Published in final edited form as:

Stat Methods Med Res. 2010 December ; 19(6): 559–574. doi:10.1177/0962280209349008.

Bayesian credible intervals for binomial proportions in a single patient trial

Jacob J Oleson

Department of Biostatistics, The University of Iowa, Iowa City, IA, USA

Abstract

Practitioners are often asking if the treatment successfully improved performance. Many times this question is directed towards the outcome of a single individual. In this article, we develop a method to assess the improvement of a single individual who is administered a test of percent correct at pre-treatment and post-treatment. A Bayesian approach is taken where the number correct is modelled as a binomial random variable and the percent correct is set to a beta prior distribution. The first model assumes percent correct at pre-test is equal to the percent correct at post-test and the posterior predictive distribution is used to evaluate the change in the number correct. We subsequently model the proportions correct at pre-test and post-test as unequal. The second model then assumes independent proportions and the third assumes correlated beta distributions for the two proportions. 95% credible intervals are calculated for the various methods for number of correct at post-test given a particular level at pre-test. An example using data from a cochlear implant clinical trial is presented where clinicians recorded percent correct in a consonant-nucleus-consonant test.

1 Introduction

Did the subject perform better after the treatment? This is a common question in practice that is directed towards individual performance rather than an average population performance. The practitioners that consistently ask this critical question are thinking in terms of a single-patient trial or n of 1 trial. The single-patient trial is typically designed as a crossover trial with a treatment measured at multiple points in time. In medical settings they may be comparing multiple medications within a single patient using random allocation. A fully randomised clinical trial is preferred to address most research questions but useful information can still be gathered on single patients. A number of clinical papers on the topic have been written using single-patient trials.^{1–3} A Bayesian structure has also been used to model binary outcomes in single-patient trials.^{4,5}

Often the question of better performance needs to be rephrased into ‘how much better?’ Simplifying the outcome into the dichotomous reject-do-not-reject decision has many flaws.⁶ Knowing the degree of improvement in the patient is more important than simply asking whether or not there was improvement. This is true whether examining single-patient trials or clinical trials. With an appropriate interval estimate we will be able to detect a clinically meaningful degree of improvement per patient.

We address the scenario of individual performance pre-treatment and post-treatment on a score measuring percentage answered correctly. The level of improvement after treatment

describes how effective the treatment was for the individual. This individual level information can, in turn, be used to find how the percentage improvement differed by demographic and explanatory variables. That is, what group of patients saw the most improvement and which groups saw the least amount of improvement?

The percentage correct scores measured pre-treatment and post-treatment can be considered as coming from two correlated binomial distributions. Two correlated binomials arise frequently in practice. The usual scenario with two correlated binomials results in a 2×2 table and counts are focused on the off diagonal elements using McNemar's test. McNemar's procedure has been extended in many ways including an approach with better small sample properties⁷ and with confidence intervals that are informative in this setting.⁸ A comparison of 10 different frequentist-based interval estimation methods for the difference between binomial proportions is available.⁹ From a Bayesian perspective an analysis is presented for 2×2 tables to compute odds ratios and risk¹⁰ and Bayesian confidence intervals in the paired binomial setting.¹¹ Bayesian methods can have frequentist properties for comparing proportions in 2×2 tables and it is recommended using Jeffrey's prior.¹² The aforementioned settings use the 2×2 setting where counts are observed within a contingency table. Situations arise, though, where we cannot place those counts in a 2×2 table since only the marginal totals are known and not the off-diagonal elements needed in procedures such as McNemar's test.

The circumstance described is common in word recognition for speech discrimination. Many speech discrimination tests consist of lists of words that are read to the subject and the subject is to recite the word that he/she believes was said. The percentage of words recited correctly or phonemes recited correctly is the outcome score. Each subject has a natural variability in his/her own responses. If someone were to take the exact same test again, that person would likely produce a different score. Because of this natural variability in subject responses, it is desirable to account for the variability when determining the degree of change in the patient speech discrimination score. Thus, consider each individual score as a random sample of n successive Bernoulli experiences where n is the total number of words given. It can also be assumed that the probability of correctly identifying a word is identical from trial to trial.

Speech discrimination tests are administered at baseline pre-treatment and again post-treatment to inspect how much scores have changed after the treatment. The list of words in a speech discrimination test may be different from pre-test to post-test but dependence obviously exists between the number recited correctly at pre-test and again at post-test. The method most commonly used to address the issue in the speech discrimination literature is a binomial confidence interval approach.^{13,14} These intervals are used by practitioners to determine what they call significant improvement. The idea is to give a range of post-treatment scores that are within reasonable variation of the pre-treatment scores. Given a particular score at pretest, if the second score is within the confidence interval, then the subject is determined to have had no significant change. A second score above the upper limit of the confidence interval signifies improvement in the speech score. Likewise, a score below the lower limit of the confidence interval indicates a significant decrease in the speech score. The confidence intervals assume a binomial distribution for the number of words correct and use an arcsin transformation¹⁵ to construct a confidence interval.

There are some flaws with this approach. First, a confidence interval gives the degree of change and not the statistical significance of a p -value.⁶ The confidence intervals can be used more effectively than just determining statistical significance. Secondly, the Thornton and Raffin approach used the arcsin transformation to approximate normality in computing the confidence intervals. Obviously with today's computing power, such a transformation

should no longer be required so an attempt was made to update the intervals.¹⁴ Pre-test values and post-test values were simulated from a binomial distribution. From those values it was computed how many post-test values were within 1.96 standard deviations of the pre-test scores. Although this approach uses modern computing to sample from binomial distributions, it assumes a normal distribution for the confidence intervals. The normal approximations could be misleading for intervals near the probability extremes of 0 and 1 and also if the word list is short, say less than 30. The approach also ignores the variability associated with the post-test value and only allows one type of variability.

In this article we propose a new approach to computing and using such intervals. Like the previous work, we too assume that the word count follows a binomial distribution but we avoid approximations to the normal distribution. Clearly the sample size is fixed and given by the number of words. There is a question on whether the trials (words) can all be considered independent. We assume that the probability of correctly answering a word is constant from one word to the next. The words are meant to be a representative sample of real world words and the learning effect should be minimal, making the independence assumption realistic.

We take a Bayesian approach and examine the distribution on the proportion correct p for a fixed sample size n . A key issue with computing such intervals is how to incorporate knowledge of the relationships between pre-test values and post-test values. We prefer to think of the problem hierarchically, which lends itself to the Bayesian paradigm. It is easier to build correlation structure and prior knowledge into a model hierarchically as well as being computationally easier. In the end, hierarchical Bayesian models are presented, which can easily handle the addition of covariates and confounding variables.

