Outcome Oriented

The Online Newsletter of the
Center for Outcome Measurement in Brain Injury **(COMBI)**

Summer 2002

Measuring Up!

The COMBI continues to add more important scales to its resource center. As of July 2002 there are currently twenty-one measures featured and detailed in the COMBI.

Agitated Behavior Scale (ABS)

Awareness Questionnaire (AQ)

Coma/Near Coma Scale (CNC)

Community Integration Questionnaire (CIQ)

The Craig Handicap Assessment and Reporting Technique (CHART)

The CHART Short Form (CHART-SF)

The Craig Hospital Inventory of Environmental Factors (CHIEF)

Disability Rating Scale (DRS)

The Family Needs Questionnaire (FNQ)

Functional Assessment Measure (FAM)

Functional Independence Measure (FIM)

Glasgow Outcome Scale (GOS)

Extended Glasgow Outcome Scale (GOS-E)

Levels of Cognitive Functioning Scale (LCFS)

Mayo Portland Adaptability Inventory (MPAI)

Neurobehavioral Functioning Inventory (NFI)

The Orientation Log (O-Log)

The Patient Competency Rating Scale (PCRS)

Satisfaction With Life Scale (SWLS)

Service Obstacle Scale (SOS)

Supervision Rating Scale (SRS)

Mysteries Revealed Inside:

Null Hypothesis Significance Testing: The Problem with P Values

Scott R. Millis, PhD
Traumatic Brain Injury National Data Center
Kessler Medical Rehabilitation Research & Education Corporation

Scott R. Millis, PhD is co-director of the Traumatic Brain Injury National Data Center and a Senior Research Scientist at Kessler Medical Rehabilitation Research and Education Corporation. Dr. Millis will be writing for Outcome Oriented on issues relating to statistics and interpretation.

Sure, it's ≤ .05, but can it be trusted?

One doesn't need to be a statistician to have heard of "P values." It is common to encounter P values in research papers that evaluate the effectiveness of medical and rehabilitation treatments. Many of us have been taught to look for the magical number of P less than .05. We may not exactly understand how or why this number of .05 has taken on such significance, but it is probably one of the few bits of information that most of us still remember from Statistics 101. We may believe that if P is less than .05, then the treatment *must* be effective; if P is .06 (or larger), then it is not. Use of the P value in this way is one aspect of null hypothesis significance testing (NHST). Carver (1978) describes NHST:

> *Statistical significance testing sets up a straw man, the null hypothesis, and tries to knock him down. We hypothesize that two means represent the sample population and that sample or chance alone can explain any difference we find between the two means. On the basis of this assumption, we are able to figure out mathematically just how often differences as large or larger than the difference we found would occur as a result of chance or sampling. (p. 381)*

If the difference between the groups on the outcome measure is likely to have occurred less than 5 times out of 100 (i.e., P < .05), we consider this outcome to be a "rare" event. Then, the conventional wisdom is to conclude that there is a "significant" difference between the groups. Conversely, if the probability value (P value) associated with the group difference is greater than .05, we may be tempted to conclude that there is "nothing going on" and that there is no difference between the groups. For example, in a randomized trial of cognitive rehabilitation for traumatic brain injury, Salazar et al. (2000) stated that "the overall benefit of an in-hospital cognitive rehabilitation for patients with moderate-to-severe TBI was similar to that of home rehabilitation" (p. 3075). With regard to return-to-work, 90% of their patients in the hospital treatment group resumed work compared to 94% of the home rehabilitation group. This 4% difference had an associated P value of .51, apparently prompting Salazar and associates to conclude that there were no differences in treatment.

Continued on Page 2

Continued from Page 1

Should one accept these conclusions? Do the P values tell the whole story? Unfortunately, P values alone are not enough when performing statistical analysis. As I will discuss in this article, P values have some significant limitations. They must be supplemented with additional statistical information lest the unwary be led astray. Paul Meehl described NHST as "a potent but sterile intellectual rake who leaves in his merry path a long train of ravished maidens but no viable scientific offspring" (Cohen, 1994, p. 997).

