An Intrasubject Replication Strategy for Evaluating Occlusal Splint Therapy

Abstract

Okeson and others have proposed that statistical comparisons between groups of treated and untreated patients be used to evaluate the effectiveness of occlusal splint therapy. This article argues that valid between-groups comparisons cannot be made by most clinicians and that such comparisons do not satisfactorily guide the treatment of individual patients. An intrasubject replication design strategy is then offered as a model for clinicians to use in evaluating splint-therapy effects in individual cases.

F. Dudley McGlynn, Ph.D.

Dr. McGlynn received his Ph.D. in clinical psychology from the University of Missouri in 1968. He has taught at the University of Missouri and at Mississippi State University. For the past five years he has served as professor in the departments of Basic Dental Sciences and Clinical Psychology at the University of Florida in Gainesville. Dr. McGlynn is doing research on facial pain, on compliance with prescribed dental regimens, and on fear/avoidance of dentists.

Dr. McGlynn is a member of the Southeastern Psychological Association and the Association for Advancement of Behavior Therapy. He is the author of numerous research articles, chapters, and reviews, and he serves on the editorial boards of four national journals.

Jeffrey E. Cassisi, M.A.

Mr. Cassisi received his bachelor's degree in psychology from the University of Florida in 1979 and his master's degree in experimental psychology from the University of Hawaii in 1981. He is a co-author of a chapter on the diagnosis and treatment of bruxism. He has carried out research on biofeedback, the measurement of dental fear, the measurement of bruxism, and the treatment of shyness.

Mr. Cassisi is presently a Ph.D. candidate in the Department of Clinical Psychology at the University of Florida, where he is specializing in behavioral dentistry and medicine.
An Intrasubject Replication Strategy for Evaluating Occlusal Splint Therapy

By F. Dudley McGlynn, Ph.D., Jeffrey E. Cassisi, M.A.

Okeson, Moody, Kemper, and Calhoun' reviewed the treatment-outcome literature on occlusal splint therapy. They characterized the literature as inconclusive and proposed a set of methodological correctives for further work in the area. In this article, we will briefly reiterate the major features of the proposal by Okeson et al. We then argue that their design is impractical from the vantage point of most clinicians and that, in fact, its experimental product only begins to answer the questions in which clinicians are interested. Finally, we offer a straightforward methodological approach that clinicians can use in evaluating results from occlusal-splint treatments. This approach will guide clinical decision-making as well as provide scientific data.

The Between-Groups Comparison Strategy

The design by Okeson et al. was developed to test the effectiveness of splint therapy by comparing it with a no-treatment control condition. They suggest that repeated pain measures be acquired from a group of subjects receiving occlusal splint therapy as well as from a group of subjects (temporarily) left untreated. Statistical comparisons of pain measures should then be made to see whether observed differences between the distributions representing treated and untreated groups are coincidental.

The bulk of the Okeson et al. narrative describes how to make comparisons between groups of treated and untreated subjects properly. Subjects are to be matched with respect to symptomatology, randomly assigned to treated versus untreated conditions, recruited in sufficient numbers to meet statistical assumptions, assessed by condition-blind examiners, and so forth. These various desiderata for making the between-groups comparison are not at issue here. Rather, our concern is with variables that govern the choice of between-groups statistical methodology per se. The Okeson et al. narrative also describes a somewhat incomplete approach to measuring facial pain (i.e., assigning numbers to it). The issue of pain assessment is also beyond the scope of this paper. Our concern is with strategies for showing the relationships between splint therapy and pain measures, whatever those measures might be.

Between-Groups Comparisons Are Impractical

Okeson et al. gloss over the problem of how many subjects are required by experimental comparisons between groups. They say only that "The sample size would be adequate to assure statistical significance in each treatment area." (There is a way to make such a determination statistically, provided that some estimator data are available and that acceptable error rates are endorsed.) Being very liberal, one might suppose that 15 subjects in each experimental condition will suffice to meet statistical assumptions. If so, 30 patients with craniomandibular disorders would be required for a meaningful experiment. (A more conservative account might call for a great many more.) Notwithstanding the recent surge of such patients into dentists' offices, the average dentist would require a very long time to attract and treat 30 of them. Very probably, the time required would be too long to preclude strong influence from extraneous variance sources that inevitably occur with time.

Okeson et al. do mention several practical problems with their proposal and some potential ethical concerns as well. They then argue that their design "will nevertheless allow researchers to adequately measure the effectiveness of various types of splints." This statement is true in an actuarial sense. Nonetheless, their design is practical only for a select few professionals who specialize in treating craniomandibular disorders.