We show three approaches that could be taken for computing the individual Bayesian credible intervals. While most of the paper focuses on the single subject design, we also extend the idea to compare treatment groups if there are enough subjects for such a comparison. The first approach, presented in Section 2.1, is the usual Beta-Binomial model that assumes the pre- and post-tests share a common percent correct. Given the score from pre-test we provide a credible interval for where to expect the post-test scores to lie if the proportions were equal. A second approach, outlined in Section 2.2, assumes that the proportions are unequal and we show how to find the probability that the post-test percent correct is higher than the pre-test percent correct. The unequal proportions are assumed correlated in Section 2.3. Finally, we take a hierarchical Bayesian approach to account for covariates in determining if post-test scores are higher than pre-test scores. This is accomplished in Section 2.4 via correlated proportions. A simulation study is presented in Section 3.1 to compare the aforementioned methods which is followed by a data analysis on speech recognition scores in Section 3.2. Section 4 concludes the article discussing the proposed technique and recapping the results.

2 Methods

2.1 Equal proportions

Assume that π_1 and π_2 are the true proportions of words recited correctly by an individual at pre-test and post-test, respectively. Let Y_1 be the observed number correct at pre-test and let Y_2 be the observed number correct at post-test. Previous approaches^{13,14} were developed assuming the Y_1 value was known and created a confidence interval for the Y_2 value. The reported confidence intervals were given for the proportion correct which is simply Y_2/n . Assume n is a fixed sample size for number of items tested. Furthermore, assume for the data model that Y_1 and Y_2 are both random and follow binomial distributions with the same probability of correct response $\pi_1 = \pi_2 = \pi$:

$$\begin{aligned} Y_1|\pi &\sim \text{Binomial}(n, \pi) \\ Y_2|\pi &\sim \text{Binomial}(n, \pi). \end{aligned} \quad (1)$$

From a Bayesian perspective a prior distribution needs to be selected for π . The beta distribution is a natural choice to model the probability correct π because it is a conjugate prior and bound by 0 and 1. Thus, $\pi \sim \text{Be}(a, b)$. The posterior distribution $\pi | Y_1$ is then

$$\pi|Y_1 \sim \text{Be}(Y_1+a, n-Y_1+b). \quad (2)$$

Bayesian inference is often framed in terms of the parameters from the posterior distribution. Consider the scenario before the post-treatment values, Y_2 , are recorded. Rather than focusing on the parameter π we can make inference on the still unknown but observable Y_2 . The inference is what to expect Y_2 to be after observing the value of Y_1 . The posterior predictive distribution of $Y_2 | Y_1$ is then found by integrating out the percentage correct π . Thus, a distribution for the yet unobserved Y_2 given the likelihood is obtained based on a particular value of Y_1 . The predictive distribution is formulated as

$$\begin{aligned} Y_2|Y_1 &= \int_{\pi} (Y_2|\pi)(\pi|Y_1)d\pi \\ &= \binom{n}{Y_2} \frac{\text{Be}(Y_1+Y_2+a, 2n-(Y_1+Y_2)+b)}{\text{Be}(Y_1+a, n-Y_1+b)} \end{aligned} \quad (3)$$

where $\text{Be}(x, y)$ is the beta function $\text{Be}(x, y) = \Gamma(x) \Gamma(y) / \Gamma(x+y)$. In this way, we develop what the most likely values of Y_2 should be assuming Y_1 and Y_2 are generated from the same distribution. If Y_2 is not within the confidence interval generated based on the observed value of Y_1 then it is conjectured that there is likely a difference due to treatment.

From the posterior predictive distribution a 95% credible interval for Y_2 given a specified value of Y_1 is easily computed. This process assumes that Y_1 and Y_2 share a common success probability π . A confidence band is constructed for Y_2 concerning every possible value of Y_1 ($0 - n$). The coverage of the intervals will be slightly unpredictable due to the discreteness of the binomial distribution.¹⁶ The lower bound of each credible interval is the whole value of Y_2 that is above 2.5% of the cumulative predictive distribution and the upper bound is the whole value of Y_2 that is immediately above 97.5% or n , whichever is smaller. If the observed value of Y_2 is outside of the 95% credible interval then there is evidence that the treatment has made an impact and we can measure the size of said impact. This approach uses the original distributions and does not rely on the normal distribution to approximate the interval. It will also work for any sample size where it should be particularly beneficial for small n and π near the endpoints of 0 and 1.

The last part of this Bayesian method is to choose the hyperparameters a and b for the $\text{Be}(a, b)$ distribution. If a and b are equal then the beta distribution is symmetric around 0.5. The choice of the prior for the beta distribution does have an impact on the width of the interval. We examined multiple non-informative priors and their respective impacts on the results. The improper prior $a = b = 0$ is not used because it yields an improper posterior at $\pi = 0$ and at $\pi = 1$. Since those endpoints are of critical interest other values were chosen. Two other common non-informative priors for the beta distribution are the uniform, $a = b = 1$, and Jeffrey's prior, $a = b = 0.5$. The posterior results were similar between the two priors but often lead to slight differences in the endpoints of the credible interval. The credible intervals using Jeffrey's prior, $a = b = 0.5$, are reported for $n = 10, 25, 50, 100$ in Tables 1

and 2 because they generally lead to narrower intervals and have better frequentist properties.¹² The intervals in the tables are listed in the column labelled $\pi_1 = \pi_2$.

2.2 Unequal proportions

The method presented in Section 2.1 assumed that the proportions π_1 and π_2 at pre-test and post-test, respectively, are equal. Another approach is to focus more on the distributions of the probabilities themselves when $\pi_1 \neq \pi_2$. Begin with the likelihood where

$$\begin{aligned} Y_1 | \pi_1 &\sim \text{Binomial}(n, \pi_1) \\ Y_2 | \pi_2 &\sim \text{Binomial}(n, \pi_2). \end{aligned} \quad (4)$$

Assume beta distribution priors for the proportion correct such that $\pi_1 \sim \text{Be}(a_1, b_1)$ and $\pi_2 \sim \text{Be}(a_2, b_2)$. Treating π_1 and π_2 as independent we would write the posterior distribution $\pi_1 | Y_1$ as $\text{Be}(Y_1 + a_1, n - Y_1 + b_1)$ and $\pi_2 | Y_2$ as $\text{Be}(Y_2 + a_2, n - Y_2 + b_2)$. If π_1 and π_2 are treated as independent then the posterior distribution of interest is simply

$$\begin{aligned} \pi_1, \pi_2 | Y_1, Y_2 &= [\pi_1 | Y_1] [\pi_2 | Y_2] \\ &= \frac{\pi_1^{Y_1+a_1-1} (1-\pi_1)^{n-Y_1+b_1-1}}{\text{Be}(Y_1+a_1, n-Y_1+b_1)} * \frac{\pi_2^{Y_2+a_2-1} (1-\pi_2)^{n-Y_2+b_2-1}}{\text{Be}(Y_2+a_2, n-Y_2+b_2)}. \end{aligned} \quad (5)$$

In this setting the predictive distribution $Y_2 | Y_1$ is not of interest since π_1 and π_2 are independent. Instead, we can find the probability that the post-test score is greater than the pre-test score, $\Pr(\pi_2 > \pi_1)$.