The Problem with P Values

- When used in the typical way, P values and NHST essentially answer an uninteresting question. That is, when P is less than .05, it suggests that "there is not nothing" (Dawes, 1991, p. 252). But is this meaningful? To know that there is "not nothing" does not seem to tell us very much. In NHST, we are testing the probability that the data (e.g., the difference between two groups or the magnitude of a correlation between two variables) could have occurred if there were no differences (or zero correlation). More specifically, we are evaluating a conditional probability statement: P(D | Ho). A common misconception is that we are evaluating P(Ho | D). However, the two conditional probability statements are not equivalent. Hence, the P value in NHST does not represent the probability that the null hypothesis is true (Cohen, 1994).
- P values alone do not provide information whether statistical results, may it be a difference between two treatments or a correlation coefficient, are large enough to have practical significance or have clinical importance (Borenstein, 1994). For example, a small difference between placebo and medication can be statistically significant if the sample size is large enough – which may be quite common in large multi-center clinical drug trials. Conversely, a small sample and a nonsignificant P value can mask a clinically meaningful effect. A nonsignificant P value does not mean that there is no difference but, rather, no evidence was found that there was a difference. For example, a Pearson correlation of .63 is statistically significant (p < .05) in a sample size of 10 but when the sample size is 500, a correlation of .09 is significant! Mathews and Altman (1996) point out, "A P value is a composite which depends not only on the size of an effect but also on how precisely the effect has been estimated (its standard error). So differences in P values can arise because of differences in effect sizes or differences in standard errors or a combination of the two" (p. 808). Kirk (1996) notes,

  > *Because the null hypothesis is always false, a decision to reject it simply indicates that the research design had adequate power to detect a true state of affairs, which may or may not be a large effect or even a useful effect. It is ironic that a ritualistic adherence to null hypothesis significance testing has led researchers to focus on controlling the Type I error that cannot occur because all null hypotheses are false (p. 747).*

- A small P value (e.g., p < .05) does not imply that the result from a single study will replicate in subsequent studies (Carver, 1978).

What Should Be Done

Should researchers abandon P values and NHST altogether? No. However, P values and NHST need to be put into proper perspective and supplemented with additional methods.

- Conventional NHST can often provide a useful starting point by indicating the direction of differences between groups and by offering a method for dealing with chance variation. As Mulaik et al. (1997) discuss, "We cannot get rid of significance tests because they provide us with the criteria by which provisionally to distinguish results due to chance variation from results that represent systematic effects in data available to us" (p. 81).
- Investigators should routinely report effect sizes for each statistical comparison. Effect sizes give an estimate of the magnitude of group differences, correlations, and related comparisons. There are a variety of effect sizes available such as Glass' D, Hedges' g, Cohen's d, r, and eta (Rosenthal et al., 2000). For example, Hedges' g can be used when comparing mean differences between two groups: it is the difference between the means divided by the pooled estimate of the population standard deviation. Effect sizes can help to determine the practical significance of statistical findings.
- Along with effect sizes, investigators should report confidence intervals (CIs). The CI is a measure of the precision of the study findings (Matthews & Altman, 1996). A technical interpretation of the 95% CI is that if the same study were done 100 times with different samples of patients, 95% of these intervals would contain the true population values (e.g., group difference, mean, proportion). The wider the CI, the less precise the estimate. CIs can be particularly useful in interpreting statistically nonsignificant results. Returning to the Salazar et al. (2000) study, they reported that the 4% difference in return-to-work rates between the hospital treatment and home treatment groups was not significant (p = .51). Yet, the associated confidence interval was [-5% to 14%]. This CI is quite wide, which suggests that one cannot necessarily conclude that the treatments are equivalent. In fact, we cannot rule out that a reasonable large treatment effect for either intervention might exist. Hedges' g for the comparison is .15 with a 95% confidence interval of [-.22 to .51]. In effect, if one relies on P values alone, one runs the risk of both over-estimating and underestimating treatment effects.
- Investigators should routinely use graphical methods to examine their data.
- There is no substitute for replicating findings except by repeating the experiment.
- Although outside the scope of this paper, investigators are encouraged to look into Bayesian statistical techniques (e.g., Gill, 2002).

References

Borenstein, M. (1994). The case for confidence intervals in controlled clinical trials. *Controlled Clinical Trials,* 15, 411-428.

Carver, R. P. (1978). The case against statistical significance testing. *Harvard Educational Review,* 48, 378-399.

Cohen, J. (1994). The earth is round (p < .05). *American Psychologist,* 49, 997-1003.

Dawes, R. M. (1991). Probabilistic versus causal thinking. In D. Cicchetti & W. M. Grove (Eds.), *Thinking clearly about psychology* (pp. 235-264). Minneapolis: University of Minnesota Press.

Gill, J. (2002). *Bayesian methods.* Boca Raton, FL: Chapman & Hall/CRC

Kirk, R. (1996). Practical significance: A concept whose time has come. *Educational and Psychological Measurement,* 56, 746-642.

Matthews, J. N. S., & Altnam, D. G. (1996). Statistics notes: Interaction 2: Compare effects sizes not p value. *British Medical Journal,* 313, 808.