Between-Groups Comparisons Answer Actuarial, Not Clinical, Questions

Using composite group data to make empirical comparisons across different experimental conditions is an ingrained habit of social scientists in America. True differences between experimental conditions are to be
found by comparing distributions of scores produced by large groups of subjects exposed unilaterally to those experimental conditions. Researchers who use this approach assume that for each of the to-be-compared conditions of the experiment a "true" population parameter does exist and can be estimated from the dispersion that occurs within distributions of experimental measures. They also assume that to-be-compared estimates of population parameters for different experimental conditions become more accurate as the samples included in those conditions become larger.

In the case of Okeson et al., splint-therapy effects and no-treatment effects are assumed to have "true" representations in the sampled population—representations that can be estimated using the dispersion in measures taken from treated and untreated experimental samples. If the samples of subjects are sufficiently large, we can arrive at probability estimates that there is a true difference between the population parameter for splint-therapy effects and the population parameter for no-treatment effects.

The bona fide experimental units of such a study are the two sample populations, not then individuals of which the two samples are comprised. The bona fide purpose of such a study is to make statements about the whole population sampled, not about the individuals who were included. This is shown conceptually in the quest for population parameters just described. It is also seen in the history of inferential statistics. The important participants in this history were gamblers, economists, and agricultural researchers, all of whom sought to make statements about event populations, not about individual instances.5

**Between-Groups Comparisons Guide Treatment Selection, Not ‘Treatment Evaluation**

Population-anchored results of the type produced by statistical designs are of value mainly when the clinician is selecting a treatment. Ideally, actuarial treatment-outcome results are weighed along with factors such as the invasiveness or reversibility of available procedures in choosing a treatment modality.

Thus a statement such as "the effects of splint therapy were significantly better than were the effects of no treatment" might be valid. This statement could guide the clinician to use splint therapy as a treatment modality if the only alternative is to leave the patient untreated. Another example might be the statement that "the effects of splint therapy were significantly better than were the effects of equilibration, and the effects of both treatments were significantly better than no treatment at all." On reading this statement, the clinician might begin treatment with splint therapy and use equilibration later on if splint therapy were found to be ineffectual.

Because between-groups statistical comparisons guide clinicians in choosing from among potential treatment approaches, we are not arguing against experimentation of the type Okeson et al. proposed. However, our thesis is that population-anchored results do not tell clinicians enough about outcomes in individual cases to serve as sufficient guides for clinical work. Even granting that the two population-anchored statements above are true, some individual patients will be as well off without treatment as with it, some patients will benefit more from equilibration than from splint therapy, and so on. Because of these actuarial uncertainties, clinicians should monitor and evaluate therapy effects carefully in each individual case. Between-groups statistical methodology is not suited to clinical treatment evaluation of this sort.

**Intrasubject Replication Methodology**

The major purpose of experimental research is to describe reliable relationships among experimental variables. The purpose of experimental-design logic is to permit descriptions of these relationships that the scientific audience (e.g., journal referees) will accept as believable or credible. Mathematical statistics, which governs the design-logic of between-groups experimentation, can be viewed as an evolving system of conventions or rules by which authors, referees, and editors can agree on the kinds of relationship descriptions that will have believability or credibility. This old and established set of conventions was made possible late in the eighteenth century when Laplace's "continuous equation of the law of error" connected probability theory with calculus.

Over the past 50 years, an entirely different system of design logic has evolved within American psychology. A discipline known as "experimental analysis of behavior" governs the design parameters of "intrasubject replication" research. The latter is a type of behavioral experimentation in which each individual subject (not a group of subjects) is the bona fide unit of study. Experimental analysis of behavior offers a recently evolved set of conventions by which authors, referees, and editors can agree on what kinds of relationship descriptions have credibility when repeated measures taken from separate individual subjects are studied. The experimental analysis of behavior also offers clinicians conventions that allow each individual patient's case to be viewed as an autonomous experiment and that provide trustworthy conclusions about treatment efficacy for that individual.
Intrasubject Replication Design Logic

According to orthodox rules, the between-groups experimenter seeks to convince the audience that his or her results are meaningful or credible by arguing mathematically that they are not likely due to coincidence. The unifying purpose of intrasubject replication design logic is also to convince the audience that results are meaningful and not accidental. In the latter case, however, the strategy is to show predicted and repeated changes in trustworthy measures of an individual's behavior following sequential changes in the experimental conditions imposed upon that individual. The fact that there is obvious experimental control over the individual's conduct argues against interpreting experimental results in terms of coincidence.