For each individual calculate $\Pr(\pi_2 > \pi_1)$ which is the probability that the post-test score is higher than the pre-test score by integrating (5) as

$$\Pr(\pi_2 > \pi_1) = \int_0^1 \int_0^{\pi_2} \frac{\pi_1^{Y_1+a_1-1} (1-\pi_1)^{n-Y_1+b_1-1}}{\text{Be}(Y_1+a_1, n-Y_1+b_1)} * \frac{\pi_2^{Y_2+a_2-1} (1-\pi_2)^{n-Y_2+b_2-1}}{\text{Be}(Y_2+a_2, n-Y_2+b_2)} d\pi_1 d\pi_2. \quad (6)$$

We assumed that $a_1 = a_2$ and $b_1 = b_2$. This integral was solved using Mathematica and choosing various values of Y_1 , Y_2 , a , b and n to find the requested probability. The probability is calculated for $n = 10, 25, 50, 100$ and Jeffrey's prior of $a = b = 0.5$. By cycling through the values of Y_2 we find the smallest value of Y_2 such that $\Pr(\pi_2 > \pi_1) > 0.025$ and the smallest value of Y_2 giving $\Pr(\pi_2 > \pi_1) > 0.975$ for every value of Y_1 . Those Y_2 results construct 95% credible intervals for Y_2 and are given in Tables 1 and 2 in the column labelled $\pi_1 \neq \pi_2$. Note that the intervals from the equal proportion method are symmetric. That is, if the interval for Y_2 when $Y_1 = k$ is (ll, ul) then the interval when $Y_1 = n - k$ is $(n - ul, n - ll)$. For example, if $n = 10$ and the interval for Y_2 at $Y_1 = 2$ is $(0, 7)$ then the interval for Y_2 at $Y_1 = 8$ is $(3, 10)$. This phenomenon is not true for the $\pi_1 \neq \pi_2$ case. In the $\pi_1 = \pi_2$ case there is one distribution and the values follow the well known characteristics of the binomial distribution. In the case of $\pi_1 \neq \pi_2$ though, there are two distributions and it should not be expected that the symmetry characteristics follow.

2.3 Correlated proportions

Assume that Y_1 and Y_2 are as specified in Equation (4). Recall the previous approach in Section 2.2 assumes that Y_1 is independent of Y_2 and that π_1 is independent of π_2 . The approach can be modified to account for the correlation within paired samples. This modification results in a more hierarchical approach to the problem.

Again let the beta distribution be the prior distribution on π_1 and π_2 . Then define

$$\pi_1 = \frac{U}{U+W} \quad \text{and} \quad \pi_2 = \frac{V}{V+W} \quad (7)$$

where U, V, W are *iid Gamma* random variables with respective shape parameters a, b, c . The scale parameters for U, V and W should be identical, assuming they are equal to one. It follows¹⁷ that $\pi_1 \sim \text{Be}(a, c)$ and $\pi_2 \sim \text{Be}(b, c)$. More importantly π_1 and π_2 are correlated. The correlation results from the choice of a, b and c .¹⁸ For given values of a and b , the larger the value of c is, the smaller the correlation between π_1 and π_2 will be. If $a = b$ then π_1 and π_2 follow the same distribution. Assume a, b and c are all fixed. The non-informative Jeffrey's prior for both π_1 and π_2 , such that a, b and c all equal 0.5, yields a correlation between π_1 and π_2 of approximately 0.46. If a different correlation is required then a, b and c can be specified to match the required correlation and obtain appropriate beta distribution priors.

Any of the values of a, b and c could be chosen to be random variables. In this type of single subject design there is often not enough data to obtain a reliable posterior estimate of the value. If more knowledge is known of the correlation then the intervals can be adjusted accordingly through judicious choices of a, b and c . Again, the hyperparameter choices are essential to the endpoints of the confidence interval, particularly for $n = 10$ and where Y_1 or Y_2 are near 0 or n .

This approach is more complicated to compute directly but is ideally suited for a hierarchical Bayes setting. The hierarchical model is easily set up in WinBUGS or programmed directly using another language such as R (<http://www.r-project.org>). To evaluate the posterior distribution of $[\pi_1, \pi_2 | \cdot]$ we use Markov chain Monte Carlo (MCMC) simulations and slice sampling.¹⁹ The joint full conditional distribution of π_1 and π_2 is

$$[\pi_1, \pi_2 | \cdot] \propto \frac{\pi_1^{Y_1+a-1} (1-\pi_1)^{n-Y_1+b+c-1} \pi_2^{Y_2+b-1} (1-\pi_2)^{n-Y_2+a+c-1}}{(1-\pi_1\pi_2)^{a+b+c}}. \quad (8)$$

The posterior distribution was approximated for every combination of Y_1 and Y_2 for $n = 10, 25, 50, 100$. Visual inspection confirmed immediate convergence of the chains as the last 10,000 of 11,000 draws were kept in every case.

Since π_1 and π_2 are not equal, we find the $\Pr(\pi_2 > \pi_1)$ for every combination of Y_1 and Y_2 for each given n and use Monte Carlo integration to approximate said probability. This is done by recording whether or not π_2 is greater than π_1 at each iteration of the Gibbs cycle. Divide the number of greater than occurrences by the length of the chain to obtain the estimate of $\Pr(\pi_2 > \pi_1)$. Again, by cycling through the values of Y_2 for different Y_1 values we find the smallest value of Y_2 such that the estimated probability is greater than 0.025 and the smallest value of Y_2 that gives the estimated probability is greater than 0.975. Those values construct a 95% credible interval for Y_2 given Y_1 and are given in Tables 1 and 2 in the column 'Correlated π 's'.

