15 (0.036)	91.8
				B	-0.15 (0.035)	91.1
				C	-0.15 (0.035)	91.8
	70	0.87 (0.010)	0	A	-0.06 (0.024)	44.8
				B	-0.06 (0.024)	46.6
				C	-0.06 (0.024)	47.7

2.3 CRISP Application

Recall the CRISP application used by Bang, and represented in Section 1.6.2 of this dissertation. Men and women over the age of 65 with high cholesterol were randomized to receive either Lovastatin to reduce cholesterol or a placebo. We use the same prior structure from Section 2.1. Of the 431 subjects in the study, 416 participated in the blinding survey at the end of the study. Simulations were performed using `WinBUGS`, with a burn-in of $2,000$ iterations, $10,000$ iterations to compute posterior estimates, no thinning, and $1,000$ replications (same as in Section 2.2).

Figures 2.5-2.8 show the dynamic trace, time series, kernel density, and autocorrelation function for the simulated values. The posterior mean of James BI is 0.75 with a standard error of 0.022. A 95% credible interval for BI is 0.71 to 0.79. The posterior mean of FBI_{Placebo} (also BI_{Placebo}) is 0.01 with a standard error of 0.053. A 95% credible interval for FBI_{Placebo} is -0.09 to 0.12. Finally, the posterior mean of $FBI_{\text{Lovastatin}}$ (also $BI_{\text{Lovastatin}}$) is 0.20 with a standard error of 0.035. A 95% credible interval for $FBI_{\text{Lovastatin}}$ is 0.14 to 0.27.

Comparing the FBI posterior estimates to the frequentist equivalents in Table 1.4, we see that the point estimates are nearly identical. The interval estimates under the Bayes approach are slightly wider, but not by much. The posterior point and interval estimates of James BI for the CRISP data match the results reported in Bang's manuscript Bang et al. (2010). Simulation results from the previous chapter confirm that the Bayesian results should be close to the frequentist values, so the addition of Bayesian methods to the CRISP study behaved as expected.

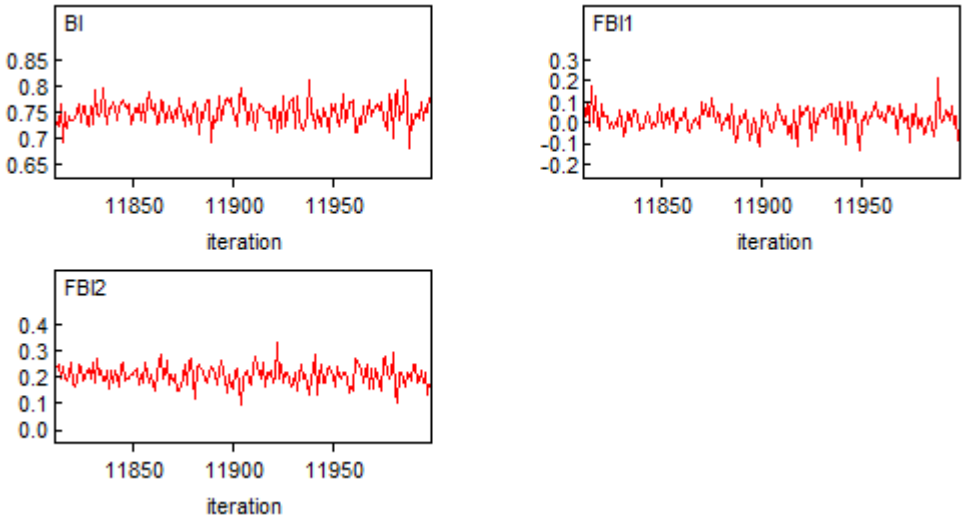

Figure 2.5: Dynamic Trace Plots for CRISP

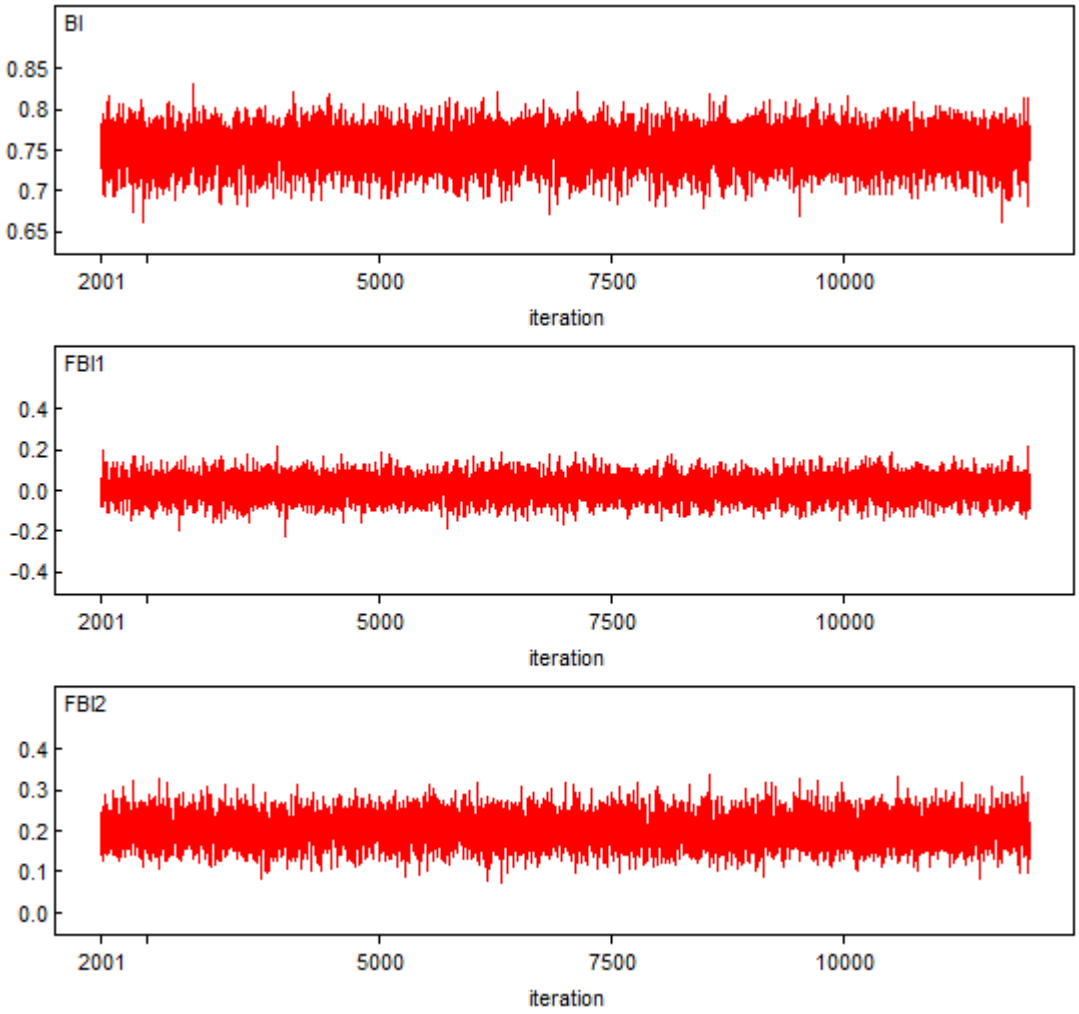

Figure 2.6: Time Series Plots for CRISP

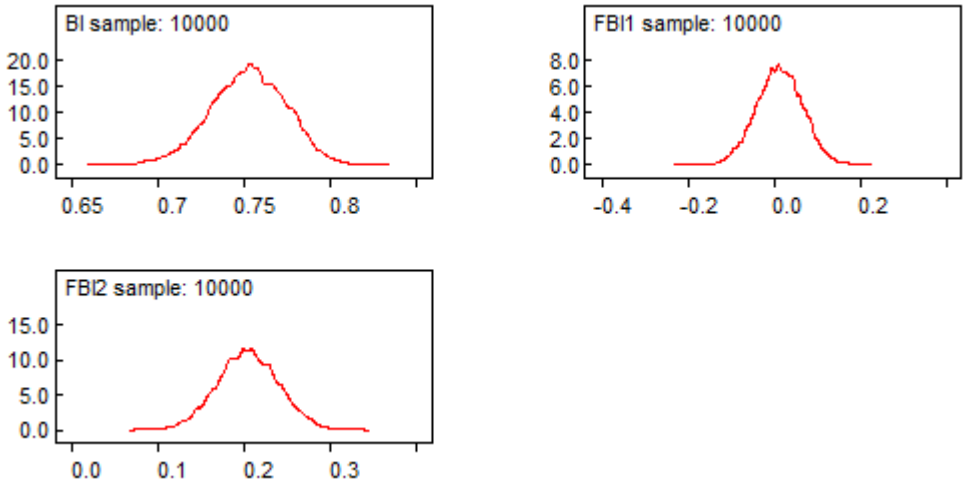

Figure 2.7: Kernel Density Plots for CRISP

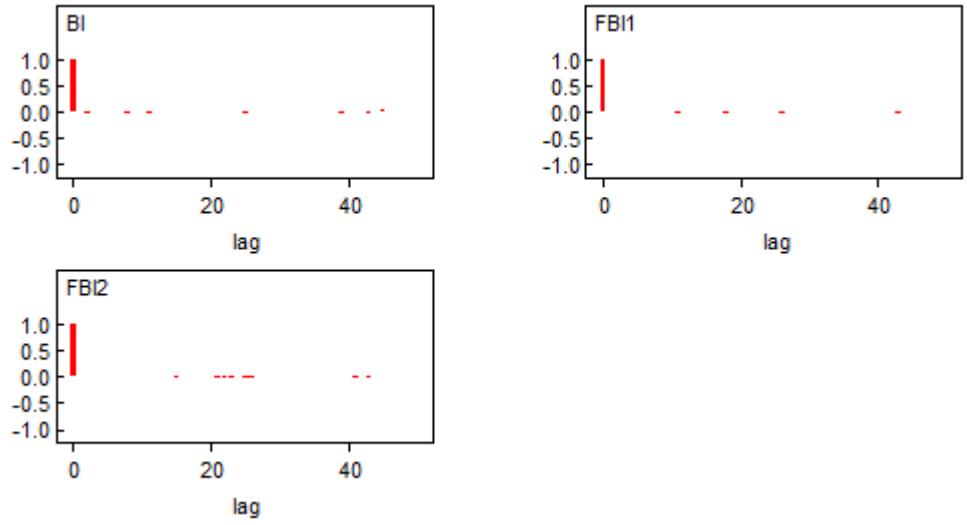

Figure 2.8: Autocorrelation Function Plots for CRISP

2.4 Discussion

There is great potential for the use of Bayesian methods in measuring blinding. The first step taken here shows that, under the most simplistic approach, a Bayesian approach tells the same story as the frequentist approach. However, there is much still open to investigation. To start, it would be of interest to change the diffuse prior structure in (2.3). For example, it would be interesting to see how the Jeffery's prior with hyperparameter $\boldsymbol{\alpha}_i = 0.5 \times \mathbf{1}'$ influences the posterior. On the other hand,

what about an informative prior that allows for unequal values of α_{ij}, $j = 1, \ldots, k$? Imagine a scenario in which researchers know that the DK response rate is likely to be extreme (either very large of very small). For example, say researchers are more prone to answer DK, or perhaps subjects in the placebo arm are expected to answer DK often. In this case, it would be interesting to use a prior structure such that the hyperparameter of the Dirichlet corresponding to the probability of a DK response, $\alpha_{i,k+1}$, is not equal to the other α_{ij} hyperparameters. It would be interesting to see how an informative prior would affect the value of the blinding indexes when we have prior information about DK response rates.

Another possibility is to model dependence between treatment arms by our choice of hyperparameters. Rather than using prior distributions with independent hyperparameters between treatment arms, a hierarchical structure can be used. It would be interesting to see how adding dependence between the arms changes the results, if at all.

Also, if an expert has an opinion about the value of the index itself rather than the guessing, a prior structure could be placed on the index. We can look at the induced prior on the response probabilities (2.3) to see how the opinion about the index transfers information back about the multinomial probabilities. For example, in a trial with two treatment arms we could start with a $uniform(-1, 1)$ prior on FBI_i and see what this implied about p_{i1}, p_{i2} and p_{i3}. From here, we would change the prior structure to something non-flat such as a beta distribution that is shifted and scaled as necessary.

These may seem like theoretical exercises, but the truth is any one of the above scenarios is possible. The struggle with choosing more informative priors is determining if anything stands a good chance of having a practical application – are researchers really less likely to respond with DK, do subjects in the placebo group respond DK more often on average, etc. Without more published findings on blinding results, it would be difficult to get away with using a more informative prior structure. However,

if a researcher is conducting a trial which includes a treatment used in another published study, then informative priors could be a useful way to incorporate knowledge from the previous study. Many of these questions can be addressed via simulation, but the applicability is dependent on researchers publishing blinding data.

CHAPTER THREE

A Guidance on Blinding Indexes for Regulatory Agencies and Clinical Trialists

3.1 Introduction

In clinical trials, blinding is desirable due to its ability to reduce information bias and improve compliance and retention. However, just because we don't tell study participants or researchers which treatment is being administered on an individual basis does not mean blinding has been successful. Once blinding has been implemented we assume it has been achieved. More dangerously, trialists carry forward making conclusions about study outcomes as if blinding were maintained throughout the study, ignoring possible biases that may have occurred due to a lack of blinding Schulz and Grimes (2002). Currently, blinding is more of a concern during trial development and initiation rather than something to investigate during the study or at trial completion. It is important for researchers to assess the success of blinding for all trials implementing single-, double-, or even triple-blinds.

There is not a standard approach for measuring blinding, nor is there a consensus on how to collect data on blinding. However, there are methods available to researchers who wish to report on blinding. A range of statistical approaches has been used, but the most recent trend is in blinding indexes: estimators that are developed specifically with the intent to measure how well a study has been blinded. We walk through the various statistical methods, including blinding indexes, in the Methods section. Then in Section 3.3 we look at simulation results to compare blinding indexes. In Section 3.4 we look at a Bayesian approach to blinding, then an application in the following section. We end with concluding remarks and goals for blinding reporting.

There is a lack of agreement about when to ask about blinding. The options are: before randomization, shortly after randomization, during trial, and/or at study completion Bang et al. (2010). Most commonly investigators ask at study completion, but this convention has been criticized. A primary concern is that, by the time the trial has ended, correct guesses could be confounded with efficacy and side-effects.

Regardless of when we administer the blinding survey, the data structure is the same. The process is slightly different, depending on which party is blinded. Study participants are asked which treatment they believe they are on. The subjects are allowed to choose among the treatment options, or say that they don't know (DK). On the other hand, if we are assessing the blind in a non-patient group (i.e. among researchers, coordinators, etc.) then we ask them about the assignment of several patients they have been involved with during the study. Experimenters are required to guess what each of their patients is assigned to, and again are allowed to answer among the treatment options or DK.

Blinding data can be easily summarized in a contingency table, where the columns and rows are represented by actual treatment assignment and guessed treatment (including DK). Correct guesses are evidence of unblinding, and incorrect guesses are used as evidence for blinding. DK responses are thought to suggest blinding because, after all, the point is that subjects should not know what group they are assigned to. We must be careful when we say subjects are *blinded* and *unblinded*. Unblinded, in this case, does not necessarily mean that an individual found out they were definitely assigned to a certain group (although this is a possibility), rather we use it to say that someone has a good notion of the treatment group to which they have been assigned.

In the next two subsections we go through a brief history of blinding as a statistical problem. First we look at some of the ways in which researchers have analyzed blinding data using common statistical approaches. Then we go in to the newest

trend in blinding reporting, which is to use a blinding index. We will convince you that using a blinding index is superior to using a traditional statistical approach, and that to use these indexes is a straightforward process that can be easily incorporated in to any study protocol where blinding and randomization are implemented.