Classifications of intrasubject replication designs are usually organized into four categories. By and large, these categories reflect different ways of showing sequential experimental control over some regularly assessed feature(s) of the subject's behavior. The "reversal" or "withdrawal" design category is of interest here because, in general, it is most readily applicable to evaluating occlusal splint effects.

The Reversal or Withdrawal Design

In using the reversal design, the experimenter's intent is to argue against coincidence by showing how the subject's behavior predictably changes with the introduction, withdrawal, and reintroduction of the independent variable condition. This independent variable condition is usually a treatment of some sort. All routine intrasubject replication designs use a baseline measurement period preceding the first introduction of treatment and repeated measurement throughout all the phases of treatment. Behaviors are recorded regularly under baseline, treatment, baseline, and treatment conditions successively.

In applied uses of this ABAB design, symptoms are regularly recorded during each of the four phases. Obvious and repeated differences between baseline and treatment phases attest to the efficacy of treatment. However, the replication of differences between baseline and treatment measures is not enough to render the ABAB design persuasive. Rather, it is the conditions under which decisions are made to introduce, withdraw, and reinstate treatment that are important. Some examples of these conditions using splint therapy and daily pain ratings are given below.

Applications to Evaluating Splint Therapy Effects

Figure 1 depicts the hypothetical course of the reversal design approach in three different splint therapy patients. The purpose of daily baseline pain recording is to find the most stable trend in the patient's pain ratings and to predict the immediate future course of the patient's pain with no treatment. (This prediction is shown by the dotted lines in the drawing.)

For the first patient, daily pain ratings are variable but stable over the course of a two-week baseline recording period. Clinically significant pain reduction is demonstrable because of a predicted short-term continuation in baseline stability. For the second patient, daily pain ratings are variable, with increasing intensity over the course of the two-week baseline recording period. Clinically significant pain reduction is even more demonstrable here. For both of these patients, the clinician can decide to introduce treatment after the two-week baseline period.

For the third patient, daily pain ratings are variable, with a diminishing trend over the two weeks of baseline recording. Clinically significant pain reduction from therapy is not readily demonstrable, because the short-term prediction is that improvement may continue without treatment. The clinician may thus decide to withhold treatment until a stable trend other than improvement appears in the daily records. For this patient, the baseline pain ratings stabilize during the third week, after which treatment is introduced (Figure 1).

As with baseline recordings, the purpose of daily pain recording during treatment is to describe a stable trend and to predict the immediate future course of pain under continued treatment conditions. For our first hypothetical patient, pain recordings during treatment stabilized between the third and fourth weeks. The prorated continuation of treatment effects is substantially below the prorated continuation of baseline values. The clinician can decide to withdraw treatment after the fourth week in hopes that an obvious reversal to baseline pain values will argue against coincidental interpretations of the effect of treatment.

For the second hypothetical patient, pain recordings during treatment stabilized during the fifth week. Again, the prorated stable trend for pain reports during future treatment is substantially lower than is the last stable trend in the baseline. Here, the clinician can decide after the fifth week to test for coincidence by withdrawing therapy for a while. For the third patient, pain recordings during treatment shifted suddenly and stabilized quickly. The clinician can decide to test for coincidence by returning to baseline conditions after only a few days of treatment.

Once baseline recording conditions have been put into effect for the second time, the experimental analysis begins anew. Daily pain ratings are monitored until a stable
trend other than improvement appears, then treatment is
instituted and monitored for the second time. (Figure 1
depicts several scenarios that are possible with the ABAB
format.)

We can easily envision clinical situations in which
facial pain patients record the intensity of their pain once
or twice daily, and clinicians evaluate these records in
making chairside decisions about when to initiate and
withdraw occlusal splint therapy. The type, frequency, and
complexity of facial pain rating formats would vary ac-
cording to the clinician's exact clinical or clinical-research
purpose, the patient's likely adherence to home monitor-
ing regimens, and so on. As we noted earlier, assessment
of pain per se is beyond the scope of this article.