2.4 Group comparisons

The methods presented farso, all reflect a form of single-patient trials. Suppose the interest is not only to examine individual scores but also to adjust for covariate effects. The data model in this setting is still binomial and is demonstrated using separate $\pi_1 = (\pi_{11}, \dots, \pi_{1m})'$ and $\pi_2 = (\pi_{21}, \dots, \pi_{2m})'$ values that are subject specific. Suppose

$$\begin{aligned} Y_{1i}|\pi_{1i} &\sim \text{Binomial}(n, \pi_{1i}) \\ Y_{2i}|\pi_{2i} &\sim \text{Binomial}(n, \pi_{2i}) \end{aligned} \quad (9)$$

with $i = 1, \dots, m$ where m is the number of subjects. Each individual is allowed their own proportions, π_{1i} and π_{2i} , in the spirit of the single patient trial.

To include covariates we choose logistic models on π_{1i} and π_{2i} such that

$$\begin{aligned} \log\left(\frac{\pi_{1i}}{1-\pi_{1i}}\right) &= \mu_1 + \theta_i + \mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_{1i} \\ \log\left(\frac{\pi_{2i}}{1-\pi_{2i}}\right) &= \mu_2 + \theta_i + \mathbf{X}_i' \boldsymbol{\beta} + \varepsilon_{2i}. \end{aligned} \quad (10)$$

Thus, the $\mu = (\mu_1, \mu_2)'$ parameter designates the overall pre-treatment and post-treatment group mean effects while θ_i represents individual specific random subject effects where $i = 1, \dots, m$. The $\boldsymbol{\beta}$ vector captures any covariate effects with ε_{1i} and ε_{2i} encapsulating any additional random error. The design matrix $\mathbf{X}$ contains the covariate level data. This type of model allows for adjustment of covariates and estimation of the correlation between pre-treatment and post-treatment scores. The correlation, ρ , could be modelled by assuming a multivariate normal distribution on μ . Rather than imposing additional parameters on the model, however, we calculate the correlation between π_1 and π_2 at each iteration of the MCMC simulations.

Flat priors are chosen for μ_1, μ_2 and $\boldsymbol{\beta}$ with a normal prior for θ_i . The error terms, ε_{1i} and ε_{2i} , are $N(0, \tau)$ where τ follows an inverse gamma distribution. This format allows demographic and explanatory variables in the model to find how the proportions are impacted by those variables. Each individual probability, $\Pr(\pi_{2i} > \pi_{1i})$, can be estimated using Monte Carlo integration as described in Section 2.3 giving a probability of improvement for each individual conditioned on selected controlling variables.

3 Data analysis

3.1 Simulation study

Ten-thousand pairs of Y_1 and Y_2 values were randomly sampled at each level of $\pi = 0.1, 0.2, \dots, 0.9$ assuming Y_1 and Y_2 share the same value of π . The simulations were run for $n = 10, 25, 50, 100$ for the intervals presented in Sections 2.1, 2.2, 2.3 as well as the Thornton and Raffin intervals. Under each sample size, the percent of Y_2 values falling outside of the intervals corresponding to its paired Y_1 value were recorded and presented in Table 3. The error percentage should be approximately 5% for the 95% intervals.¹⁶ The values in Table 3 are aggregated across the levels of π to observe the overall error rate. The Thornton-Raffin method produces the narrowest intervals which corresponds to the highest error rate for each sample size. Error rates being greater than 0.05 is an indication that they are perhaps too narrow. The $\pi_1 = \pi_2$ error rates are nearly identical to the correlated π 's error rate while the $\pi_1 \neq \pi_2$ error rates tend to be the highest of the three proposed Bayesian intervals but still less than target rate of 0.05.

It is also important to examine how the error rates function at specific levels of π . Ten thousand pairs of Y_1 and Y_2 were sampled at each of nine different values of π . For each simulation the value of Y_1 was recorded and its corresponding Y_2 value compared with the interval. Some results are presented in Figure 1 for $n = 10$. There are four figures reflecting results when $\pi = 0.2, 0.4, 0.6, 0.8$. On the x axis are the sampled Y_1 values and on the y axis are the Y_2 error rates at that level of Y_1 . Within each figure are four lines, one for each method of confidence interval. The error rates are lowest when Y_1 is near $n * \pi$. The further

Y_1 is from $n * \pi$ the higher the error rate becomes. The Thornton–Raffin intervals typically increase most rapidly away from $n * \pi$.

Many adjustments can be made to the intervals. The priors for the presented intervals are non-informative. Informative priors can shift the intervals so that they are not symmetric and wider/narrower at higher percentages.

3.2 Hybrid 10 Cochlear implant clinical trial

Cochlear implants (CIs) have been designed to assist individuals who are severely to profoundly deaf with verbal communication. The device receives acoustic signals through an externally worn microphone. These signals are processed to filter and transmit those components of sound critically important for speech perception. From there, those components are transmitted via electrical signals to an array of electrodes in the cochlea, resulting in electrical stimulation of the auditory nerve. The central auditory pathway then receives the signal for interpretation which does not produce an exact replica of normal hearing. However, when using modern CIs the majority of post-lingually deafened implant recipients score above 80% on high-context sentences in quiet listening conditions, even without visual cues.²⁰

Investigators at the University of Iowa have been researching a way of improving high frequency speech sounds.^{21–24} The Iowa/Nucleus 10 mm hybrid cochlear implant is under an FDA clinical trial IDE. Eighty-seven adults with severe to profound hearing loss of about 2000 Hz were enrolled in the FDA clinical trial in the United States and implanted with the Iowa/Nucleus 10mm hybrid implant. The poorer hearing ear received the device. Initial implantations took place in 1999 and enrollment in the clinical trial was closed as of 2008. Longitudinal follow-ups are still being administered. Before surgery, each subject was administered the consonant-nucleus-consonant (CNC) word score test. Two lists of 50 monosyllabic words were administered. Each word is digitally recorded and the recording is played to the patient. The patient is to repeat the word back. The number of words repeated correctly was recorded and used to report the proportion correct.

The clinical trial is a longitudinal study where CNC word scores were collected pre-operatively and again post-operatively. As of November 2008, 68 subjects had follow-up values out to at least 1 year. Subjects were not tested at both 9 and 12 months and for the purposes of this example we use the available 9- and 12-month scores to be our post-test value.