3.2.1 Traditional Statistical Approaches

Early analysis of blinding data was quite simple. An excellent review of statistical methods for blinding assessment is given in Bang 2010 Bang et al. (2010), we merely summarize their work here. Hughes and Krahn Hughes and Krahn (1985) looked at the proportions of correct and incorrect guesses. Blinding was deemed unsuccessful if the proportion of correct guesses was larger than the proportion of incorrect guesses. They, along with Margraf et al. Margrat et al. (1991) and others, used a traditional Chi-square test on the contingency table of blinding data. Kolahi et al. Kolohi et al. (2009) used McNemar's test for the case where there were two arms in the trial. Wisner et al. Wisner et al. (2001) reported the κ statistic as an improvement on the previous approaches which merely test for significance but in no way measure the degree of blinding. A criticism of the κ statistic as a measure of blinding is that is measures *agreement* which is not the desirable outcome for blinding survey responses.

A pitfall with the above mentioned approaches is that each one neglects the DK responses. We have already stated that a DK response should be indicative of blinding, but up to now we have thrown away all of that information. We seem to accept that we need to compare correct guesses to incorrect guesses in order to determine if people are adequately blinded (random guessing should also indicate blinding). It is not important, however, to compare the proportion of DK responses with the proportion of correct and incorrect guesses. Thus, even though we could have included the DK responses when performing the Chi-square test, it makes little sense to do so. We transition out of traditional approaches and look at a few approaches that are specific to our problem.

3.2.2 Blinding Indexes

James et al. James et al. (1996) realized the value in using a κ-like statistic to *measure* blinding as opposed to only testing for it. A modification was made to the κ statistic and James's blinding index, BI, was born. With the James BI we can measure blinding in a meaningful way. The index is on a scale from zero to one, with smaller numbers representing unblinding and larger numbers representing blinding. The middle, one-half, represents random guessing. Thus a value of one-half may seem adequate, accepting that random guessing acts as a surrogate for blinding. In the presence of DK responses, the index shifts toward one. James BI places weight on various responses. For example, a correct guess (undesirable) is assigned a weight of zero whereas a DK response is assigned a weight of one. The weight of an incorrect guess depends on the guess itself. James et al. distinguish types of incorrect guesses. For example if there are multiple treatment groups then a subject could guess the wrong treatment. If said subject is on a treatment but guesses a different treatment, this carries less weight than if he or she is assigned to the placebo group but guesses an active treatment. The weight structure for the different types of incorrect guesses does not often change our conclusion, as James et al. have shown in a simulation study.

The James index has received slight criticism because of the influence DK responses on determining its value. Recall that according to the standard survey structure there are two indicators of blinding: DK responses and random guessing among non-DK responses. The James index focuses on DK responses to indicate blinding. For example, if nobody answers DK in a blinding assessment *and* subjects are randomly guessing, then the James BI would have a value of one-half. But a "complete blind" is indicated by a value of one, which only occurs when all respondents give a DK response James et al. (1996). Therefore, Bang et al. Bang et al. (2004) took a different approach to creating a blinding index. First off, the Bang blinding index,

BI_i, is on a scale from negative one to one, with zero representing "blinding." Bang's index seems to have the following advantages over the James index:

(1) Mathematically, zero represents the case of random guessing or all DK responses, or a mix of random guessing and DK responses. Thus, for both indicators of blinding the value of the Bang index approaches the same constant (not the case in James).

(2) Also, BI_i can detect something known as *reverse unblinding* – where guessers are incorrect more often than they would be if they were randomly guessing. Opposite guessing could be an indication of adequate blinding, because subjects clearly are unable to determine their treatment allocation. However, there are practical reasons why reverse unblinding is undesirable; for example in some studies the dropout rate in a placebo group can be higher, so if respondents believe they are in a placebo group then trial enrollment and retainment could suffer Montgomery (1999). Note that reverse unblinding is suggested by negative values of the blinding index.

(3) Additionally, the Bang index is arm-specific, meaning that it is calculated for each treatment arm independently. The advantage here is that we can look at blinding success within each arm. This is important because not all arms are necessarily blinded/unblinded to the same degree. A disadvantage is that Bang does not give an overall blinding index for the study as a whole, but we believe that the advantage of arm-specific investigation outweighs the disadvantage of no study-wide index score.

(4) Lastly, the Bang index is very interpretable. The value of the index represents the *proportion of correct guesses beyond random guessing.* Imagine we have study with a placebo group and a control group and the Bang blinding index for these groups is calculated to be 0.15 and 0.8, respectively. We interpret these values as correct guessing 15% and 80% beyond what random chance would suggest, respectively. Although the James index serves well as an

index, it is unclear what the practical difference between 0.6 and 0.7 is for the James BI.

There are, however, some limitations to Bang's approach.

(1) First of all, the Bang index is only valid for a binary response (placebo/treatment), ignoring DK. The theory behind BI_i assumes that within each treatment arm, the probability of guessing correctly and the probability of guessing incorrectly are equal, with random guessing. This is true only if: there are only two arms, *or* the subjects are only allowed to guess between two options. James et al. had their own way of addressing different types of incorrect answers using the weight structure, but to Bang et al. the answer is either correct or incorrect. If the survey is designed so that each person is only given two options (e.g. placebo or experimental treatment), then the index is valid no matter the number of arms in the trial. Also, if there are only two arms to begin with then the Bang index is always valid. However, it is untrue to assume equal probabilities of guessing correctly and incorrectly if given more than two options, with only one being a correct response. For such a scenario, we cannot apply Bang and therefore are forced to use James's BI.

(2) Second, Bang admits that the purpose of the index is not for hypothesis testing. James et al. make a similar claim about their index. However, the advantage Bang's index holds over James's index is the interpretability. Neither were developed for the sole purpose of hypothesis testing, and each index places an emphasis on understanding the point estimate. So in this regard we still find Bang's approach to be preferable to James's approach.

We should note that Bang et al. proposed a variant of the BI_i index which uses weights. The weights are not determined by the guess, but rather by the guesser. Some people may be more confident in their response than others, and therefore Bang decided these should have more weight in the index.

We tend to prefer the Bang approach to measuring blinding, but there are some properties that need addressing before trying to make it the standard in blinding reporting. Most obviously, we require an extension to the Bang BI_i that allows for more than two study arms.

We have developed a statistic that stays true to the Bang paradigm of measuring blinding success, but that has been generalized to measure blinding for any amount of study arms. For k arms theFBI ranges from $-(k-1)^{-1}$ to one, with zero still serving as the reference for blinding (random guessing and/or DK responses). This index can detect both unblinding and reverse unblinding, preserves the null constant representing blinding success, is treatment-arm specific, is interpretable, and can be used for any number of treatment groups. We can calculate a blinding index under the Bang paradigm for any study which looks at blinding. FBI is merely a generalization of Bang's BI_i, and in the case of two treatment arms the indexes are exactly the same. From this point forward, we refer to James's blinding index as BI and the Williamson blinding index as FBI. Keep in mind that any reference to FBI also implies Bang's BI_i when there are $k = 2$ arms.

The lower bound of FBI is dependent on the number of treatment arms. James et al. James et al. (1996) suggest this is an undesirable property for an index (referring to the κ statistic). Negative values for FBI suggest *reverse unblinding* and therefore it is not of great concern for us to measure the degree of reverse unblinding, merely being able to detect it (as indicated by a negative value of the index) is advantageous. Therefore, the "lack of interpretability" for negative values of FBI is not alarming. We did investigate a weight structure to FBI. However, the mathematics yield an index with undesirable properties. Therefore, we present no weighted alternative to the FBI.

3.2.3 Blinding Surveys

One of the benefits of the blinding indexes is that they are capable of being used no matter when the data are collected. An additional benefit of FBI is that the host

can decide before hand a degree of unblinding that would be acceptable (e.g. setting a maximum unblinding threshold). This degree of unblinding could account for biases in survey administration timing. For example, if measuring post-trial responses a larger value of FBI may be more acceptable than if the data were collected shortly after randomization. Thresholds could be set for the James BI as well, but the interpretability of FBI yields a less abstract definition of the blinding threshold.

3.3 Simulation Findings

We already mentioned simulation work conducted by James et al. that showed blinding results are robust to the choice of the weight structure. In this simulation, James fixed the weights for correct guesses and DK responses at zero and one, respectively. Then, assuming guessing an incorrect treatment is better indication of blinding that guessing the correct treatment but incorrect dose, the weights of these two incorrect guess types were varied between 0.2 and 0.9 with only the constraint that the weight of the former be larger than the weight of the latter James et al. (1996).

Bang et al. performed simulations to compare their approach with James's approach. Recall that Bang's index measures each arm independently, and therefore we end up with two BI_i for each BI in the simulation. Additionally, there are three blinding scenarios for any arm: random guessing, unblinded, and reverse unblinded. Bang considers three levels of DK responses: 0% to represent no DK responses, 30% to represent a moderate amount of DK responses, and 70% to represent a large amount of DK responses. There are three blinding scenarios each of which can occur at the three DK response levels, thus forming nine possible "cases" for the simulation. Bang et al. report six of these cases in their manuscript Bang et al. (2004). In fact this is more than necessary to look at the Bang index: remember each "case" yields two BI_i values, one for each arm, which gives twelve in total using the six cases. There are only nine unique simulation scenarios for the Bang index (three blinding scenarios matched with three DK response levels). Because the Bang index

is arm-specific, which group represents the placebo arm and which one represents the experimental arm is irrelevant. However, because of the weight structure, the cases are not interchangeable for the James index. Thus, although reporting six cases does not fully cover the unique setting for the James index, Bang more than sufficiently shows the properties of BI_i.

Bang discovers that under most blinding scenarios the James index shows significant "blind success". As we have already stated, DK responses increase the James index. For all blinding scenarios paired with 30% or 70% DK responses (with the exception of one: both arms being unblinded), the James index had a 0% rejection rate. Furthermore, BI is not built to detect opposite guessing (reverse unblinding), and this scenario actually inflates the index making it appear that blinding was achieved. It is not necessarily incorrect to interpret opposite guessing as successful blinding, but looking at James's BI alone we would not know where the evidence of blinding is coming from. Bang's index was powerful to detect all three blinding scenarios, regardless of the percent DK responses.

We recreated Bang's simulations, but using $k = 3$, 4, and 5 arms. The results were similar: FBI is powerful to detect each of the blinding scenarios at various levels of DK responses. It was of interest to us to keep an eye on the unblinding group as the number of arms increases. Recall the lower bound of FBI approaches zero as the number of arms increases. We worried there would be a significant loss of power as we increased the number of arms. It turns out this is true, for an interesting reason. FBI is based on the assumption of random guessing, that is to say that subjects should be guessing in each group with an equal probability (aside from DK). At some point, there become too many treatment groups to tell the difference between opposite guessing and random guessing. For example, let's say that a certain arm is *reverse unblinded* to a degree such that only 10% of them are guessing correctly (of the subjects who are guessing and not responding "DK"). A 10% accuracy rate is to be expected in a group that is guessing *randomly* among $k = 10$ arms. Essentially,

the accuracy of the *reverse unblinded* group and of the *randomly guessing* group are the same, meaning there is no mathematical difference between opposite guessing and random guessing. This problem seems to be paradoxical: if a group is reverse unblinded but responding as would be expected for random guessing, are they reverse unblinded? Simply put, the power of FBI does decrease as the number of treatment arms increases, all else constant.

Our simulations and Bang's simulations proved some of our concerns about the James index. First of all, we can now see how sensitive BI is to DK responses. Additionally, we see that the James index does not properly describe what is happening when the blinding scheme between treatment arms is not the same. If all arms happen to be unblinded to some degree, or all arms are randomly guessing, then the James index is fine. But when behavior differs between arms, the James index tries to compromise and leaves us with a less clear picture of what is going on. The James index does not do a great job at harmonizing the two indicators of blinding: random guessing and DK responses. A plus side to BI is that it can distinguish between random guessing and DK responses. We just criticized this in the previous sentence, but it could also be considered a good thing. FBI yields a value "close" to zero in either case, which would suggest to us blinding success. Theoretically the James index can tell us which of the indicators of blinding is dominant (but we'll mention one final time that DK response overwhelm random guessing).

Finally, to summarize our simulation findings, FBI is shown to have decent power even for small sample sizes. Sample size is not a big issue in blinding, because usually we are measuring and reporting blinding in Phase III studies which have larger enrollment. However, there can be opportunity for using the index on small sets of data. For example, if we administer the blinding surveys shortly after randomization we might want to do an interim analysis when the number of subjects per arm is not substantial. Also, it is not impossible that we investigate blinding in a Phase II trial, or even a very small Phase III trial. And finally, we could apply the index to subsets

of the study population (e.g.: look at blinding in males versus females, or across other demographic or geographic variables).

3.4 Bayesian Approach to Blinding Indexes

A Bayesian approach to measuring blinding has potential to add valuable information. For the first time Bayesian methods have been used with the blinding indexes. The prior structure examined is simple, a conjugate Dirichlet prior with equal hyperparameters so that we have a "flat" prior. Posterior point and interval estimates under the Bayesian approach matched very closely the results from the frequentist simulations.

However, the potential influence of Bayesian methods is much more interesting. With different prior structures, it is possible we could gain additional information about blinding success. This could be a particularly interesting approach for small sample sizes, modeling dependence between treatment arms under the Bang approach, and for cases where DK responses are extreme (very low or very high). Additionally, if investigators believe they know something about the distribution of the index rather than the guesses themselves, we could look at the induced prior on the multinomial frequencies. There is decent potential for Bayes in this field.

3.5 Application

A four month double-blind, placebo-controlled trial was conducted between September 2006 and February 2011 on 138 overweight but otherwise healthy, middle-aged to older adults to determine if omega-3 polyunsaturated fatty acid (n-3 PUFA) supplementation would decrease serum cytokine production and depressive symptoms. The trial consisted of three arms: 2.5 g/d n-3 PUFAs, 1.25 g/d n-3 PUFAs, and placebo capsules mirroring the proportion of fatty acids in the average American diet. The manufacturer of the drugs added fish flavoring to the placebo capsules to aid in blinding. The study investigators were interested in the success of the blind

among study participants and experimenters. Blinding surveys were administered to participants at the final visit during the study Kiecolt-Glaser et al. (2012).

Table 3.1: Omega-3 and Inflammation Participant Responses

Actual Assignment	Guessed Assignment				Total
	1.25 g/d	2.5 g/d	Placebo	DK	
1.25 g/d *n*-3 PUFAs	17	3	12	12	44
2.5 g/d *n*-3 PUFAs	14	5	9	16	44
Placebo	9	4	17	14	44
Total	40	12	38	42	132

Participant responses to the blinding questionnaire are summarized in Table 3.1. The authors reported the James index for the participants, omitting one observation. Unsure of which observation was dropped, we recalculated the index for the full data in Table 3.1. The James index for study participants is 0.61 (95% CI: 0.53–0.68). Kiecolt-Glaser et al. report that "blinding is considered adequate if the index is greater than one-half." Because participants (and experimenters) were allowed to guess between all three arms, the investigators were limited to use the James index as apposed to Bang et al.'s alternative.

With the generalized index, we may now look at blinding results under Bang's paradigm. The Williamson blinding index for participants at the end of the study in the 50% fish group (1.25 g/d *n*-3 PUFAs) is 0.22 (95% CI: −0.02–0.45). For the 100% Fish group (2.5 g/d *n*-3 PUFAs), the index is −0.15 (95% CI: −0.31–0.02). Finally, for the placebo group the FBI index is 0.24 (95% CI: 0.01–0.46). The only significant finding is in the placebo arm, but note that the lower confidence bound is very close to 0. Additionally, note that the confidence bounds for the two active treatment arms are also close to 0. In this case, the two approaches agree: there are minimal concerns that unblinding occurred in this study.

We also examined the experimenters' responses using the FBI statistic. There was an overwhelming rate of DK responses among experimenters, which yielded large values of BI. Values of FBI were close to zero for all treatment arms. In the experimenter group, our approaches again agree that unblinding is not a concern.