Variations of the Reversal Design

The ABAB design exemplified above represents a
larger class of design logic in which arguments against
coincidental therapy effects are based on repeated and
predicted instances in which the patient's symptoms track
the introduction and withdrawal of therapy. Nothing inher-
et to this inferential approach prevents the clinician from
beginning with a treatment phase (BABA design). This
might, for example, be done for a patient whose intense
pain prompts the clinician to forgo a pretreatment baseline
assessment. There is also no reason the clinician could not
add a third baseline/therapy sequence (ABABAB design).
This might be done, for example, when the graphs from the
ABAB series are not wholly convincing and the clinician
wants further evidence of specific therapy effects. The
practitioner could also compare the effects of different
reversible treatments. He or she might, for example, use an
AB,B₂AB₂ series. In such a case, the clinician might first
take baseline recordings (A), then institute splint therapy
(BO and find that it fails, next institute Valium* therapy

---

Figure 1
Example results of a reversal design approach to evaluating occlusal splint effects in individual patients.
Special Considerations in Reversal Designs

The reversal design is not without shortcomings. The persuasiveness of the design depends upon observing declines in symptoms with treatment and increases in symptoms when treatment is withdrawn. Because of the latter requirement, the design is not applicable to treatments that have potentially permanent effects. For example, a reversal design could not be used to evaluate the effects of grinding down occlusal interferences, because the interferences cannot be replaced for a return to baseline conditions. However, this method can be used to evaluate the effects of occlusal splint therapy because the splint can be completely withdrawn so that the symptoms can return to baseline levels.

A related problem is that symptom measures do not always revert to baseline levels even when theoretically reversible therapy modes are used. When this occurs, the initial correspondence between treatment implementation and symptom change could be interpreted as coincidence.

A third potential problem in using ABAB designs (and all other intrasubject replication designs) is extreme variability in data within phases. As a general rule, extreme variability obscures the trends that are needed to guide shifts between phases and to evaluate the effects of interventions. For some facial pain patients, day-to-day variability will be so great as to preclude an intrasubject replication approach altogether. For others, variability will be sufficiently great to extend the durations of phases beyond what the practitioner endorses as clinically feasible. For a great many patients, however, these problems will not arise. Moreover, there are various approaches that clinicians can use in overcoming day-to-day variability in pain ratings (e.g., averaging multiple ratings to produce a single data point per day, averaging data points across two or three days, and so forth).

Another problem common within contemporary ABAB design logic is that visual inspection of graphs is the sole means of classifying symptom changes as clinically significant or not. There are currently no criteria or "inspection rules" that would render visual inspection uniform in different settings, at different times, and so forth. Users of ABAB design logic admit that the treatment under study must produce rather dramatic changes if visual inspection for therapy effects is to work. By applying statistics to intrasubject replication designs, resistance to this arises largely from the view that reliance on raw visual inspection of graphs serves to focus attention on symptom changes that are clinically significant, while statistics might draw attention to changes that are statistically "significant" yet clinically trivial.

A final criticism of reversal designs is that ethical questions arise from a decision to withdraw a successful treatment, even temporarily. No doubt clinical situations do exist (such as stable, intense pain during baseline recording) that would forestall use of a reversal strategy for ethical reasons. The ethical argument against reversal methodology would also have much merit if reversals were done solely for research purposes. In the clinical context, however, the patient's best interest is served when the clinician is free to discover accidental relationships between therapies and changes in symptoms.

Overview

Statistical comparisons of pain ratings among treated and untreated patient populations will serve to describe occlusal splint effects actuarially and to provide clinicians with general guidelines about the feasibility of the approach. However, the information distilled from between-groups comparisons leaves significant uncertainty about splint therapy effects in individual cases.

The proposals in this article are offered as algorithms for the clinical management of occlusal splint therapy which allow the clinician to fill some of the gaps in actuarial data. Dentists treating facial pain patients can conveniently monitor and evaluate splint therapy effects in individual patients using the intrasubject reversal design logic we have just described. We feel that dentists would be well-served by so doing.

Of course, problems will inevitably arise. Patient A might forget to record her pain each day. Patient B might allow as how all this recording stuff is too much trouble. No data exist on the frequencies of these types of problems. However, compliance with simple recording regimens should be fairly high among chronic facial pain patients, and there are ways to deal easily with some compliance problems. (For example, a receptionist could telephone Patient A to record her pain ratings.)

The clinical management proposals made here can also be viewed as bona fide experimental designs for studying individual subjects. The dentist who clinically assesses splint therapy effects through daily pain measurement in an ABAB design is doing therapy-outcome research.
fact, a major benefit of the intrasubject replication strategy is that clinicians without access to large numbers of facial pain patients can participate in research by providing detailed reports of ABAB protocols. As a research stratagem, the intrasubject replication format is complementary to the type of research proposed by Okeson et al. The two approaches answer different questions, and both are of interest to the clinician.

References