One goal of this study is to determine what proportion of the patients improved and how much they improved by. This summary information was broken down by explanatory variables including age of the patient at which the surgery took place and age at which the patient became profoundly deaf. The time between those two ages is known as duration of deafness. Duration of deafness tends to have the most impact in addressing amount of change in CNC score from pre-implant to post-implant.²⁵ It is potentially of interest to find an exclusion criterion based upon patient age or duration of deafness for when it is no longer productive to implant a hybrid device.

Given the pre-test score for our 68 subjects we compare their post-test score with the credible intervals in Table 2. The individual scores are presented in Table 4 if the increase from pre-test to post-test was no more than 25. For the 35 subjects who had an increase of more than 25 words correct their post-test score is outside of the interval. In Table 4 ‘Yes’ is marked in the column if the individual’s post-test score is within the interval for their pre-test score, otherwise ‘No’ is marked when the post-test score is outside the interval. We find that 18 of the 68 (26%) subjects have not increased their scores by 12 months under all three

approaches. The Thornton–Raffin intervals would yield similar results in this sample except that the post-test score of ID 52 would have also been outside of the interval. In this data analysis there were only eighteen subjects who did not show significant improvement with six of them performing the same or worse. This indicates that the hybrid implant is doing well with consonant recognition in adults.²⁵

It is also critical to examine covariates to help inform why some individuals are not improving. The analysis uses the hierarchical approach from Section 2.4 to estimate $\Pr(\pi_{2i} > \pi_{1i})$ for each individual subject conditional on duration of deafness. Convergence was met quickly and the chain was run for 11,000 iterations in WinBUGS with the first 1000 iterations used as burn-in.

The estimated posterior mean of the duration of deafness variable, β , was -0.004 with a 95% credible interval $(-0.017, 0.008)$. Although the interval was not completely below zero we did see an indication that as duration increases the CNC score should be expected to decrease. The value of β implies that the mean percentage correct will decrease by 0.4% for every additional year of deafness.

The estimates of the mean parameters were $\hat{\mu}_1 = -0.62$ and $\hat{\mu}_2 = 0.64$. The random subject effects, θ_i , ranged from -0.90 to 1.10 with a variance of 0.37 . The estimated model variance was $\hat{\tau} = 0.46$. The correlation between π_1 and π_2 was also calculated. Note that the Pearson correlation between Y_1 and Y_2 was 0.44 and the model estimate was 0.45 with a 95% credible interval $(0.38, 0.53)$.

In addition the model also allows for a subject-level comparison between pre-test and post-test scores while controlling for covariates. The last column in Table 4 labelled Group reflects the estimated $\Pr(\pi_{2i} > \pi_{1i})$. If the probability is greater than 0.975 then we would have evidence of improvement which matches ‘No’ in the previous columns. Any values less than 0.975 would correspond to Y_2 being within the intervals. The results after adjusting for duration of deafness exhibit some minor differences from the previous methods. Subjects 10 and 67 were within the limits using the previous methods whereas the method adjusting for duration of deafness suggests an improvement in CNC scores for these two individuals. The two cases are similar to each other with both having a 13% change. ID 10 had a very small duration of deafness while ID 67 had an average duration of deafness. While a change in 13 words was outside the range for those two individuals, a change in 14 words was within the range for ID 52. This individual had a larger than average duration of deafness.

4 Discussion

Being able to determine whether a patient significantly improved or not is important to practitioners. In this article, we produced intervals for pre-post proportion correct testing using a method that does not rely on normal distribution approximations. The narrowest intervals tend to be the $\pi_1 = \pi_2$ approach outlined in Section 2.1. The unequal π approach in Section 2.2 is more general and the inherent correlation between π_1 and π_2 can be incorporated into the model as shown in Section 2.3.

The posterior predictive distribution naturally lends itself to be interpreted for the question of improvement in percent correct scores for an individual. The intervals produced for $n = 10, 25, 50, 100$ differ slightly from previously published intervals^{13,14} with our intervals typically being wider since they account for the uncertainty in both Y_1 and Y_2 . All of the intervals for $n = 10$ are approximately one unit longer and most of the intervals for $n = 25$ are one unit longer. The intervals as n becomes larger become more similar. This is expected as the other approaches assume normality so the intervals should be similar for $n = 100$ with discrepancies most common at $Y_1 = 0$ and $Y_1 = n$. The equal π and the correlated π

approaches had the smallest error rates with the equal π being slightly smaller. Although the intervals are similar they are not exactly the same. The equal π intervals are symmetric from $Y_1 = 0$ to n . The correlated π approach assumed a correlation of 0.46 and yielded a very similar error rate as the equal π approach. If stronger prior information is required for a different correlation or beta distributions that are centered somewhere other than 0.5 then the correlated π approach is easily modified.

We note that it may not be possible to show significant increases in some cases. For example, when $n = 10$ a score of six or greater at pre-test can not show significant improvement. Researchers using an $n = 10$ scale should use this system only when the pre-test score is five or less. This is of less concern for studies with larger n , especially the frequently used $n = 100$ scale where it takes a score of 96 or better to have no chance to show improvement.

A problem with analyses in this cochlear implant CNC score setting is the lack of a true comparison group. There is clearly within-subject variability as scores will vary from one test to the next. A test-retest study in a steady state would be informative to gain an understanding of that natural variability. In the current climate that is far too difficult as the patients are already taxed in excess as they participate in multiple studies every time they visit the clinic and these tests can take upwards of 8 h to administer.

The group comparison is an important model when assessing group summary information and adjusting for covariates. The model also lends itself to other comparisons that can be made. For instance, it may be clinically important to compare subsets of important demographic variables such as gender, race or age.

If more information is available to use an informative prior on the beta distribution then the endpoints of the credible intervals are likely to be affected. Jeffrey's prior was used in this article because of the appealing non-informative nature of it and it gave narrower intervals than other non-informative priors.

Other approaches are possible to construct similar intervals. We chose to find the interval for the number of post-treatment trials correct, $Y_2|Y_1$. The difference between Y_1 and Y_2 could also be used although the difference would not be constant across the space of Y_2 . Intervals for π_1 and π_2 may also be of interest in certain settings and could be easily formulated with the framework devised here.

Acknowledgments

This research was supported in part by research grant 5 P50 DC00242 from the National Institutes on Deafness and Other Communication Disorders, National Institutes of Health; grant MO1-RR-59, National Center for Research Resources, General Clinical Research Centers Program, National Institutes of Health; the Lions Clubs International Foundation; and the Iowa Lions Foundation.