3.6 Suggested Use of Blinding Indexes

We have considered two ways to measure blinding: the James method and the Bang method (with the Williamson generalization). The Williamson (Bang) FBI allows us to look at blinding in each arm independently and can detect reverse unblinding, whereas James BI gives a study-wide measure of blinding and distinguished between blinding due to random guessing and blinding as a result of DK responses. We have seen that these two methods do not necessarily agree with one another. We therefore recommend using both indexes when assessing a blind. There is not much to consider when both of the indexes agree with one another, but it can be equally advantageous to see when the two methods disagree Bang et al. (2010). Thus, although the indexes were developed under different philosophies of blinding success, it seems more practical that they be used *complementary* to one another.

Blinding is a complex issue, and we hope that study sponsors and investigators begin to recognize the importance of reporting blinding results at the end of any trial which implements a blind. It can give insight to other investigators who wish to maintain blinds in similar studies, especially for determining a threshold of acceptable unblinding *a priori*. The interpretability of FBI is especially useful when trying to define a threshold. We are not limited to accepting or rejecting blinding with the use of these indexes, rather we can measure the degree to which studies are unblinded. To us, this is the biggest improvement blinding indexes hold over the traditional statistical approaches – we don't have to classify a trial as blinded or unblinded, instead we can describe the blind in meaningful terms and determine if the results meet a standard established in the protocol.

Since James et al. introduced their blinding index, there have been an increasing amount of articles dedicating to how to assess blinding in randomized, controlled trials. Researchers such as Park Park et al. (2008) and Bang Bang and Park (2013) have been creating awareness for the need to report blinding with trial findings. They have also debated the appropriateness of blinding questionnaires, when they should be administered, and how freely we should accept DK responses. Park is leading a project, `blindingindex.org`, where clinical trialists can deposit data on the success of blinding. It is evident that blinding is becoming a new hot-topic in the world of clinical trials.

Statistically we have developed indexes which aim to measure the success of blinding. First was the James index, followed by the Bang index which was then generalized by Williamson et al. These indexes can be applied to data collected at any time in the trial, and can be applied to any blinded group (subjects, researchers, etc.). For larger studies, we could even assess blinding across various demographic and geographic groups. For example, we could assess blinding separately at different clinical sites James et al. (1996). Our goal is that it becomes standard for researchers to report at least one of the aforementioned blinding indexes when summarizing study findings.

It will require additional research and expertise to decide what to do with our information on blinding. Should it simply serve us so that we may try to implement better blinds in the future, or should we leverage the degree of unblinding against study conclusions? Zhang et al. (2013) have done research in this area, suggesting a causal model between placebo and treatment-specific effects. To be able to effectively determine the role blinding will play in the future of trial conclusions, we need more trialists to assess blinding. With the current statistical approaches, we have provided them easy, efficient reporting tools that can be used for all types of randomized, controlled trials.

APPENDICES

APPENDIX A

R Functions to Compute Blinding Indexes

A.1 Function to Compute the James Blinding Index

```
James.BI <- function(d,w){
#######################################################################
# Function: James.BI                                                  #
#                                                                     #
# Purpose: Function to calculate the BI statistic and asymptotic      #
#          variance proposed in James et al. (1996)                   #
#                                                                     #
#######################################################################
# INPUT                                                               #
#    d           A (k+1 x k) matrix of frequencies:                   #
#                    - columns are actual treatment                   #
#                    - rows are guessed treatment                     #
#                    - last row is "dont know"                        #
#    w           A (k+1 x k) matrix of weights corresponding          #
#                to the types of responses in the matrx d.            #
#                                                                     #
# OUTPUT                                                              #
#    BI.hat      The blinding index statistic                         #
#    var.asymp   Asymptotic variance of BI.hat                        #
#                                                                     #
#######################################################################
# To call function:                                                   #
#    James.BI(d,w)                                                    #
```

```
#                                                                      #
########################################################################
  # Initial definitions
  k = dim(d)[2]
  N = sum(d)
  n0 = sum(d[k+1,])
  p = d/N
  PDK = n0/N
  pD0 = pDe = 0

  # pD0
  for(i in 1:k){
    for(j in 1:k){
      pD0 = pD0 + w[i,j]*p[i,j]/(1-PDK)
    }
  }

  # pDe   for(i in 1:k){
  pi. = sum(p[i,])
  for(j in 1:k){
    p.j = sum(p[,j])
    p0j = p[k+1,j]
    pDe = pDe + w[i,j]*pi.*(p.j-p0j)/(1-PDK)^2
    }
  }

  # Blinding Index
  kD = (pD0-pDe)/pDe
  BI = (1+PDK+(1-PDK)*kD)/2
```

```
  # Asymptotic variance
  var.pt1 = 0
  for(i in 1:k){
    for(j in 1:k){
      inner = 0
      for(r in 1:k){
        inner = inner + sum(p[r,]*w[r,j]+(sum(p[,r])-p[k+1,r])*w[i,r])
      }
      var.pt1 = var.pt1 + p[i,j]*(1-PDK)^2*((1-PDK)*w[i,j]-(1+kD)*inner)^2/
                          (4*(sum(p[i,])*(sum(p[,j])-p[k+1,j])*w[i,j])^2)
    }
  }
  var.asymp = (var.pt1 + PDK*(1-PDK)-(1-PDK)*(1+kD)*(PDK+(1-PDK)*(1+kD)/4))/N

  # Output
  list(BI.hat=BI, var.asymp=var.asymp)
}
```

1.1.1 Example Run

```
> # Data from James et al. (1996), p. 1425
> w = matrix(c(0,.5,.75,1,.5,0,.75,1,.75,.75,0,1),nrow=4)
> w
     [,1] [,2] [,3]
[1,] 0.00 0.50 0.75
[2,] 0.50 0.00 0.75
[3,] 0.75 0.75 0.00
[4,] 1.00 1.00 1.00
> d11 = matrix(c(29,0,0,4,0,29,0,4,0,0,29,4),nrow=4)
> d11
```

```
     [,1] [,2] [,3]
[1,]   29    0    0
[2,]    0   29    0
[3,]    0    0   29
[4,]    4    4    4
>
> # James Blinding Index
> James.BI(d11,w)
$BI.hat
[1] 0.1212121


$var.asymp
[1] NaN
```

A.2 Function to Jackknife the James Blinding Index

```
jack.BI <- function(d,w){
#####################################################################
# Function: jack.BI                                                 #
#                                                                   #
# Purpose: Function to jackknife the BI statistic                   #
#                                                                   #
# Written by: Forrest Williamson                                    #
#                                                                   #
#####################################################################
# INPUT                                                             #
#    d                 A (k+1 x k) matrix of frequencies:           #
#                          - columns are actual treatment           #
#                          - rows are guessed treatment             #
#                          - last row is "dont know"                #
```

```
#    w                A (k+1 x k) matrix of weights          #
#                     corresponding to the types of          #
#                     responses in the matrx d.              #
#                                                            #
# OUTPUT                                                     #
#    BI.hat           The blinding index estimate            #
#    jack.estimates   A (k+1 x k) matrix of the new          #
#                     jackknife estimates corresponding      #
#                     to omiting one observation from        #
#                     each element of d.                     #
#    jack.mean        Jackknife mean estimate                #
#    jack.var         Jackknife variance estimate            #
#    jack.se          Jackknife standard error estimate      #
#    jack.CI95        95% jackknife confidence interval;     #
#                     jack.mean(+-)1.96*jack.se              #
#    reject           Indicator:                             #
#                       1   reject H0                        #
#                       0   FTR H0                          #
#                                                            #
##############################################################
# To call function:                                          #
#    jack.BI(d,w)                                            #
#                                                            #
##############################################################

  # Initial definitions
  k = dim(d)[2]
  N = sum(d)
```

```
# Compute overall blinding index
BI.hat = James.BI(d,w)$BI.hat

# Matrix of new BI values (leave one out per cell)
T = matrix(NA,nrow=k+1,ncol=k)
for(i in 1:(k+1)){
  for(j in 1:k){
    dummy  = matrix(0,nrow=k+1,ncol=k)
    if(d[i,j]>0){ dummy[i,j] = -1 }       #no negatives
    d.temp = d + dummy
    T[i,j] = James.BI(d.temp,w)$BI.hat
  }
}

# Matrix of jackknife estimates
jack.estimates = N*BI.hat - (N-1)*T

# Jackknife mean, variance, se, and 95% CI
jack.mean = sum(d*jack.estimates)/N
jack.var  = sum(d*(jack.estimates-jack.mean)^2)/(N-1)
jack.se   = sqrt(jack.var/N)
jack.CI95 = jack.mean + 1.96*c(-jack.se,jack.se)

# Reject at 5% level (one-sided)?
if(jack.mean+qnorm(.95)*jack.se<0.5) {reject=1} else {reject=0}

# Names
list(BI.hat=BI.hat, jack.estimates=jack.estimates,
```

```
       jack.mean=jack.mean, jack.var=jack.var, jack.se=jack.se,
       jack.CI95=jack.CI95, reject=reject)
}
```

1.2.1 Example Run

```
> # Cooperative Study No.107 (James et al. p.1427)
> study = matrix(c(41,66,30,44,27,72,24,51,22,36,64,52),nrow=4)
> study
     [,1] [,2] [,3]
[1,]   41   27   22
[2,]   66   72   36
[3,]   30   24   64
[4,]   44   51   52
>
> # James BI
> w = matrix(c(0,.5,.75,1,.5,0,.75,1,.75,.75,0,1),nrow=4)
> w
     [,1] [,2] [,3]
[1,] 0.00 0.50 0.75
[2,] 0.50 0.00 0.75
[3,] 0.75 0.75 0.00
[4,] 1.00 1.00 1.00
> jack.BI(study,w)
$BI.hat
[1] 0.5564209

$jack.estimates
           [,1]      [,2]       [,3]
[1,] 0.02150165 0.6829181  0.8186271
```

```
[2,] 0.57112654 0.1016612  0.8026661
[3,] 0.77415572 0.8699067 -0.1264497
[4,] 1.00000000 1.0000000  1.0000000

$jack.mean
[1] 0.5562377

$jack.var
[1] 0.176728

$jack.se
[1] 0.01827784

$jack.CI95
[1] 0.5204131 0.5920623

$reject
[1] 0
```

A.3 Function to Compute the Bang Blinding Index

```
Bang.BI <- function(d, alpha=0.05){
#########################################################################
# Function: Bang.BI                                                     #
#                                                                       #
# Purpose: Function to calculate the "2x3-format" BI statistic          #
#          and variance proposed in Bang et al. (2004)                  #
#                                                                       #
# Written by: Forrest Williamson                                        #
#                                                                       #
```

```
#####################################################################
# INPUT                                                             #
#    d          A (k+1 x k) matrix of frequencies:                  #
#                    - columns are actual treatment                 #
#                    - rows are guessed treatment                   #
#                    - last row is "dont know"                      #
#    alpha      Significance level for 2-sided CI                   #
#                                                                   #
# OUTPUT                                                            #
#    BI.hat     A k-dimensional vector of blinding index            #
#               statistic by treatment arm                          #
#    var.hat    A k-dimensional vector of estimated                 #
#               variances corresponding to each BI.hat              #
#    CI         A 2-sided 95% confidence interval for               #
#               each BI.hat                                         #
#                                                                   #
#####################################################################
# To call function:                                                 #
#    Bang.BI(d,.05)                                                 #
#                                                                   #
#####################################################################
  # Define number of treatment arms
  k = dim(d)[2]

  # Create empty vectors
  BI.hat = var.hat = c()

  # Evaluate the statistic within each treatment arm
  for(i in 1:k) {
```

```
    # P vector (P_k|i) of length k
    ni = apply(d,2,sum)[i]
    P  = d[1:k,i]/ni

    # Blinding Index in treatment arm i
    r.hat = d[i,i] / sum(d[1:k,i])
    BI.hat[i] = (2*r.hat-1)*(sum(P))

    # Variance of the blinding index in treatment arm i
    var.hat[i] = (sum(P*(1-P)) + 2*prod(P))/ni
  }

  # Confidence intervals
  z  = qnorm(1-alpha/2)
  CI = cbind(BI.hat - z*sqrt(var.hat), BI.hat + z*sqrt(var.hat))

  # Output
  list(BI.hat=BI.hat,var.hat=var.hat,CI=CI)
}
```

1.3.1 Example Run

```
> # CRISP data from Bang 2004
> CRISP = matrix(c(82,25,170,27,29,83),nrow=3)
> CRISP
     [,1] [,2]
[1,]   82   27
[2,]   25   29
[3,]  170   83
>
```

```
> # Results
> Bang.BI(CRISP)
$BI.hat
[1] 0.20577617 0.01438849

$var.hat
[1] 0.001241653 0.002896911

$CI
             [,1]      [,2]
[1,]  0.13671275 0.2748396
[2,] -0.09110258 0.1198796
```

A.4 Function to Compute the Williamson Blinding Index

```
FBI = function(DatMat,alpha=0.05) {
#################################################################
# Function: FBI                                                 #
#                                                               #
# Purpose: To calculate the FBIi indexes for all groups,        #
#          as well as the variance for each FBIi.               #
#                                                               #
# Written by: Forrest Williamson                                #
#                                                               #
#################################################################
# INPUT                                                         #
#    DatMat   The frequency contingency table, as an R          #
#             matrix object, where the rows are the actual      #
#             treatment assignments and the columns are the     #
#             guessed assignments (in the same order as the     #
```

```
#             rows), and the last column must be the DK     #
#             responses. Thus, dim(DatMat) = k x (k+1).     #
#    alpha    Probability of a Type I Error for a 2-sided   #
#             test (double desired significance level for   #
#             1-sided test). Default = .05                  #
#                                                           #
# OUTPUT                                                    #
#    FBI.pe   A vector of length k of point estimates of    #
#             the FBI index, in the same order as the       #
#             input matrix, DatMat.                         #
#    FBI.var  A vector of length k of variance estimates    #
#             under the null hypothesis of random guessing. #
#    Reject   A vector of length k binary responses:        #
#             1 = reject null hypothesis, 0 = FTR H0 at the #
#             specified signifigance level, alpha (2-sided).#
#                                                           #
#############################################################
# To call function:                                         #
#    library(gdata)                                         #
#    FBI(data,0.05)                                         #
#                                                           #
#############################################################


  # Dimension
  k = dim(DatMat)[1]


  # Critical value
  cv = qnorm(1-alpha/2)
```

```
  # Compute FBI and Var for each group
  FBI.pe  = c()
  FBI.var = c()
  Reject  = c()

  for(i in 1:k) {
    # Definitions
    N = sum(DatMat[i,])
    p = sum(DatMat[i,1:k]) / (k*N)

    # FBI measures
    FBI.pe[i]  = (k*DatMat[i,i] - sum(DatMat[i,1:k])) / ((k-1)*N)
    FBI.var[i] = (k^2*P[i]*(1-P[i]) + P%*%(1-P) + 2*k*sum(P[i]*P) -
                    2*sum(lowerTriangle(P%*%t(P)))) / ((k-1)*N)

    # CI
    lower = max(-1/(k-1), FBI.pe[i] - cv*sqrt(FBI.var[i]))
    upper = min(FBI.pe[i] + cv*sqrt(FBI.var[i]), 1)

    # Reject or FTR
    if(lower<0 & 0<upper) {Reject[i]=0
    } else {Reject[i]=1}
  }

  # Output
  results = list(FBI.pe=FBI.pe, FBI.var=FBI.var, Reject=Reject)
  return(results)
}
```

1.4.1 Example Run

```
> # Cooperative Study No.107 (James et al. p.1427)
> study = t(matrix(c(41,66,30,44,27,72,24,51,22,36,64,52),nrow=4))
> study
     [,1] [,2] [,3] [,4]
[1,]   41   66   30   44
[2,]   27   72   24   51
[3,]   22   36   64   52
>
> #FBIi
> FBI(t(study))
$FBI.pe
[1] -0.03867403  0.26724138  0.20114943

$FBI.var
[1] 0.00770607 0.01191196 0.01106221

$Reject
[1] 0 1 0
```