References

1. Guyatt G, Sackett D, Adachi J, et al. A clinician's guide for conducting randomized trials in individual patients. *Canadian Medical Association Journal*. 1988; 139:497–503. [PubMed: 3409138]
2. Larson EB, Ellsworth AJ, Oas J. Randomized clinical trials in single patients during a 2-year period. *Journal of the American Medical Association*. 1993; 270:2708–12. [PubMed: 8133588]
3. Kravitz RL, Duan N, White RH. N-of-1 trials of expensive biological therapies. *Archives of Internal Medicine*. 2008; 168:1030–3. [PubMed: 18504330]
4. Ezzet F, Whitehead J. A random effects model for binary data from crossover clinical trials. *Applied Statistics*. 1992; 41:117–26.

5. Schluter PJ, Ware RS. Single patient (n -of-1) trials with binary treatment preference. *Statistics in Medicine*. 2005; 24(17):2625–36. [PubMed: 16118811]
6. Kirk RE. Practical significance: a concept whose time has come. *Educational and Psychological Measurement*. 1996; 56:746–59.
7. Tango T. Equivalence test and confidence interval for the difference in proportions for the paired-sample design. *Statistics in Medicine*. 1998; 17:891–908. [PubMed: 9595618]
8. Lloyd CJ. Confidence intervals for the difference between two correlated proportions. *Journal of the American Statistical Association*. 1990; 85:1154–58.
9. Newcombe RG. Improved confidence intervals for the difference between binomial proportions based on paired data. *Statistics in Medicine*. 1998; 17:2635–50. [PubMed: 9839354]
10. Hashemi L, Nandram B, Goldberg R. Bayesian analysis for a single 2×2 Table. *Statistics in Medicine*. 1997; 16:1311–28. [PubMed: 9232755]
11. Shi L, Bai P. Bayesian confidence interval for the difference of two proportions in the matched-paired design. *Communications in Statistics - Theory and Methods*. 2008; 37(13):2034–51.
12. Agresti A, Min Y. Frequentist performance of bayesian confidence intervals for comparing proportions in 2×2 contingency tables. *Biometrics*. 2005; 61(2):515–23. [PubMed: 16011699]
13. Thornton AR, Raffin MJM. Speech-discrimination scores modeled as a binomial variable. *Journal of Speech and Hearing Research*. 1978; 21:507–18. [PubMed: 713519]
14. Carney E, Schlauch RS. Critical difference table for word recognition testing derived using computer simulation. *Journal of Speech, Language, and Hearing Research*. 2007; 50:1203–9.
15. Freeman MF, Tukey JW. Transformations related to the angular and the square root. *Annals of Mathematical Statistics*. 1950; 21:607–11.
16. Brown LD, Cai TT, DasGupta A. Confidence intervals for a binomial proportion and asymptotic expansions. *The Annals of Statistics*. 2002; 30(1):160–201.
17. Robert, CP. Appendix A. In *the Bayesian choice: a decision-theoretic motivation*. Springer-Verlag; New York: 1994. p. 382
18. Olkin I, Liu R. A bivariate beta distribution. *Statistics & Probability Letters*. 2003; 62:407–12.
19. Gilks, WR.; Richardson, S.; Spiegelhalter, DJ. *Markov chain Monte carlo in practice*. Chapman and Hall; London: 1996.
20. Wilson, B. Cochlear implant technology. In: Niparko, JK.; Kirk, KI.; Mellon, NK.; Robbins, AM.; Tucci, DL.; Wilson, BS., editors. *Cochlear implants: principles and practices*. Lippincott, Williams, and Wilkins; New York: 2000. p. 109-18.
21. Gantz BJ, Turner CW. Combining acoustic and electrical hearing. *Laryngoscope*. 2003; 113:1726–30. [PubMed: 14520097]
22. Gantz BJ, Turner CW. Combining acoustic and electric speech processing: Iowa/Nucleus hybrid implant. *Acta Otolaryngologica*. 2004; 24:344–47.
23. Gantz BJ, Turner CW, Gfeller KE, Lowder MW. Preservation of Hearing in cochlear implant surgery: advantages of combined electrical and acoustical speech processing. *Laryngoscope*. 2005; 115:796–802. [PubMed: 15867642]
24. Gantz BJ, Turner CW, Gfeller KE. Acoustic plus electric speech processing: results of a multicenter clinical trial of the Iowa/Nucleus hybrid implant. *Audiology and Neurotology*. 2006; 11:63–8. [PubMed: 17063013]
25. Gantz BJ, Hansen MR, Turner CW, Oleson JJ, Reiss LA, Parkinson AJ. Hybrid 10 clinical trial: Preliminary results. *Audiology and Neurotology*. 2009; 14:32–8. [PubMed: 19390173]

Figure 1.

Error rates for $n = 10$ at $\pi = 0.2, 0.4, 0.6, 0.8$. The y axis represents the Y_2 error rate and the x axis contains the values of Y_1 .

Table 1

95% Credible intervals on Y_2 for a given value of Y_1 . The intervals in the $\pi_1 = \pi_2$, $\pi_1 \neq \pi_2$ and correlated π 's columns correspond to the methods described in Sections 2.1, 2.2 and 2.3, respectively

$n = 10$				$n = 50$			
Y_1	$\pi_1 = \pi_2$	$\pi_1 \neq \pi_2$	Correlated π 's	Y_1	$\pi_1 = \pi_2$	$\pi_1 \neq \pi_2$	Correlated π 's
0	(0, 3)	(0, 3)	(0, 3)	0	(0, 3)	(0, 4)	(0, 4)
1	(0, 5)	(0, 5)	(0, 5)	1	(0, 6)	(0, 6)	(0, 6)
2	(0, 6)	(0, 7)	(0, 7)	2	(0, 8)	(0, 8)	(0, 8)
3	(0, 7)	(1, 8)	(0, 8)	3	(0, 9)	(0, 10)	(0, 10)
4	(0, 8)	(1, 9)	(1, 9)	4	(0, 11)	(1, 11)	(1, 11)
5	(1, 9)	(2, 9)	(2, 10)	5	(1, 12)	(1, 13)	(1, 13)
6	(2, 10)	(2, 10)	(2, 10)	6	(1, 14)	(2, 14)	(2, 14)
7	(3, 10)	(3, 10)	(3, 10)	7	(2, 15)	(2, 16)	(2, 15)
8	(4, 10)	(4, 10)	(4, 10)	8	(2, 16)	(3, 17)	(3, 17)
9	(5, 10)	(6, 10)	(5, 10)	9	(3, 18)	(3, 18)	(3, 18)
10	(7, 10)	(8, 10)	(7, 10)	10	(3, 19)	(4, 19)	(4, 19)
				11	(4, 20)	(5, 20)	(5, 20)
				12	(5, 21)	(5, 22)	(5, 22)
				13	(5, 22)	(6, 23)	(6, 23)
				14	(6, 23)	(7, 24)	(7, 24)
				15	(7, 25)	(7, 25)	(8, 25)
				16	(8, 26)	(8, 26)	(8, 26)
				17	(9, 27)	(9, 27)	(9, 27)
				18	(9, 28)	(10, 28)	(10, 28)
				19	(10, 29)	(11, 29)	(10, 29)
				20	(11, 30)	(11, 30)	(11, 30)
				21	(12, 31)	(12, 31)	(12, 31)
				22	(13, 32)	(13, 32)	(13, 32)
				23	(14, 33)	(14, 33)	(14, 33)
				24	(14, 34)	(15, 34)	(15, 34)
$n = 25$							
Y_1	$\pi_1 = \pi_2$	$\pi_1 \neq \pi_2$	Correlated π 's				

Y_1	$n = 10$			$n = 50$		
	$\pi_1 = \pi_2$	$\pi_1 \neq \pi_2$	Correlated π 's	Y_1	$\pi_1 = \pi_2$	Correlated π 's
0	(0, 3)	(0, 4)	(0, 4)	25	(15, 35)	(16, 35)
1	(0, 6)	(0, 6)	(0, 6)	26	(16, 36)	(17, 36)
2	(0, 7)	(0, 8)	(0, 8)	27	(17, 36)	(18, 37)
3	(0, 9)	(0, 9)	(0, 9)	28	(18, 37)	(19, 38)
4	(0, 10)	(1, 11)	(1, 11)	29	(19, 38)	(20, 39)
5	(1, 11)	(1, 12)	(1, 12)	30	(20, 39)	(20, 40)
6	(1, 13)	(2, 13)	(2, 13)	31	(21, 40)	(22, 41)
7	(2, 14)	(2, 14)	(2, 14)	32	(22, 41)	(23, 41)
8	(2, 15)	(3, 15)	(3, 15)	33	(23, 41)	(24, 42)
9	(3, 16)	(4, 16)	(4, 16)	34	(24, 42)	(25, 43)
10	(4, 17)	(4, 17)	(4, 17)	35	(25, 43)	(26, 44)
11	(5, 18)	(5, 18)	(5, 19)	36	(27, 44)	(27, 45)
12	(5, 19)	(6, 19)	(6, 19)	37	(28, 45)	(28, 45)
13	(6, 20)	(7, 20)	(7, 20)	38	(29, 45)	(29, 46)
14	(7, 20)	(8, 21)	(8, 21)	39	(30, 46)	(30, 47)
15	(8, 21)	(9, 22)	(9, 22)	40	(31, 47)	(31, 47)
16	(9, 22)	(10, 22)	(9, 22)	41	(32, 47)	(32, 48)
17	(10, 23)	(11, 23)	(10, 23)	42	(34, 48)	(34, 49)
18	(11, 23)	(12, 24)	(12, 24)	43	(35, 48)	(35, 49)
19	(12, 24)	(13, 24)	(13, 25)	44	(36, 49)	(37, 50)
20	(14, 24)	(14, 25)	(14, 25)	45	(38, 49)	(38, 50)
21	(15, 25)	(15, 25)	(15, 25)	46	(39, 50)	(39, 50)
22	(16, 25)	(17, 25)	(16, 25)	47	(41, 50)	(41, 50)
23	(18, 25)	(18, 25)	(18, 25)	48	(42, 50)	(43, 50)
24	(19, 25)	(20, 25)	(20, 25)	49	(44, 50)	(44, 50)
25	(22, 25)	(22, 25)	(20, 25)	50	(47, 50)	(46, 50)

Table 2

95% Credible intervals on Y_2 for a given value of Y_1 . The intervals in the $\pi_1 = \pi_2$, $\pi_1 \neq \pi_2$ and correlated π 's columns correspond to the methods described in Sections 2.1, 2.2 and 2.3, respectively

Y_1	$n = 100; Y_1 \leq 50$			$n = 100; Y_1 > 50$		
	$\pi_1 = \pi_2$	$\pi_1 \neq \pi_2$	Correlated π 's	$\pi_1 = \pi_2$	$\pi_1 \neq \pi_2$	Correlated π 's
0	(0, 3)	(0, 4)	(0, 4)	51 (37, 65)	(38, 65)	(38, 65)
1	(0, 6)	(0, 6)	(0, 6)	52 (38, 66)	(39, 66)	(38, 66)
2	(0, 8)	(0, 8)	(0, 8)	53 (39, 67)	(40, 67)	(40, 67)
3	(0, 10)	(0, 10)	(0, 10)	54 (40, 68)	(41, 68)	(40, 68)
4	(0, 11)	(1, 12)	(1, 12)	55 (41, 68)	(42, 69)	(41, 69)
5	(1, 13)	(1, 13)	(1, 13)	56 (42, 69)	(43, 70)	(42, 70)
6	(1, 14)	(2, 15)	(2, 15)	57 (43, 70)	(44, 71)	(44, 71)
7	(2, 16)	(2, 16)	(2, 16)	58 (44, 71)	(45, 72)	(45, 71)
8	(2, 17)	(3, 18)	(3, 17)	59 (45, 72)	(46, 73)	(45, 73)
9	(3, 18)	(3, 19)	(3, 19)	60 (46, 73)	(47, 74)	(46, 73)
10	(3, 20)	(4, 20)	(4, 20)	61 (47, 74)	(48, 74)	(48, 74)
11	(4, 21)	(4, 22)	(4, 22)	62 (48, 75)	(49, 75)	(48, 75)
12	(4, 22)	(5, 23)	(5, 23)	63 (49, 76)	(50, 76)	(49, 76)
13	(5, 24)	(6, 24)	(5, 24)	64 (50, 77)	(51, 77)	(51, 77)
14	(6, 25)	(6, 25)	(6, 25)	65 (51, 78)	(52, 78)	(51, 78)
15	(6, 26)	(7, 27)	(7, 27)	66 (52, 78)	(53, 79)	(53, 79)
16	(7, 27)	(8, 28)	(8, 28)	67 (53, 79)	(54, 80)	(54, 80)
17	(8, 29)	(8, 29)	(8, 29)	68 (55, 80)	(55, 81)	(55, 81)
18	(9, 30)	(9, 30)	(9, 30)	69 (56, 81)	(56, 81)	(56, 82)
19	(9, 31)	(10, 31)	(10, 32)	70 (57, 82)	(57, 82)	(57, 82)
20	(10, 32)	(11, 33)	(10, 33)	71 (58, 83)	(58, 83)	(58, 84)
21	(11, 33)	(11, 34)	(11, 34)	72 (59, 83)	(59, 84)	(59, 84)
22	(12, 34)	(12, 35)	(12, 35)	73 (60, 84)	(60, 85)	(60, 85)
23	(12, 36)	(13, 36)	(13, 36)	74 (61, 85)	(61, 86)	(61, 86)
24	(13, 37)	(14, 37)	(14, 37)	75 (62, 86)	(63, 86)	(62, 86)
25	(14, 38)	(15, 38)	(15, 38)	76 (63, 87)	(64, 87)	(64, 87)