A.5 Function to Jackknife the Williamson Blinding Index

```
jack.FBI = function(DatMat,Alpha=.05) {
#################################################################
# Function: jack.FBI                                            #
#                                                               #
# Purpose: to obtain simultaneous jackknife interval            #
#          estimates for the  FBIi statistics, using a          #
#          Bonferroni correction.                               #
#                                                               #
```

```
# Written by: Forrest Williamson                                #
#                                                               #
#################################################################
# INPUT                                                         #
#    DatMat   The frequency contingency table, as an R          #
#             matrix object, where the rows are the actual      #
#             treatment assignments and the columns are the     #
#             guessed assignments (in the same order as the     #
#             rows), and the last column must be the DK         #
#             responses. Thus, dim(DatMat) = k x (k+1).         #
#    Alpha    Global probability of a Type I Error for k        #
#             2-sided tests. Default = .05                      #
#                                                               #
# OUTPUT                                                        #
#    FBI.pe   A vector of length k of point estimates of        #
#             the FBI indexes, in the same order as the         #
#             input matrix, DatMat.                             #
#    FBI.j    A vector of length k of point estimates of        #
#             the jackknife mean FBI index, in the same         #
#             order as the input matrix, DatMat.                #
#    FBI.cs   A matrix of dimension k x 2 of confidence         #
#             sets (1st and 2nd columns represent lower and     #
#             upper bounds, respectively, and the rows          #
#             represent the groups in the same order as         #
#             DatMat). Note: Bonferroni Multiple Correction     #
#             Factor used to create simultaneous intervals.     #
#    Reject   A vector of length k binary responses:            #
#             1 = reject null hypothesis, 0 = FTR H0 at the     #
#             specified global significance level, alpha.       #
```

```
#                                                                #
##################################################################
# To call function:                                              #
#    jack.FBI(data,0.05)                                         #
##################################################################
  # Dimension
  k = dim(DatMat)[1]

  # Bonferroni correction
  alpha.star = 1 - (1-Alpha)^(1/k)

  # Critical Value
  cv = qnorm(1-alpha.star/2)

  # Set up null vectors and matrices
  FBI.pe = c()
  FBI.j  = c()
  FBI.cs = matrix(ncol=2,nrow=k)
  Reject = c()

  for(i in 1:k) {
    # Definitions
    N   = sum(DatMat[i,])
    nii = DatMat[i,i]

    # FBI
    T0         = (k*DatMat[i,i] - sum(DatMat[i,1:k])) / ((k-1)*N)
    FBI.pe[i] = T0
```

```
  # T1 - omit 1 from correct guess, T2 = omit 1 from incorrect guess
  T1 = ((k-1)*(nii-1) - (sum(DatMat[i,1:k])-nii)) / ((k-1)*(N-1))
  T2 = ((k-1)*nii - (sum(DatMat[i,1:k])-nii-1)) / ((k-1)*(N-1))

  # 'IF' statement to protect against when there 100% DK responses
  if(sum(DatMat[i,1:k]) == 0) {Ti = 0
  } else {Ti = c(rep(T1,nii), rep(T2,(sum(DatMat[i,1:k])-nii)))}

  # Pseudo values
  Ji = N*T0 - (N-1)*Ti

  # Jackknife mean
  J         = mean(Ti)
  FBI.j[i]  = mean(Ti)

  # Jackknife variance
  Sj2 = sum((Ji-J)^2)/(N-1)

  # Confidence interval (adj.)
  SEj        = sqrt(Sj2/N)
  lower      = J - cv*SEj
  upper      = J + cv*SEj
  FBI.cs[i,] = c(lower, upper)

  # Reject
  if(lower<0 & 0<upper) {Reject[i]=0
  } else {Reject[i]=1}
}
```

```
  # Output
  results = list(FBI.pe=FBI.pe, FBI.j=FBI.j,
                 FBI.cs=FBI.cs, Reject=Reject)
  return(results)
}
```

1.5.1 Example Run

```
> # Cooperative Study No.107 (James et al. p.1427)
> study = t(matrix(c(41,66,30,44,27,72,24,51,22,36,64,52),nrow=4))
> study
     [,1] [,2] [,3] [,4]
[1,]   41   66   30   44
[2,]   27   72   24   51
[3,]   22   36   64   52
>
> #FBIi
> jack.FBI(study)
$FBI.pe
[1] -0.03867403  0.26724138  0.20114943

$FBI.j
[1] -0.03860503  0.26660087  0.20065384

$FBI.cs
            [,1]       [,2]
[1,] -0.14498048 0.06777042
[2,]  0.15253322 0.38066853
[3,]  0.08603318 0.31527450
```

```
$Reject

[1] 0 1 1
```

APPENDIX B

R Code for Chapter 1 Simulations

B.1 Cases

Below we provide R code for Case 1 with 0% DK responses (ref. Table 1.10). Each of the ten cases and special examinations have similar code, with the only difference being the probabilities of the multinomial distributions. The R directory is set so that the appropriate functions (ref. A) can be read in.

```
#########################################################################
# CASE 1: rrr                                                           #
#########################################################################
# Read in functions
source("jack.BI.R")
source("jack.FBI.R")

# Global parameters
its = 1000
N   = 200
w = matrix(c(0,.75,.75,1,.75,0,.5,1,.5,.75,0,1),nrow=4)

#############
## DK = 0% ##
#############
# Set seed
set.seed(11100)
```

```
# Multinomial probabilities
pA = c(1/3,1/3,1/3,0)
pB = c(1/3,1/3,1/3,0)
pC = c(1/3,1/3,1/3,0)

# Simulation objects (Point Estimates & Rejection vectors)
FBI.peA  = FBI.peB  = FBI.peC  = James.pe  = c()
FBI.RejA = FBI.RejB = FBI.RejC = James.Rej = c()

# Simulation
for(i in 1:its) {
  # Simulate data
  A = t(rmultinom(1,N,pA))
  B = t(rmultinom(1,N,pB))
  C = t(rmultinom(1,N,pC))
  DatMat = rbind(A,B,C)

  # FBI
  fbi= FBI(DatMat)

  # Save FBI results
  FBI.peA[i]  = fbi$FBI.pe[1]
  FBI.peB[i]  = fbi$FBI.pe[2]
  FBI.peC[i]  = fbi$FBI.pe[3]
  FBI.RejA[i] = fbi$Reject[1]
  FBI.RejB[i] = fbi$Reject[2]
  FBI.RejC[i] = fbi$Reject[3]

  # James
```

```
  James = jack.BI(t(DatMat),w)

  # Save James BI results
  James.pe[i]  = James$BI.hat
  James.Rej[i] = James$reject
}

# Create data frame
Case1.0 = data.frame(James.pe=James.pe, James.Rej=James.Rej,
                     FBI.peA=FBI.peA, FBI.peB=FBI.peB,
                     FBI.peC=FBI.peC, FBI.RejA=FBI.RejA,
                     FBI.RejB=FBI.RejB, FBI.RejC=FBI.RejC)

##################
## Write Output ##
##################
write.csv(Case1.0, "Case1.0.csv")
```

B.2 Sample Size and Power

The following two sections provide R code used to create Figures 1.2 and 1.3, respectively

2.2.1 Over DK Response Rates

```
#########################################################################
# Sample Size & Power Simulation                                        #
#########################################################################
# Read in functions
source("jack.FBI.R")
```

```
# Global parameters
its = 1000
N   = c(10,50,100,200,500,750,1000)

##################
## DK = 0%, k=3 ##
##################
# Set seed
set.seed(12345)

# Multinomial probabilities
pA = c(1/3,1/3,1/3,0)
pB = c(.1,.8,.1,0)
pC = c(.4,.4,.2,0)

# Higher level simulation objects
Power = matrix(nrow=3,ncol=length(N))

# Simulation
for(j in 1:length(N)) {
  # Lower level simulation objects
  FBI.RejA = FBI.RejB = FBI.RejC = c()

  for(i in 1:its) {
    # Simulate data
    A = t(rmultinom(1,N[j],pA))
    B = t(rmultinom(1,N[j],pB))
    C = t(rmultinom(1,N[j],pC))
    DatMat = rbind(A,B,C)
```

```
    # FBI
    fbi= FBI(DatMat)

    # Save FBI results
    FBI.RejA[i] = fbi$Reject[1]
    FBI.RejB[i] = fbi$Reject[2]
    FBI.RejC[i] = fbi$Reject[3]
  }

  Power[,j] = c(1-mean(FBI.RejA), mean(FBI.RejB), mean(FBI.RejC))
}

# Create data frame
SS.0 = as.data.frame(Power,row.names=c('R','U','O'))

####################
## DK = 25%, k=3 ##
####################
# Set seed
set.seed(12345)

# Multinomial probabilities
pA = c(1/4,1/4,1/4,1/4)
pB = c(.075,.6,.075,.25)
pC = c(.3,.3,.15,.25)

# Higher level simulation objects
Power = matrix(nrow=3,ncol=length(N))
```

```
# Simulation
for(j in 1:length(N)) {
  # Lower level simulation objects
  FBI.RejA = FBI.RejB = FBI.RejC = c()

  for(i in 1:its) {
    # Simulate data
    A = t(rmultinom(1,N[j],pA))
    B = t(rmultinom(1,N[j],pB))
    C = t(rmultinom(1,N[j],pC))
    DatMat = rbind(A,B,C)

    # FBI
    fbi= FBI(DatMat)

    # Save FBI results
    FBI.RejA[i] = fbi$Reject[1]
    FBI.RejB[i] = fbi$Reject[2]
    FBI.RejC[i] = fbi$Reject[3]
  }

  Power[,j] = c(1-mean(FBI.RejA), mean(FBI.RejB), mean(FBI.RejC))
}

# Create data frame
SS.25 = data.frame(Power,row.names=c('R','U','O'))

###################
```

```
## DK = 70%, k=3 ##
####################
# Set seed
set.seed(12345)

# Multinomial probabilities
pA = c(.1,.1,.1,.7)
pB = c(.03,.24,.03,.7)
pC = c(.12,.12,.06,.7)

# Higher level simulation objects
Power = matrix(nrow=3,ncol=length(N))

# Simulation
for(j in 1:length(N)) {
  # Lower level simulation objects
  FBI.RejA = FBI.RejB = FBI.RejC = c()

  for(i in 1:its) {
    # Simulate data
    A = t(rmultinom(1,N[j],pA))
    B = t(rmultinom(1,N[j],pB))
    C = t(rmultinom(1,N[j],pC))
    DatMat = rbind(A,B,C)

    # FBI
    fbi= FBI(DatMat)
```

```
    # Save FBI results
    FBI.RejA[i] = fbi$Reject[1]
    FBI.RejB[i] = fbi$Reject[2]
    FBI.RejC[i] = fbi$Reject[3]
  }

  Power[,j] = c(1-mean(FBI.RejA), mean(FBI.RejB), mean(FBI.RejC))
}

# Create data frame
SS.70 = data.frame(Power,row.names=c('R','U','O'))

#########################################################
## Create Graphics                                     ##
#########################################################
# Global plotting parameters
colA = "red"
colB = "blue"
colC = "green"
ltyA = 1
ltyB = 2
ltyC = 3
lwd  = 2

# Plot: Power vs. Sample Size by %DK
par(mfrow=c(2,2))
plot(N,SS.0[1,],main="0% DK",xlab="Subjects per Arm",ylab="Power",
     ylim=c(0,1),xlim=c(0,1000),'l',col=colA,lwd=lwd,xaxt='n')
  axis(1,at=N)
```

```
  lines(N, SS.0[2,], col=colB, lty=ltyB, lwd=lwd)
  lines(N, SS.0[3,], col=colC, lty=ltyC, lwd=lwd)
plot(N,SS.25[1,],main="25% DK",xlab="Subjects per Arm",ylab="Power",
     ylim=c(0,1),xlim=c(0,1000),'l',col=colA,lwd=lwd,xaxt='n')
  axis(1,at=N)
  lines(N, SS.25[2,], col=colB, lty=ltyB, lwd=lwd)
  lines(N, SS.25[3,], col=colC, lty=ltyC, lwd=lwd)
plot(N,SS.70[1,],main="70% DK",xlab="Subjects per Arm",ylab="Power",
     ylim=c(0,1),xlim=c(0,1000),'l',col=colA,lwd=lwd,xaxt='n')
  axis(1,at=N)
  lines(N, SS.70[2,], col=colB, lty=ltyB, lwd=lwd)
  lines(N, SS.70[3,], col=colC, lty=ltyC, lwd=lwd)
frame()
  legend(0,1,c("Random","Unblinded","Opposite"),col=c(colA,colB,colC),
         lty=c(ltyA,ltyB,ltyC),lwd=lwd)
```

2.2.2 Over Number of Trial Arms

```
#########################################################################
# Sample Size & Power Simulation                                        #
#########################################################################
# Read in functions
source("jack.FBI.R")

# Global parameters
its = 1000
N   = c(10,50,100,200,500,750,1000)

###################
## DK = 70%, k=3 ##
```

```
###################
# Set seed
set.seed(12345)

# Multinomial probabilities
pA = c(.1,.1,.1,.7)
pB = c(.03,.24,.03,.7)
pC = c(.12,.12,.06,.7)

# Higher level simulation objects
Power = matrix(nrow=3,ncol=length(N))

# Simulation
for(j in 1:length(N)) {
  # Lower level simulation objects
  FBI.RejA = FBI.RejB = FBI.RejC = c()

  for(i in 1:its) {
    # Simulate data
    A = t(rmultinom(1,N[j],pA))
    B = t(rmultinom(1,N[j],pB))
    C = t(rmultinom(1,N[j],pC))
    DatMat = rbind(A,B,C)

    # FBI
    fbi= FBI(DatMat)

    # Save FBI results
    FBI.RejA[i] = fbi$Reject[1]
```

```
    FBI.RejB[i] = fbi$Reject[2]
    FBI.RejC[i] = fbi$Reject[3]
  }

  Power[,j] = c(1-mean(FBI.RejA), mean(FBI.RejB), mean(FBI.RejC))
}

# Create data frame
SS.70 = data.frame(Power,row.names=c('R','U','O'))

####################
## DK = 70%, k=4 ##
####################
# Set seed
set.seed(12345)

# Multinomial probabilities
pA = c(.3/4,.3/4,.3/4,.3/4,.7)
pB = c(.02,.24,.02,.02,.7)
pC = c(.08,.08,.06,.08,.7)

# Higher level simulation objects
Power = matrix(nrow=3,ncol=length(N))

# Simulation
for(j in 1:length(N)) {
  # Lower level simulation objects
  FBI.RejA = FBI.RejB = FBI.RejC = c()
```

```
  for(i in 1:its) {
    # Simulate data
    A = t(rmultinom(1,N[j],pA))
    B = t(rmultinom(1,N[j],pB))
    C = t(rmultinom(1,N[j],pC))
    DatMat = rbind(A,B,C,A)

    # FBI
    fbi= FBI(DatMat)

    # Save FBI results
    FBI.RejA[i] = fbi$Reject[1]
    FBI.RejB[i] = fbi$Reject[2]
    FBI.RejC[i] = fbi$Reject[3]
  }

  Power[,j] = c(1-mean(FBI.RejA), mean(FBI.RejB), mean(FBI.RejC))
}

# Create data frame
SS.70b = data.frame(Power,row.names=c('R','U','O'))

##########################
## DK = 70%, k=5 (ruorr)##
##########################
# Set seed
set.seed(12345)
```

```
# Multinomial probabilities
pA = c(.3/5,.3/5,.3/5,.3/5,.3/5,.7)
pB = c(.06/4,.24,.06/4,.06/4,.06/4,.7)
pC = c(.06,.06,.06,.06,.06,.7)

# Higher level simulation objects
Power = matrix(nrow=3,ncol=length(N))

# Simulation
for(j in 1:length(N)) {
  # Lower level simulation objects
  FBI.RejA = FBI.RejB = FBI.RejC = c()

  for(i in 1:its) {
    # Simulate data
    A = t(rmultinom(1,N[j],pA))
    B = t(rmultinom(1,N[j],pB))
    C = t(rmultinom(1,N[j],pC))
    DatMat = rbind(A,B,C,A,A)

    # FBI
    fbi= FBI(DatMat)

    # Save FBI results
    FBI.RejA[i] = fbi$Reject[1]
    FBI.RejB[i] = fbi$Reject[2]
    FBI.RejC[i] = fbi$Reject[3]
  }
```

```
  Power[,j] = c(1-mean(FBI.RejA), mean(FBI.RejB), mean(FBI.RejC))
}

# Create data frame
SS.70c = data.frame(Power,row.names=c('R','U','O'))

#########################################################
## Create Graphics                                     ##
#########################################################
# Global plotting parameters
colA = "red"
colB = "blue"
colC = "green"
ltyA = 1
ltyB = 2
ltyC = 3
lwd  = 2

# Plot: Power vs. Sample Size by k
par(mfrow=c(2,2))
plot(N,SS.70[1,],main="3 Arms",xlab="Subjects per Arm",ylab="Power",
     ylim=c(0,1),xlim=c(0,1000),'l',col=colA,lwd=lwd,xaxt='n')
  axis(1,at=N)
  lines(N, SS.70[2,], col=colB, lty=ltyB, lwd=lwd)
  lines(N, SS.70[3,], col=colC, lty=ltyC, lwd=lwd)
plot(N,SS.70b[1,],main="4 Arms",xlab="Subjects per Arm",ylab="Power",
     ylim=c(0,1),xlim=c(0,1000),'l',col=colA,lwd=lwd,xaxt='n')
  axis(1,at=N)
```

```
  lines(N, SS.70b[2,], col=colB, lty=ltyB, lwd=lwd)
  lines(N, SS.70b[3,], col=colC, lty=ltyC, lwd=lwd)
plot(N,SS.70c[1,],main="5 Arms",xlab="Subjects per Arm",ylab="Power",
     ylim=c(0,1),xlim=c(0,1000),'l',col=colA,lwd=lwd,xaxt='n')
  axis(1,at=N)
  lines(N, SS.70c[2,], col=colB, lty=ltyB, lwd=lwd)
  lines(N, SS.70c[3,], col=colC, lty=ltyC, lwd=lwd)
frame()
  legend(0,1,c("Random","Unblinded","Opposite"),col=c(colA,colB,colC),
         lty=c(ltyA,ltyB,ltyC),lwd=lwd)
```