Y_1	$n = 100; Y_1 \leq 50$			$n = 100; Y_1 > 50$		
	$\pi_1 = \pi_2$	$\pi_1 \neq \pi_2$	Correlated π 's	$\pi_1 = \pi_2$	$\pi_1 \neq \pi_2$	Correlated π 's
26	(15, 39)	(15, 39)	(16, 39)	77	(64, 88)	(65, 88)
27	(16, 40)	(16, 40)	(16, 40)	78	(66, 88)	(66, 89)
28	(17, 41)	(17, 42)	(17, 42)	79	(67, 89)	(67, 90)
29	(17, 42)	(18, 43)	(18, 43)	80	(68, 90)	(68, 90)
30	(18, 43)	(19, 44)	(18, 44)	81	(69, 91)	(69, 91)
31	(19, 44)	(19, 45)	(19, 45)	82	(70, 91)	(71, 92)
32	(20, 45)	(20, 46)	(21, 46)	83	(71, 92)	(72, 93)
33	(21, 47)	(21, 47)	(21, 47)	84	(73, 93)	(73, 94)
34	(22, 48)	(22, 48)	(22, 48)	85	(74, 94)	(74, 94)
35	(22, 49)	(23, 49)	(23, 49)	86	(75, 94)	(76, 95)
36	(23, 50)	(24, 50)	(23, 50)	87	(76, 95)	(76, 96)
37	(24, 51)	(25, 51)	(25, 51)	88	(78, 96)	(78, 96)
38	(25, 52)	(26, 52)	(25, 52)	89	(79, 96)	(79, 97)
39	(26, 53)	(26, 53)	(26, 53)	90	(80, 97)	(80, 97)
40	(27, 54)	(27, 54)	(28, 54)	91	(82, 97)	(82, 98)
41	(28, 55)	(28, 55)	(28, 55)	92	(83, 98)	(83, 99)
42	(29, 56)	(29, 56)	(29, 56)	93	(84, 98)	(84, 99)
43	(30, 57)	(30, 57)	(30, 57)	94	(86, 99)	(86, 100)
44	(31, 58)	(31, 58)	(31, 58)	95	(87, 99)	(87, 100)
45	(32, 59)	(32, 59)	(32, 59)	96	(89, 100)	(89, 100)
46	(32, 60)	(33, 60)	(33, 60)	97	(90, 100)	(90, 100)
47	(33, 61)	(34, 61)	(34, 61)	98	(92, 100)	(92, 100)
48	(34, 62)	(35, 62)	(35, 63)	99	(94, 100)	(94, 100)
49	(35, 63)	(36, 63)	(36, 63)	100	(97, 100)	(96, 100)
50	(36, 64)	(37, 64)	(37, 64)			

Table 3

Error rates from simulations of Y_1 and Y_2 values generated from $\text{Binomial}(n, \pi)$. Reported are the percentage of times that Y_2 fell outside the interval even though $\pi_1 = \pi_2$

n	$\pi_1 = \pi_2$	$\pi_1 \neq \pi_2$	Correlated π 's	Thornton–Raffin
10	0.026	0.042	0.026	0.055
25	0.034	0.041	0.038	0.067
50	0.040	0.044	0.041	0.057
100	0.043	0.047	0.044	0.055

CNC pre-test and post-test scores from Hybrid clinical trial for subjects with change of 25 or less in word correct score. 'Yes' denotes a post-test score is within the interval using the denoted method from Table 2. 'No' denotes that the post-test is not within the credible interval in Table 2. The Group column gives the estimated $\Pr(\pi_{2i} > \pi_{1i})$ for individual i

Table 4

ID	Pre-test	Post-test	Change	$\pi_1 = \pi_2$	$\pi_1 \neq \pi_2$	Correlated π 's	Group
6	59	64	5	Yes	Yes	Yes	0.862
8	56	62	6	Yes	Yes	Yes	0.888
10	51	64	13	Yes	Yes	Yes	0.985
12	33	4	-29	No	No	No	0.000
13	41	33	-8	Yes	Yes	Yes	0.235
14	38	41	3	Yes	Yes	Yes	0.789
16	28	37	9	Yes	Yes	Yes	0.954
17	20	39	19	No	No	No	0.999
18	25	45	20	No	No	No	0.999
22	21	13	-8	Yes	Yes	Yes	0.192
23	21	42	21	No	No	No	0.999
24	30	51	21	No	No	No	0.999
25	69	86	17	No	No	No	0.999
27	62	72	10	Yes	Yes	Yes	0.963
28	29	29	0	Yes	Yes	Yes	0.658
29	36	46	10	Yes	Yes	Yes	0.961
30	79	92	13	No	No	No	0.999
34	77	86	9	Yes	Yes	Yes	0.975
36	22	42	20	No	No	No	0.999
38	43	68	25	No	No	No	0.999
41	23	45	22	No	No	No	0.999
44	29	23	-6	Yes	Yes	Yes	0.319
45	19	39	20	No	No	No	0.999
51	35	51	16	No	No	No	0.995
52	26	40	14	No	No	No	0.992
55	75	90	15	No	No	No	0.999
56	66	80	14	No	No	No	0.995

Oleson

ID	Pre-test	Post-test	Change	$\pi_1 = \pi_2$	$\pi_1 \neq \pi_2$	Correlated π 's	Group
57	22	27	5	Yes	Yes	Yes	0.888
60	27	34	7	Yes	Yes	Yes	0.925
62	57	65	8	Yes	Yes	Yes	0.935
64	54	62	8	Yes	Yes	Yes	0.928
67	48	61	13	Yes	Yes	Yes	0.986
68	34	29	-5	Yes	Yes	Yes	0.378