B.3 Distribution of FBI_i

Figures 1.1, 1.4 and 1.5 plot simulated values of FBI_i for the three blinding scenarios (blinded, unblinded, reverse unblinded) across different levels of percent-DK responses for $k = 3$, 4 and 5 trial arms, respectively. The code below was used to create Figure 1.1. The code to recreate the other figures is similar, with only the multinomial probabilities changed to accommodate more trial arms. (Note: to generate the data, use the code in Appendix 2.2.1.)

```
#####################################################
## Create graphics                                 ##
#####################################################
# Global plotting params
f1 = 2 ; c1 = 1.3

# Mean (SE) labels
x11 = bquote(.(pe3.0[1])~" ("~.(se3.0[1])~")")
x12 = bquote(.(pe3.25[1])~" ("~.(se3.25[1])~")")
x13 = bquote(.(pe3.70[1])~" ("~.(se3.70[1])~")")
x21 = bquote(.(pe3.0[2])~" ("~.(se3.0[2])~")")
```

```
x22 = bquote(.(pe3.25[2])~" ("~.(se3.25[2])~")")

x23 = bquote(.(pe3.70[2])~" ("~.(se3.70[2])~")")

x31 = bquote(.(pe3.0[3])~" ("~.(se3.0[3])~")")

x32 = bquote(.(pe3.25[3])~" ("~.(se3.25[3])~")")

x33 = bquote(.(pe3.70[3])~" ("~.(se3.70[3])~")")


# Histogram matrix

par(mfrow=c(3,3),oma=c(0,0,0,0)+0.1,mar=c(4,5,1,1)+0.1)

hist(SS.0[,1],ylab="Random",,main="0% DK",xlim=c(-.5,1),

     font.lab=f1,cex.lab=c1,xlab=x11)

hist(SS.25[,1],ylab="",,main="25% DK",xlim=c(-.5,1),cex.lab=c1,xlab=x12)

hist(SS.70[,1],ylab="",,main="70% DK",xlim=c(-.5,1),cex.lab=c1,xlab=x12)

hist(SS.0[,2],ylab="Unblinded",,main="",xlim=c(-.5,1),

     font.lab=f1,cex.lab=c1,xlab=x21)

hist(SS.25[,2],ylab="",,main="",xlim=c(-.5,1),cex.lab=c1,xlab=x22)

hist(SS.70[,2],ylab="",,main="",xlim=c(-.5,1),cex.lab=c1,xlab=x23)

hist(SS.0[,3],ylab="Opposite",,main="",xlim=c(-.5,1),

     font.lab=f1,cex.lab=c1,xlab=x31)

hist(SS.25[,3],ylab="",,main="",xlim=c(-.5,1),cex.lab=c1,xlab=x32)

hist(SS.70[,3],ylab="",,main="",xlim=c(-.5,1),cex.lab=c1,xlab=x32)
```

APPENDIX C

R2WinBUGS Code for Chapter 2 Simulations

C.1 Two Study Arms

3.1.1 WinBUGS Script

```
model;
{
  # Dirichlet prior on response probabilities
  p1[1:3]~ddirch(alpha1[])
  p2[1:3]~ddirch(alpha2[])

  # Likelihood of response probabilities
  y1[1:3]~dmulti(p1[1:3],n1)
  y2[1:3]~dmulti(p2[1:3],n2)

  # FBIi statistics for each arm (k=2)
  FBI1<-(p1[1]-p1[2])
  FBI2<-(p2[2]-p2[1])

  # Proportion of DK responses
  pDk<-(p1[3]*n1+p2[3]*n2)/N

  # Weighted proportion of observed guesses
  pDo<-(w1[1]*p1[1]*n1/N+w2[1]*p2[1]*n2/N+
        w1[2]*p1[2]*n1/N+w2[2]*p2[2]*n2/N)/(1-pDk)

  # Row proportions
```

```
  p1.<-(p1[1]*n1+p2[1]*n2)/N
  p2.<-(p1[2]*n1+p2[2]*n2)/N

  # Column proportions
  p.1<-(p1[1]*n1+p1[2]*n1)/N
  p.2<-(p2[1]*n2+p2[2]*n2)/N

  # Weighted proportion of expected guesses
  pDe<-(w1[1]*p1.*p.1+w2[1]*p1.*p.2+
        w1[2]*p2.*p.1+w2[2]*p2.*p.2)/((1-pDk)*(1-pDk))

  # Kappa-statistic variant
  kD<-(pDo-pDe)/pDe

  # James's BI
  BI<-1/2*(1+pDk+(1-pDk)*kD)
}

list(alpha1 = c(1, 1, 1), alpha2 = c(1, 1, 1),
y1=c(82, 25, 170),
y2=c(27, 29, 83),
n1=277, n2=139, N=416,
w1=c(0, .5, 1),
w2=c(.5, 0, 1))
```

3.1.2 R2WinBUGS Program

```
#########################################################################
# CASE 1: rr                                                            #
#########################################################################
```

```
# Global parameters
its = 1000
alpha1 = alpha2 = c(1,1,1)
n1 = n2 = 250
w1 = c(0,.5,1)
w2 = c(.5,0,1)
N = n1+n2

#############
## DK = 0% ##
#############
# Set seed
set.seed(11100)

# Simulation objects (posterior point estimates & rejection vectors)
FBI1.post = FBI2.post = BI.post = c()
FBI1.rej  = FBI2.rej  = BI.rej  = c()

# BUGS lists
params = list("BI","FBI1","FBI2")
data   = list("alpha1","alpha2","y1","y2",
              "n1","n2","N","w1","w2")
inits  = list(p1=rep(1/3,3),p2=rep(1/3,3))

# Multinomial probabilities
pA = c(50,50,0)
pB = c(50,50,0)

for(i in 1:its){
```

```
  # Simulate data
  y1 = as.vector(rmultinom(1,n1,pA))
  y2 = as.vector(rmultinom(1,n2,pB))

  # BUGS simulation
  sims = bugs(
    data=data,inits=list(inits),
    parameters.to.save=params,
    model.file="computing_BIs.txt",
    n.chains=1,n.burnin=2000,n.iter=12000,n.thin=1,
    DIC=FALSE)
  Post = sims$sims.matrix

  # Save posterior means
  BI.post[i]   = mean(Post[,1])
  FBI1.post[i] = mean(Post[,2])
  FBI2.post[i] = mean(Post[,3])

  # Save rejection indicator
  if(quantile(Post[,1],.95) < 0.5) {BI.rej[i] = 1
    } else {BI.rej[i] = 0}
  if(quantile(Post[,2],.025)<0 & 0<quantile(Post[,2],.975)) {
    FBI1.rej[i] = 0 } else {FBI1.rej[i] = 1}
  if(quantile(Post[,3],.025)<0 & 0<quantile(Post[,3],.975)) {
    FBI2.rej[i] = 0 } else {FBI2.rej[i] = 1}
}

# List output
case1.0.bayes2 = data.frame(BI.post=BI.post,BI.rej=BI.rej,
```

```
                        FBI1.post=FBI1.post,FBI1.rej=FBI1.rej,
                        FBI2.post=FBI2.post,FBI2.rej=FBI2.rej)
```

C.2 Three Study Arms

3.2.1 WinBUGS Script

```
model;
{
  # Dirichlet prior on response probabilities
  p1[1:4]~ddirch(alpha1[])
  p2[1:4]~ddirch(alpha2[])
  p3[1:4]~ddirch(alpha3[])

  # Likelihood of response probabilities
  y1[1:4]~dmulti(p1[1:4],n1)
  y2[1:4]~dmulti(p2[1:4],n2)
  y3[1:4]~dmulti(p3[1:4],n3)

  # FBIi statistics for each arm (k=3)
  FBI1<-(2*p1[1]-(p1[2]+p1[3]))/2
  FBI2<-(2*p2[2]-(p2[1]+p2[3]))/2
  FBI3<-(2*p3[3]-(p3[1]+p3[2]))/2

  # Proportion of DK responses
  pDk<-(p1[4]*n1+p2[4]*n2+p3[4]*n3)/N

  # Weighted proportion of observed guesses
  pDo<-(w1[1]*p1[1]*n1/N+w2[1]*p2[1]*n2/N+w3[1]*p3[1]*n3/N+
        w1[2]*p1[2]*n1/N+w2[2]*p2[2]*n2/N+w3[2]*p3[2]*n3/N+
        w1[3]*p1[3]*n1/N+w2[3]*p2[3]*n2/N+w3[3]*p3[3]*n3/N)/(1-pDk)
```

```
  # Row proportions
  p1.<-(p1[1]*n1+p2[1]*n2+p3[1]*n3)/N
  p2.<-(p1[2]*n1+p2[2]*n2+p3[2]*n3)/N
  p3.<-(p1[3]*n1+p2[3]*n2+p3[3]*n3)/N

  # Column proportions
  p.1<-(p1[1]*n1+p1[2]*n1+p1[3]*n1)/N
  p.2<-(p2[1]*n2+p2[2]*n2+p2[3]*n2)/N
  p.3<-(p3[1]*n3+p3[2]*n3+p3[3]*n3)/N

  # Weighted proportion of expected guesses
  pDe<-(w1[1]*p1.*p.1+w2[1]*p1.*p.2+w3[1]*p1.*p.3+
        w1[2]*p2.*p.1+w2[2]*p2.*p.2+w3[2]*p2.*p.3+
        w1[3]*p3.*p.1+w2[3]*p3.*p.2+w3[3]*p3.*p.3)/((1-pDk)*(1-pDk))

  # Kappa-statistic variant
  kD<-(pDo-pDe)/pDe

  # James's BI
  BI<-1/2*(1+pDk+(1-pDk)*kD)
}

list(alpha1 = c(1, 1, 1, 1),
alpha2 = c(1, 1, 1, 1),
alpha3 = c(1, 1, 1, 1),
y1=c(41, 66, 30, 44),
y2=c(27, 72, 24, 51),
y3=c(22, 36, 64, 52),
```

```
n1=181, n2=174, n3=174, N=529,
w1=c(0, .5, .75, 1),
w2=c(.5, 0, .75, 1),
w3=c(.75, .75, 0, 1))
```

3.2.2 R2WinBUGS Script

```
#############################################################################
# CASE 1: rrr                                                               #
#############################################################################
# Global parameters
its = 1000
alpha1 = alpha2 = alpha3 = c(1,1,1,1)
n1 = n2 = n3 = 200
w1 = c(0,.5,.75,1)
w2 = c(.5,0,.75,1)
w3 = c(.75,.75,0,1) N = n1+n2+n3

#############
## DK = 0% ##
#############
# Set seed
set.seed(11100)

# Simulation objects (posterior point estimates & rejection vectors)
FBI1.post = FBI2.post = FBI3.post = BI.post = c()
FBI1.rej  = FBI2.rej  = FBI3.rej  = BI.rej  = c()

# BUGS lists
params = list("BI","FBI1","FBI2","FBI3")
```

```
data   = list("alpha1","alpha2","alpha3","y1","y2","y3",
              "n1","n2","n3","N","w1","w2","w3")
inits  = list(p1=c(.25,.25,.25,.25),p2=c(.25,.25,.25,.25),
              p3=c(.25,.25,.25,.25))

# Multinomial probabilities
pA = c(1/3,1/3,1/3,0)
pB = c(1/3,1/3,1/3,0)
pC = c(1/3,1/3,1/3,0)

for(i in 1:its){

  # Simulate data
  y1 = as.vector(rmultinom(1,n1,pA))
  y2 = as.vector(rmultinom(1,n2,pB))
  y3 = as.vector(rmultinom(1,n3,pC))

  # BUGS simulation
  sims = bugs(
    data=data,inits=list(inits),
    parameters.to.save=params,
    model.file="computing_BIs.txt",
    n.chains=1,n.burnin=2000,n.iter=12000,n.thin=1,
    DIC=FALSE)
  Post = sims$sims.matrix

  # Save posterior means
  BI.post[i]   = mean(Post[,1])
```

```
  FBI1.post[i] = mean(Post[,2])
  FBI2.post[i] = mean(Post[,3])
  FBI3.post[i] = mean(Post[,4])

  # Save rejection indicator
  if(quantile(Post[,1],.95) < 0.5) {BI.rej[i] = 1
    } else {BI.rej[i] = 0}
  if(quantile(Post[,2],.025)<0 & 0<quantile(Post[,2],.975)) {
    FBI1.rej[i] = 0 } else {FBI1.rej[i] = 1}
  if(quantile(Post[,3],.025)<0 & 0<quantile(Post[,3],.975)) {
    FBI2.rej[i] = 0 } else {FBI2.rej[i] = 1}
  if(quantile(Post[,4],.025)<0 & 0<quantile(Post[,4],.975)) {
    FBI3.rej[i] = 0 } else {FBI3.rej[i] = 1}
}

# List output
case1.0.bayes3 = data.frame(BI.post=BI.post,BI.rej=BI.rej,
                            FBI1.post=FBI1.post,FBI1.rej=FBI1.rej,
                            FBI2.post=FBI2.post,FBI2.rej=FBI2.rej,
                            FBI3.post=FBI3.post,FBI3.rej=FBI3.rej)
```

APPENDIX D

Convergence Plots for Chapter 2 Simulations

To show convergence was achieved and that thinning was unnecessary, we provide descriptive plots from WinBUGS after random iterations from the simulation study presented in Section ####. We have chosen to use Case 6 from Table #### so that we may consider each of the blinding scenarios considered (random, unblinded, opposite guessing). Thus, FBI_1 represents random responses, FBI_2 represents an unblinded group, and FBI_3 represents a reverse unblinded group. To stay consistent with the simulation work presented in this research, we consider Case 6 for each of the three levels of DK responses reported in Section ####. We use the same simulation parameters as in Section ####: thinning set to 1, a burn-in of $2,000$ and $10,000$ updates. Plots provided include: dynamic trace, time series, kernel density, and autocorrelation function.

D.1 Low DK Response Rate

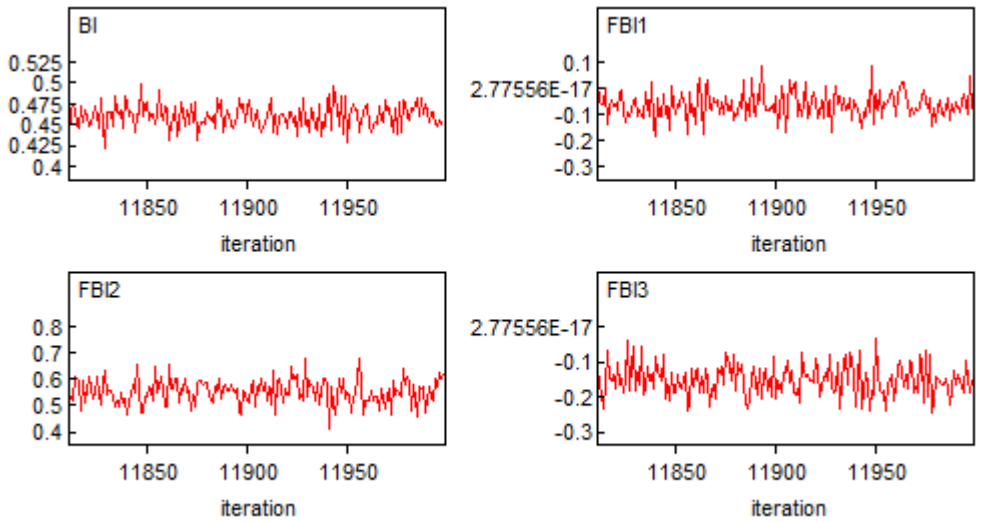

Figure D.1: Dynamic Trace (DK= 0%)

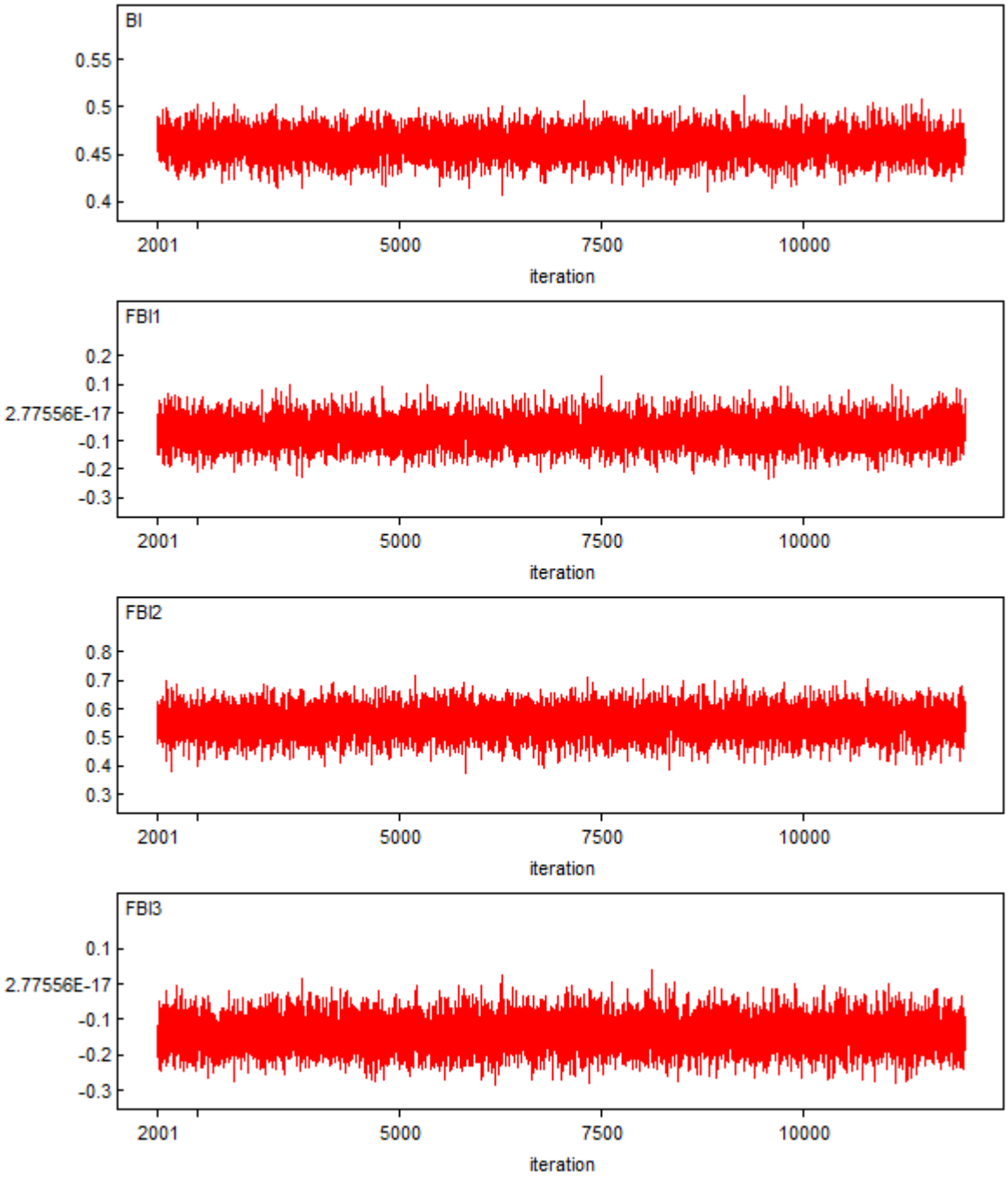

Figure D.2: Time Series (DK= 0%)

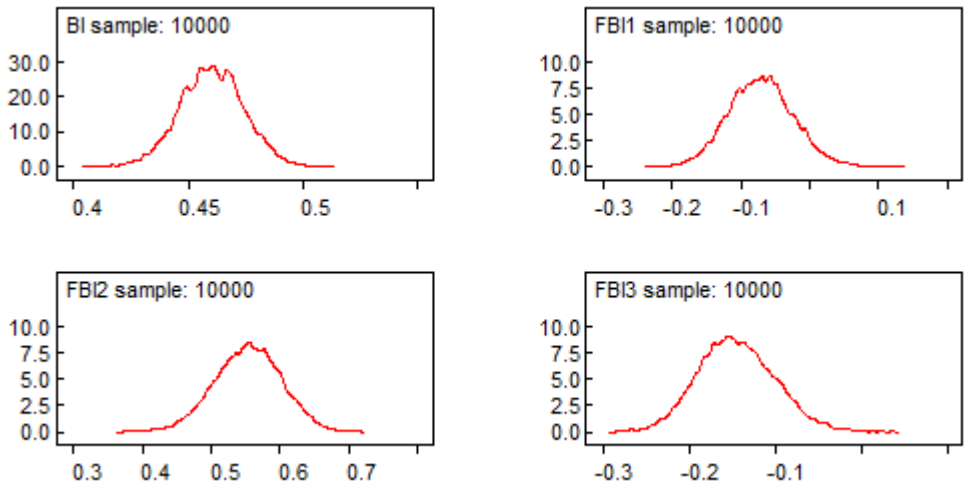

Figure D.3: Kernel Density (DK= 0%)

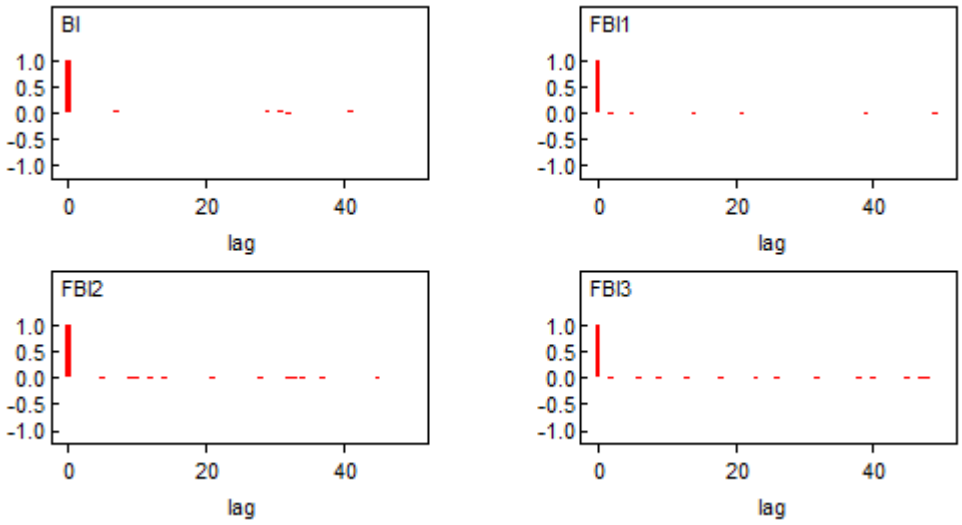

Figure D.4: Autocorrelation Function (DK= 0%)

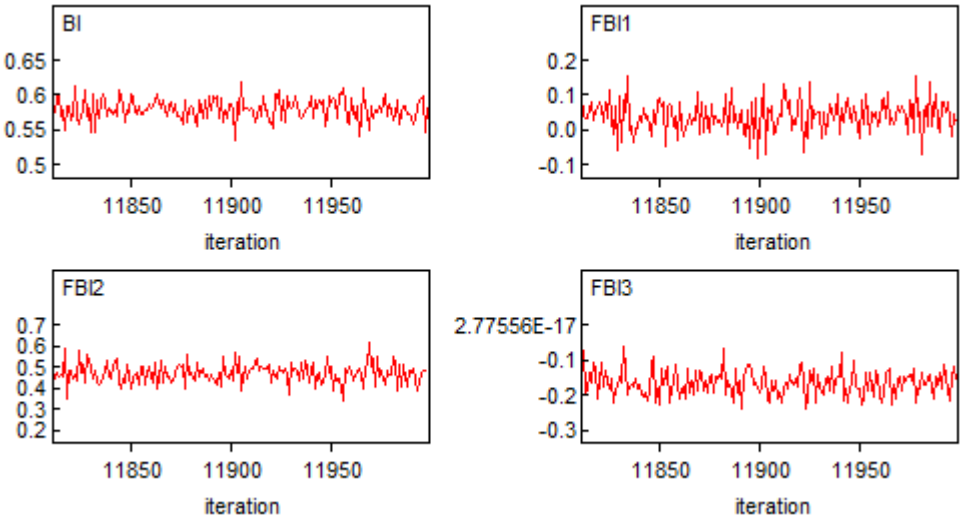

Figure D.5: Dynamic Trace (DK= 25%)

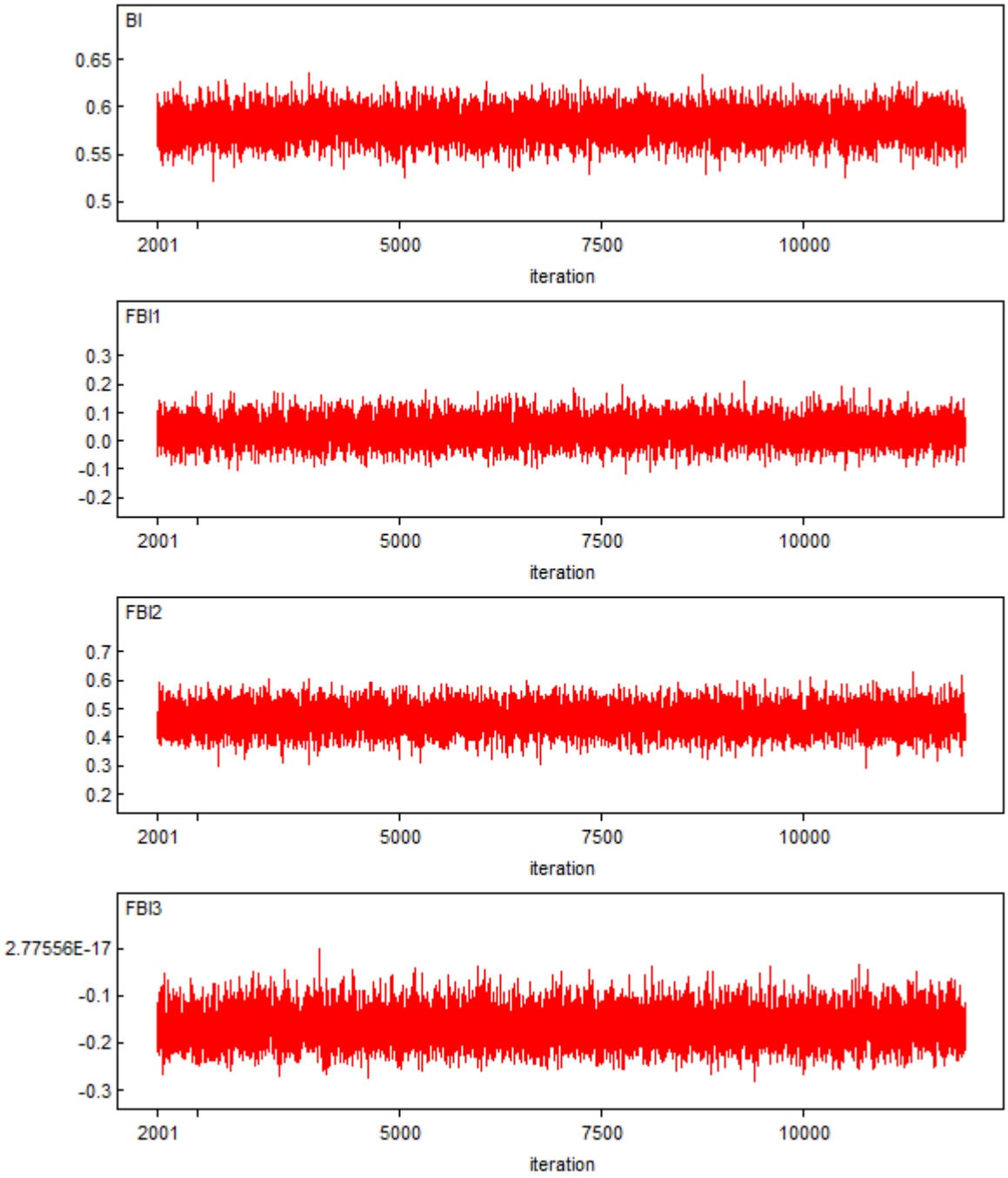

Figure D.6: Time Series (DK= 25%)

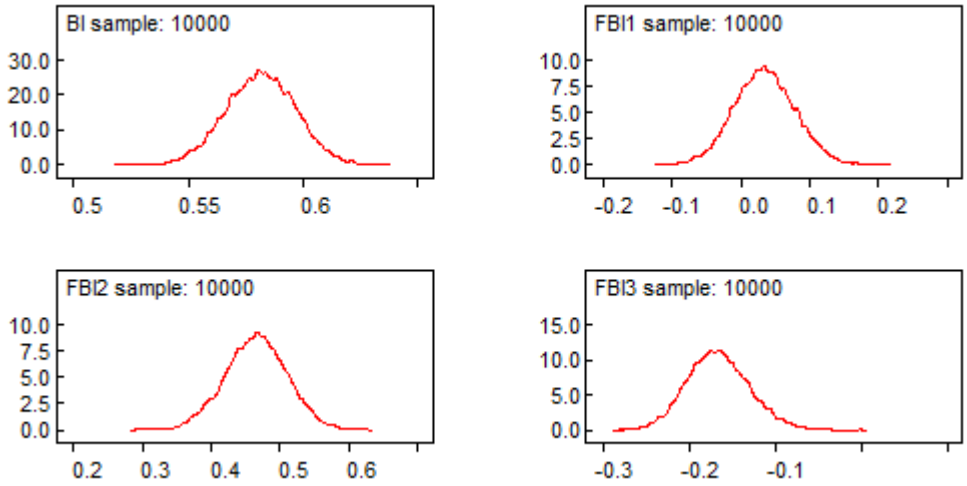

Figure D.7: Kernel Density (DK= 25%)

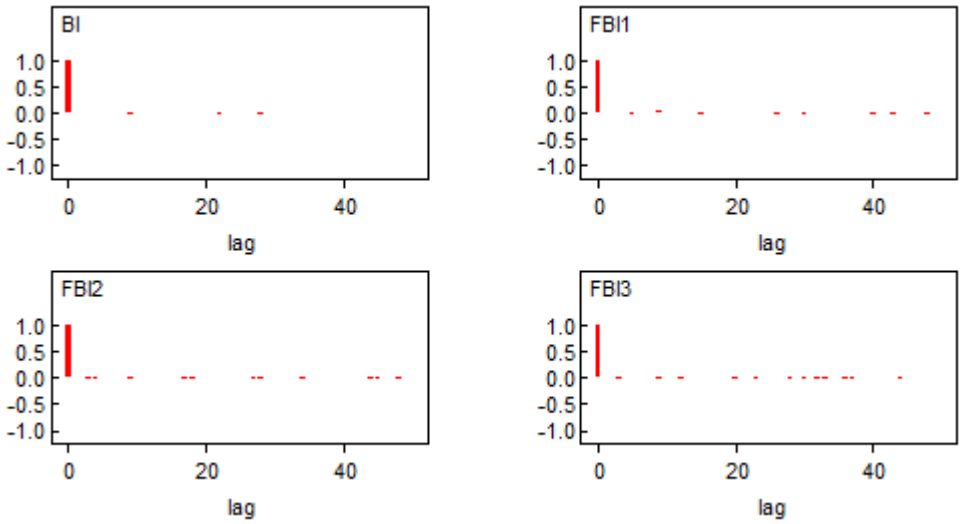

Figure D.8: Autocorrelation Function (DK= 25%)

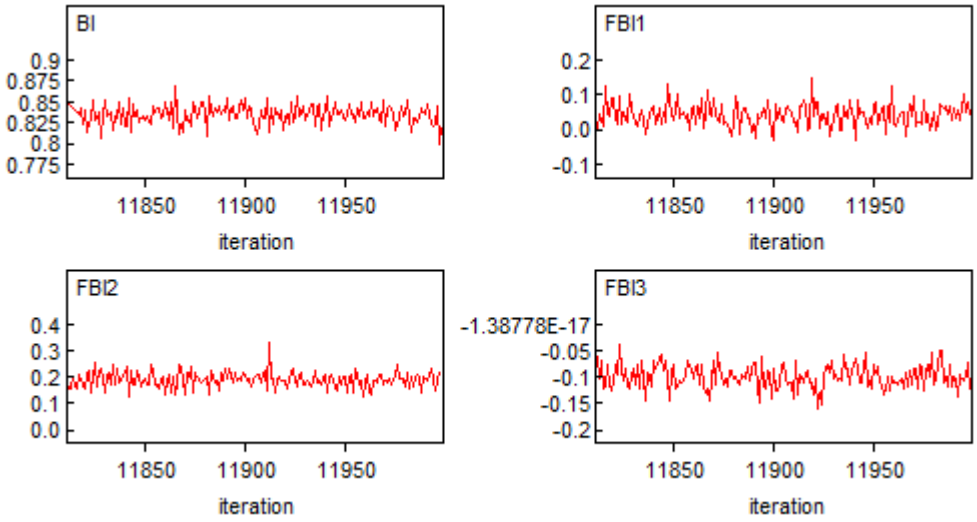

Figure D.9: Dynamic Trace (DK= 70%)

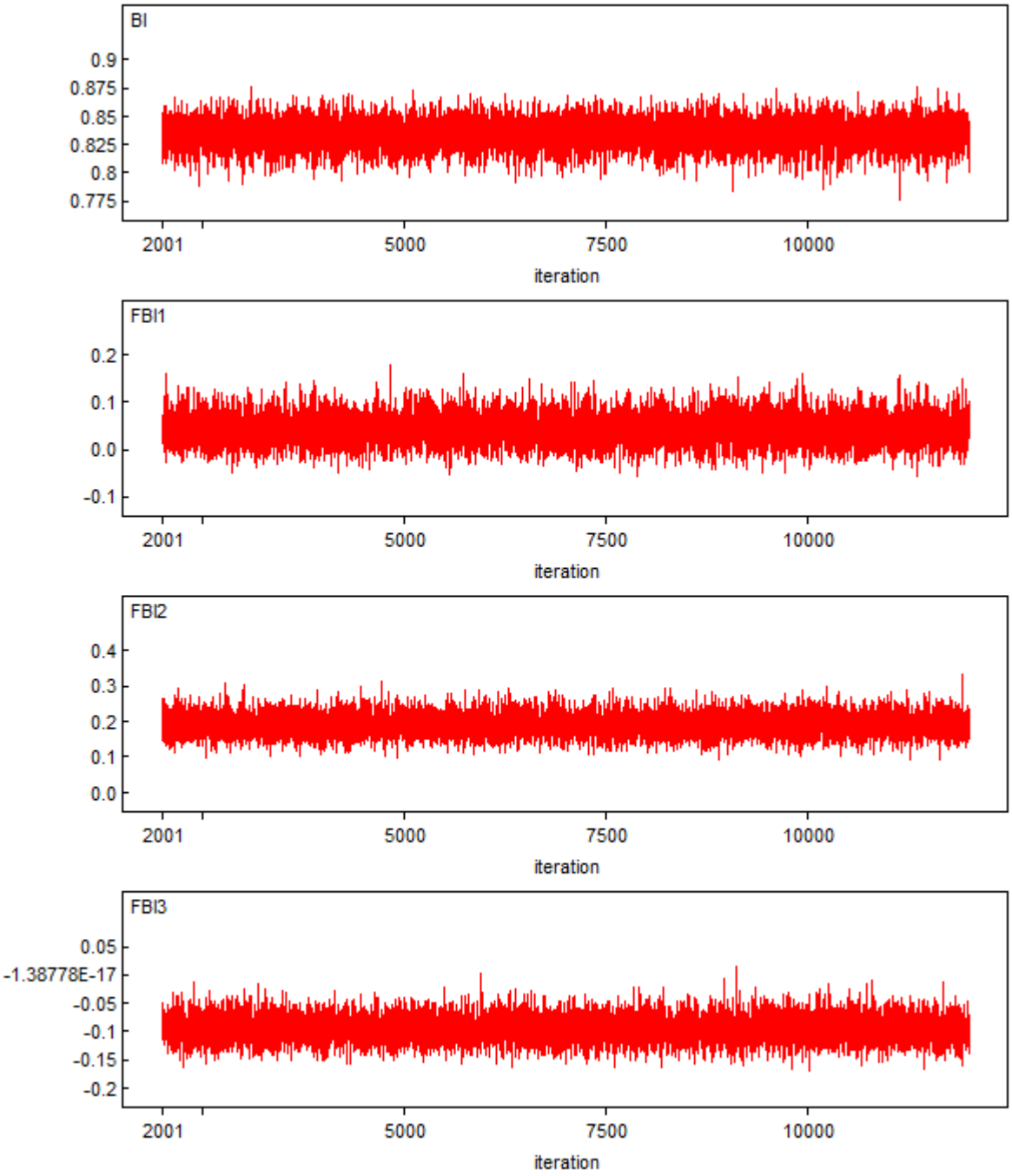

Figure D.10: Time Series (DK= 70%)

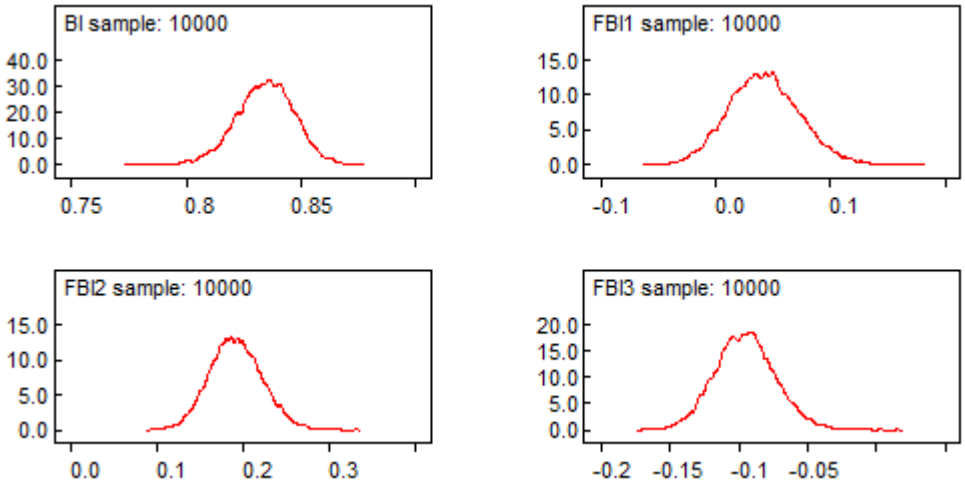

Figure D.11: Kernel Density (DK= 70%)

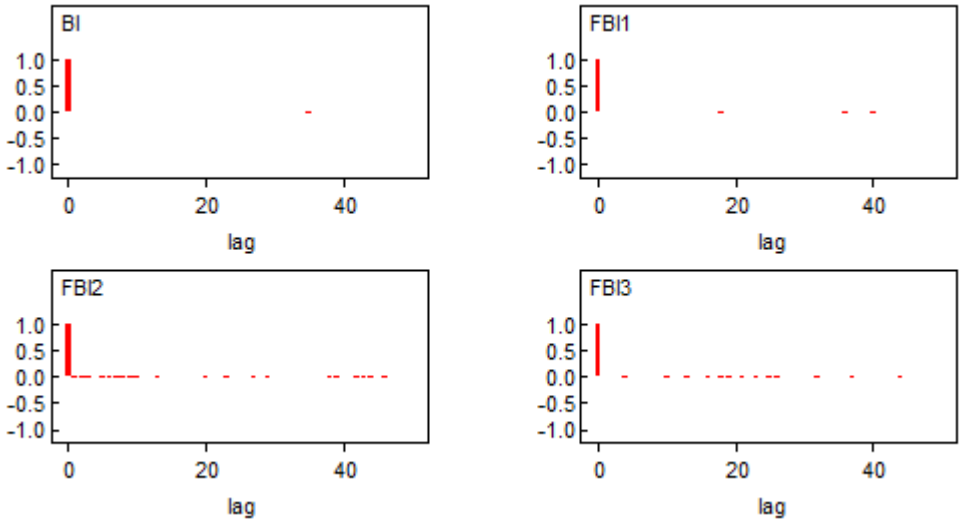

Figure D.12: Autocorrelation Function (DK= 70%)

APPENDIX E

Supplemental Material

E.1 Bang's Published Simulation Results

Table E.1: Reproduced Table of Simulation Results (Bang 2004)

Case	DK (%)	James' BI		FBI/BI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
1	0	0.50 (0.022)	5	A	0.00 (0.063)	7.2
				B	0.00 (0.063)	5.2
	30	0.65 (0.021)	0	A	0.00 (0.053)	5.8
				B	0.00 (0.053)	3.2
	70	0.85 (0.016)	0	A	0.00 (0.034)	4.4
				B	0.00 (0.034)	4.4
2	0	0.90 (0.013)	0	A	-0.80 (0.037)	0
				B	-0.80 (0.038)	0
	30	0.90 (0.013)	100	A	-0.50 (0.042)	0
				B	-0.50 (0.042)	0
	70	0.90 (0.013)	0	A	-0.10 (0.034)	0
				B	-0.10 (0.034)	0
3	0	0.10 (0.013)	100	A	0.80 (0.037)	100
				B	0.80 (0.038)	100
	30	0.40 (0.022)	99.4	A	0.50 (0.042)	100
				B	0.50 (0.042)	100
	70	0.80 (0.018)	0	A	0.10 (0.034)	92.4
				B	0.10 (0.034)	92
4	0	0.50 (0.013)	6.6	A	-0.80 (0.037)	0
				B	0.80 (0.038)	100
	30	0.65 (0.017)	0	A	-0.50 (0.042)	0
				B	0.50 (0.042)	100
	70	0.85 (0.015)	0	A	-0.10 (0.034)	0
				B	0.10 (0.034)	92
5	0	0.70 (0.018)	0	A	0.00 (0.063)	7.2
				B	-0.80 (0.038)	0
	30	0.77 (0.015)	0	A	0.00 (0.053)	5.8
				B	-0.50 (0.042)	0
	70	0.88 (0.019)	0	A	0.00 (0.034)	4.4
				B	-0.10 (0.034)	0
6	0	0.30 (0.019)	100	A	0.00 (0.063)	7.2
				B	0.80 (0.038)	100
	30	0.53 (0.021)	0	A	0.00 (0.053)	5.8
				B	0.50 (0.042)	100
	70	0.83 (0.016)	0	A	0.00 (0.034)	4.4
				B	0.10 (0.034)	92

**Results based on 500 iterations, 1-sided tests performed*

E.2 Bang's Simulations Reproduced

Table E.2: Simulation Results from Bang Cases

Case	DK (%)	James' BI		FBI/BI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
1	0	0.50 (0.023)	5.4	A	0.00 (0.064)	5.6
				B	0.00 (0.063)	4.8
	30	0.65 (0.022)	0	A	0.00 (0.054)	6.4
				B	0.00 (0.054)	5.2
	70	0.85 (0.016)	0	A	0.00 (0.037)	6.3
				B	0.00 (0.034)	3.5
2	0	0.90 (0.013)	0	A	-0.80 (0.038)	100
				B	-0.80 (0.037)	100
	30	0.90 (0.013)	0	A	-0.50 (0.043)	100
				B	-0.50 (0.041)	100
	70	0.90 (0.014)	0	A	-0.10 (0.035)	83.1
				B	-0.10 (0.034)	84
3	0	0.10 (0.13)	100	A	0.80 (0.038)	100
				B	0.80 (0.037)	100
	30	0.40 (0.021)	100	A	0.50 (0.040)	100
				B	0.50 (0.042)	100
	70	0.80 (0.018)	0	A	0.10 (0.035)	82.1
				B	0.10 (0.035)	83.7
4	0	0.50 (0.013)	5	A	-0.80 (0.038)	100
				B	0.80 (0.037)	100
	30	0.65 (0.017)	0	A	-0.50 (0.043)	100
				B	0.50 (0.042)	100
	70	0.85 (0.015)	0	A	-0.10 (0.035)	81.9
				B	0.10 (0.034)	82.4
5	0	0.70 (0.019)	0	A	0.00 (0.066)	6.1
				B	-0.80 (0.038)	100
	30	0.78 (0.018)	0	A	0.00 (0.054)	6.2
				B	-0.50 (0.041)	100
	70	0.87 (0.015)	0	A	0.00 (0.035)	5.4
				B	-0.10 (0.035)	84.4
6	0	0.30 (0.019)	100	A	0.00 (0.066)	6.1
				B	0.80 (0.038)	100
	30	0.52 (0.022)	5	A	0.00 (0.052)	5
				B	0.50 (0.043)	100
	70	0.83 (0.017)	0	A	0.00 (0.036)	5.5
				B	0.10 (0.035)	82.9

E.3 Section 2.2.1 Simulation Results without Multiple Comparisons Correction

Table E.3: Bayes Simulation Results for $k = 2$ Arms

Case	DK (%)	James' BI		FBI/BI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
1	0	0.50 (0.022)	4.4	A	0.00 (0.063)	5.6
				B	0.00 (0.064)	5.9
	30	0.65 (0.022)	0	A	0.00 (0.054)	5.4
				B	0.00 (0.051)	5.3
	70	0.85 (0.016)	0	A	0.00 (0.034)	5.3
				B	0.00 (0.036)	5.3
2	0	0.90 (0.013)	0	A	-0.79 (0.037)	100
				B	-0.79 (0.037)	100
	30	0.90 (0.013)	0	A	-0.49 (0.041)	100
				B	-0.50 (0.043)	100
	70	0.90 (0.014)	0	A	-0.10 (0.034)	82.1
				B	-0.10 (0.033)	83.2
3	0	0.11 (0.013)	100	A	0.79 (0.037)	100
				B	0.79 (0.037)	100
	30	0.40 (0.022)	99.9	A	0.49 (0.043)	100
				B	0.50 (0.043)	100
	70	0.80 (0.017)		A	0.10 (0.033)	82.6
				B	0.10 (0.034)	83.5
4	0	0.50 (0.013)	3	A	-0.79 (0.037)	100
				B	-0.79 (0.037)	100
	30	0.65 (0.016)	0	A	-0.49 (0.043)	100
				B	-0.49 (0.042)	100
	70	0.85 (0.015)	0	A	-0.10 (0.033)	83.8
				B	-0.10 (0.033)	83.9
5	0	0.70 (0.018)	0	A	0.00 (0.061)	3.9
				B	-0.79 (0.037)	100
	30	0.77 (0.017)	0	A	0.00 (0.052)	5.2
				B	-0.49 (0.041)	100
	70	0.87 (0.014)	0	A	0.00 (0.034)	5.1
				B	-0.10 (0.035)	82.3
6	0	0.30 (0.018)	100	A	0.00 (0.061)	3.6
				B	0.79 (0.037)	100
	30	0.53 (0.020)	0.1	A	0.00 (0.050)	4.2
				B	0.49 (0.041)	100
	70	0.80 (0.017)	0	A	0.00 (0.034)	3.7
				B	0.10 (0.034)	83.9

E.4 Section 1.10.1 Simulation Results without Multiple Comparisons Correction

Table E.4: Simulation Results Without Multiple Comparison Correction

Case	DK (%)	James' BI		FBI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
1	0	0.50 (0.015)	4.7	A	0.00 (0.052)	6.5
				B	0.00 (0.050)	4.9
				C	0.00 (0.050)	4.8
	25	0.62 (0.015)	0	A	0.00 (0.045)	7
				B	0.00 (0.043)	4.5
				C	0.00 (0.043)	4.3
	70	0.85 (0.012)	0	A	0.00 (0.027)	4.1
				B	0.00 (0.027)	4.5
				C	0.00 (0.027)	4.9
2	0	0.40 (0.015)	100	A	0.00 (0.050)	5.5
				B	0.00 (0.050)	5.6
				C	0.70 (0.044)	100
	25	0.55 (0.017)	0.4	A	0.00 (0.044)	5.6
				B	0.00 (0.043)	5.3
				C	0.52 (0.043)	100
	70	0.82 (0.014)	0	A	0.00 (0.027)	4.3
				B	0.00 (0.027)	4.9
				C	0.21 (0.032)	100
3	0	0.54 (0.014)	0	A	0.00 (0.051)	4.9
				B	0.00 (0.050)	5.8
				C	-0.20 (0.041)	99.4
	25	0.66 (0.014)	0	A	0.00 (0.042)	4.2
				B	0.00 (0.042)	4.3
				C	-0.15 (0.037)	97.5
	70	0.86 (0.011)	0	A	0.00 (0.027)	4.8
				B	0.00 (0.027)	4.7
				C	-0.06 (0.025)	66.9
4	0	0.28 (0.013)	100	A	0.00 (0.048)	4.9
				B	0.70 (0.043)	100
				C	0.70 (0.044)	100
	25	0.46 (0.018)	77.4	A	0.00 (0.042)	5.3
				B	0.53 (0.043)	100
				C	0.53 (0.044)	100
	70	0.78 (0.016)	0	A	0.00 (0.028)	5.8
				B	0.21 (0.032)	100
				C	0.21 (0.031)	100
5	0	0.57 (0.013)	0	A	0.00 (0.050)	4.5
				B	-0.20 (0.042)	99.7
				C	-0.20 (0.042)	99.2
	25	0.68 (0.014)	0	A	0.00 (0.043)	1.1
				B	-0.15 (0.038)	92.5
				C	-0.15 (0.038)	93.1
	70	0.87 (0.011)	0	A	0.00 (0.027)	5.3
				B	-0.06 (0.024)	69.4
				C	-0.06 (0.024)	67

Table E.4: Simulation Results Without Multiple Comparison Correction (cont'd)

Case	DK (%)	James' BI		FBI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
6	0	0.41 (0.014)	100	A	0.00 (0.051)	5.8
				B	0.70 (0.044)	100
				C	-0.20 (0.042)	99.4
	25	0.56 (0.016)	0	A	0.00 (0.044)	5.2
				B	0.52 (0.042)	100
				C	-0.15 (0.037)	96.5
	70	0.82 (0.014)	0	A	0.00 (0.027)	6
				B	0.21 (0.033)	100
				C	-0.06 (0.024)	69.4
7	0	0.15 (0.012)	100	A	0.70 (0.042)	100
				B	0.70 (0.043)	100
				C	0.70 (0.042)	100
	25	0.36 (0.018)	100	A	0.52 (0.042)	100
				B	0.53 (0.044)	100
				C	0.52 (0.043)	100
	70	0.74 (0.016)	0	A	0.21 (0.032)	100
				B	0.21 (0.032)	100
				C	0.21 (0.032)	100
8	0	0.29 (0.012)	100	A	0.70 (0.042)	100
				B	0.70 (0.042)	100
				C	-0.20 (0.041)	99.4
	25	0.46 (0.017)	67.7	A	0.53 (0.043)	100
				B	0.53 (0.044)	100
				C	-0.15 (0.038)	96.6
	70	0.80 (0.015)	0	A	0.21 (0.033)	100
				B	0.21 (0.034)	100
				C	-0.06 (0.024)	66.7
9	0	0.44 (0.013)	99.7	A	0.70 (0.044)	100
				B	-0.20 (0.043)	99.2
				C	-0.20 (0.042)	99.2
	25	0.60 (0.014)	0	A	0.52 (0.042)	100
				B	-0.15 (0.037)	97.7
				C	-0.15 (0.036)	97.1
	70	0.84 (0.012)	0	A	0.21 (0.033)	100
				B	-0.06 (0.025)	70.9
				C	-0.06 (0.024)	69.8
10	0	0.60 (0.013)	0	A	-0.20 (0.043)	99.4
				B	-0.20 (0.043)	99.5
				C	-0.20 (0.042)	99.5
	25	0.70 (0.013)	0	A	-0.15 (0.037)	97.4
				B	-0.15 (0.037)	96.4
				C	-0.15 (0.036)	96.6
	70	0.88 (0.010)	0	A	-0.06 (0.024)	69.1
				B	-0.06 (0.024)	69.6
				C	-0.06 (0.025)	69.3

E.5 Section 2.2.2 Simulation Results without Multiple Comparisons Correction

Table E.5: Bayes Simulation Results for $k = 3$ Arms

Case	DK (%)	James' BI		FBI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
1	0	0.50 (0.015)	3.3	A	0.00 (0.050)	4.9
				B	0.00 (0.049)	4.7
				C	0.00 (0.050)	5.6
	25	0.62 (0.015)	0	A	0.00 (0.042)	4.5
				B	0.00 (0.043)	5
				C	0.00 (0.044)	5.1
	70	0.85 (0.012)	0	A	0.00 (0.027)	5.4
				B	0.00 (0.028)	4.7
				C	0.00 (0.027)	4.8
2	0	0.36 (0.013)	100	A	0.00 (0.047)	3.9
				B	0.00 (0.051)	6.3
				C	0.68 (0.043)	100
	25	0.52 (0.016)	0.3	A	0.00 (0.044)	5.5
				B	0.00 (0.041)	4.7
				C	0.51 (0.042)	100
	70	0.80 (0.015)	0	A	0.00 (0.027)	4.6
				B	0.00 (0.027)	4.5
				C	0.21 (0.032)	100
3	0	0.54 (0.014)	0	A	0.00 (0.049)	5.2
				B	0.00 (0.049)	5.1
				C	-0.20 (0.042)	98.8
	25	0.66 (0.014)	0	A	0.00 (0.043)	5.6
				B	0.00 (0.041)	4.5
				C	-0.15 (0.037)	96.4
	70	0.84 (0.011)	0	A	0.00 (0.027)	5.6
				B	0.00 (0.027)	4.5
				C	-0.06 (0.024)	65
4	0	0.26 (0.013)	100	A	0.00 (0.049)	5.2
				B	0.69 (0.042)	100
				C	0.69 (0.042)	100
	25	0.44 (0.043)	94.7	A	0.00 (0.043)	4.7
				B	0.52 (0.040)	100
				C	0.51 (0.041)	100
	70	0.77 (0.015)	0	A	0.00 (0.026)	3.8
				B	0.21 (0.030)	100
				C	0.21 (0.030)	100
5	0	0.57 (0.013)	0	A	0.00 (0.050)	0
				B	-0.20 (0.042)	99.3
				C	-0.20 (0.042)	99.2
	25	0.68 (0.014)	0	A	0.00 (0.041)	4.8
				B	-0.15 (0.037)	95.8
				C	-0.15 (0.037)	96.2
	70	0.87 (0.01)	0	A	0.00 (0.027)	5.3
				B	-0.06 (0.024)	65
				C	-0.06 (0.024)	65.4

Table E.5: Bayes Simulation Results for $k = 3$ Arms (cont'd)

Case	DK (%)	James' BI		FBI		
		BI (SEE)	Rejection (%)	Assignment	FBI (SEE)	Rejection (%)
6	0	0.44 (0.013)	99.7	A	0.00 (0.047)	4.4
				B	0.69 (0.041)	100
				C	-0.19 (0.042)	98.9
	25	0.58 (0.016)	0	A	0.00 (0.044)	6.5
				B	0.51 (0.042)	100
				C	-0.15 (0.037)	96
	70	0.83 (0.013)	0	A	0.00 (0.027)	4.5
				B	0.21 (0.031)	100
				C	-0.06 (0.023)	63.6
7	0	0.16 (0.013)	100	A	0.68 (0.042)	100
				B	0.69 (0.042)	100
				C	0.69 (0.042)	100
	25	0.37 (0.018)	100	A	0.51 (0.043)	100
				B	0.51 (0.041)	100
				C	0.51 (0.041)	100
	70	0.74 (0.017)	0	A	0.21 (0.032)	100
				B	0.21 (0.033)	100
				C	0.21 (0.031)	100
8	0	0.34 (0.013)	100	A	0.68 (0.042)	100
				B	0.69 (0.040)	100
				C	-0.20 (0.042)	98.6
	25	0.50 (0.016)	3.3	A	0.52 (0.041)	100
				B	0.52 (0.042)	100
				C	-0.15 (0.037)	95.7
	70	0.80 (0.014)	0	A	0.21 (0.031)	100
				B	0.21 (0.031)	100
				C	-0.06 (0.024)	63.4
9	0	0.47 (0.013)	74	A	0.69 (0.041)	100
				B	-0.20 (0.041)	99.1
				C	-0.20 (0.042)	98.7
	25	0.60 (0.014)	0	A	0.51 (0.042)	100
				B	-0.15 (0.036)	96.1
				C	-0.15 (0.037)	95.7
	70	0.84 (0.012)	0	A	0.21 (0.032)	100
				B	-0.06 (0.024)	64
				C	-0.06 (0.024)	60.8
10	0	0.60 (0.012)	0	A	-0.20 (0.042)	99
				B	-0.20 (0.040)	99
				C	-0.20 (0.041)	99.6
	25	0.70 (0.012)	0	A	-0.15 (0.036)	96.9
				B	-0.15 (0.035)	96
				C	-0.15 (0.035)	97.6
	70	0.87 (0.010)	0	A	-0.06 (0.024)	64.9
				B	-0.06 (0.024)	65
				C	-0.06 (0.024)	65.9

BIBLIOGRAPHY

(2006), *Guidance for Industry: Patient-reported outcome measures: Use in medical product development to support labeling claims*, US Food and Drug Administration.

Bang, H., Flaherty, S. P., Kolahi, J., and Park, J. (2010), "Blinding assessment in clinical trials: A review of statistical methods and a proposal of blinding assessment protocol," *Clinical Research and Regulatory Affairs*, 27, 42–51.

Bang, H., Ni, L., and Davis, C. E. (2004), "Assessment of blinding in clinical trials," *Controlled Clinical Trials*, 25, 143–156.

Bang, H. and Park, J. (2013), "Blinding in clinical trials: A practical approach," *Journal of Alternative and Complementary Medicine*, 19, 367–369.

Byington, R. P., Curb, J. D., and Mattson, M. E. (1985), "Assessment of double-blindness at the conclusion of the Beta-Blocker Heart Attack Trial," *Journal of the American Medical Association*, 253, 1733–1736.

Fuller, R. K., Branchy, L., Brightwell, D., Derman, R. M., Emrick, C. D., and et al. (1986), "Disulfiram treatment of alcoholism: A Veterans Administration cooperative study," *Journal of the American Medical Association*, 256, 1449–1455.

Hrobjartsson, A., Forfang, E., Haahr, M. T., Als-Nielson, B., and Brorson, S. (2007), "Blinded trials taken to the test: an analysis of randomized clinical trials that report tests for the success of blinding," *International Journal of Epidemiology*, 36, 654–663.

Hughes, J. R. and Krahn, D. (1985), "Blindness and the validity of the double-blind procedure," *Journal of Clinical Psychopharmacology*, 5, 138–142.

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-center clinical trial: Disulfiram for alcohol cessation — A VA Cooperative Study," *Statistics in Medicine*, 15, 1421–1434.

Kiecolt-Glaser, J. K., Belury, M. A., Andridge, R., Malarkey, W. B., Hwang, B. S., and Glaser, R. (2012), "Omega-3 supplementation lowers inflammation in healthy middle-aged and older adults: A randomized controlled trial," *Brain, Behavior, and Immunity*, 26, 988–995.

Kolohi, J., Bang, H., and Park, J. (2009), "Towards a proposal for assessment of blinding success in clinical trials: up-to-date review," *Community Dental Oral Epidemiology*, 37, 477–484.

LaRosa, J. C., Applegate, W., Crouse, J. R. r., Hunninghake, D. B., Knopp, R., Eckfeldt, J. H., Davis, C. E., and Gordon, D. J. (1994), "Cholesterol lowering in the elderly: Results of the Cholesterol Reduction in Seniors Program (CRISP) pilot study," *Arch Intern Med*, 154, 529–539.

Margrat, J., Ehlers, A., and Roth, W. T. (1991), "How 'blind' are double-blind studies?" *Journal of Consulting and Clinical Psychology*, 59, 184–187.

Montgomery, S. A. (1999), "Alternatives to placebo-controlled trials in psychiatry," *European Neurpsychopharmacology*, 9, 265–269.

Park, J., Bang, H., and Canette, I. (2008), "Blinding in clinical trials, time to do it better," *Complementary Therapies in Medicine*, 16, 121–123.

Schulz, K. F. and Grimes, D. A. (2002), "Blinding in randomised trials: Hiding who got what," *The Lancet*, 359, 696–700.

Wisner, K. L., Perel, J. M., Peindl, K. S., Hanusa, B. H., FIndling, R. L., and Rapport, D. (2001), "Prevention of recurrent postpartum depression: a randomised clinical trial," *Journal of Clinical Psychiatry*, 62, 82–86.

Zhang, Z., Kotz, R. M., Wang, C., Ruan, S., and Ho, M. (2013), "A causal model for joint evaluation of placebo and treatment-specific effects in clinical trials," *Biometrics*, 69, 318–327.