Mulaik, S. A., Raju, N. S., & Harshman, R. A. (1997). There is a time and place for significance testing. In L. L. Harlow, S. A. Mulaik, & J. H. Steiger (Eds.), *What if there were no significance tests* (pp. 65-116). Hillsdale, NJ: Erlbaum.

Rosenthal, R., Rosnow, R. L., & Rubin, D. B. (2000). *Contrasts and effect sizes in behavioral research.* NY: Cambridge.

Salazar, A. M., Warden, D. L., Schwab, K, Spector, J, Braverman, S., Walter, J., et al. (2000). Cognitive rehabilitation for traumatic brain injury. *JAMA,* 283, 3075-3081. ☑

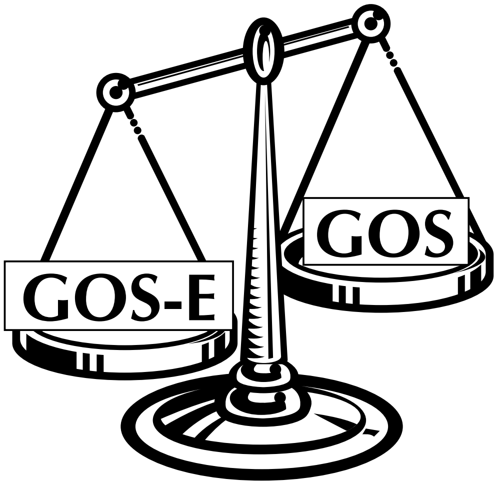
Extended Glasgow Outcome Scale

A New Measure for the COMBI

The Extended Glasgow Outcome Scale (GOS-E) was developed to address the limitations of the original GOS, including the use of broad categories that are insensitive to change and difficulties with reliability due to lack of a structured interview format. The GOS-E extends the original 5 GOS categories to 8. The 8 categories are: Dead, Vegetative State, Lower Severe Disability, Upper Severe Disability, Lower Moderate Disability, Upper Moderate Disability, Lower Good Recovery, and Upper Good Recovery. A structured interview has been provided to improve reliability of rating. Good interrater reliability and content validity have been demonstrated for the GOS-E. Compared to the GOS, the GOS-E has been shown to be more sensitive to change in mild to moderate TBI.

Information regarding the GOS-E was contributed by Angelle Sander, Ph.D. at The Institute for Rehabilitation and Research, Houston, Texas.

Wilson JTL, Pettigrew LEL, Teasdale GM. Structured interviews for the Glasgow Outcome Scale and the Extended Glasgow Outcome Scale: Guidelines for their use. Journal of Neurotrauma 1998;15:573-585. ☑



The GOS-E, definitely more filling.

Assessing The COMBI

Log Files 101

Did you know that every time you access a web page, a record of what you did is created? These records, called log files, give webmasters a lot of information about you and what you looked at on the site. We use the log files to assess how the COMBI is being used.

The Stats

In the last seven months (December 01–June 02) the COMBI has logged in 53,245 visitors. That's over 265 users a day! During this period 142,448 pages of information were reviewed (that's 1,421 megabytes of data).

The COMBI logs show that 86% of our users are within the United States and 14% are from 65 other countries. The COMBI is especially popular in Canada, the United Kingdom, Australia, Italy, and Japan. Our biggest referrals come from MSN.com, Google, the Brain Attack Coalition (www.stroke-site.org), AOL, and Yahoo.

The COMBI newsletter, *Outcome Oriented*, is primarily disseminated in Portable Document Format (PDF) from the website. Over the last seven months, 3,741 newsletters were downloaded by COMBI users.

The COMBI continues to be very successful as a dissemination effort. In the past seven months over 12,000 rating forms were downloaded. Itemized scale activity is summarized in the table below .
But please, no wagering. ☑

Scale Activity (Number of Visitors & Downloads)
December 2001–June 2002

Scale	Visitors	Downloads
ABS	1508	525
AQ	1143	1848
CHART	935	1103
CHART-SF	627	720
CHIEF	565	621
CIQ	1161	682
CNC	1226	914
DRS	1730	279
FAM	1517	1177
FIM	5274	na
FNQ	710	na
GOS	5796	na
GOS-E	510	na
LCFS	1260	363
MPAI	997	1204
NFI	558	na
O-LOG	601	468
PCRS	934	1568
SOS	433	306
SRS	654	388
SWLS	2036	na

Future Directions

This is the last *Outcome Oriented* newsletter for this funding cycle (1997-2002). If this project is refunded, the COMBI will continue to add new measures and act as a resource for the rehabilitation community. All of the current COMBI contributors have offered to continue to support the information available on the website.

We are looking to add more training and testing materials for COMBI measures, and to make the existing materials more interactive (automatic email of results from testing exercises).

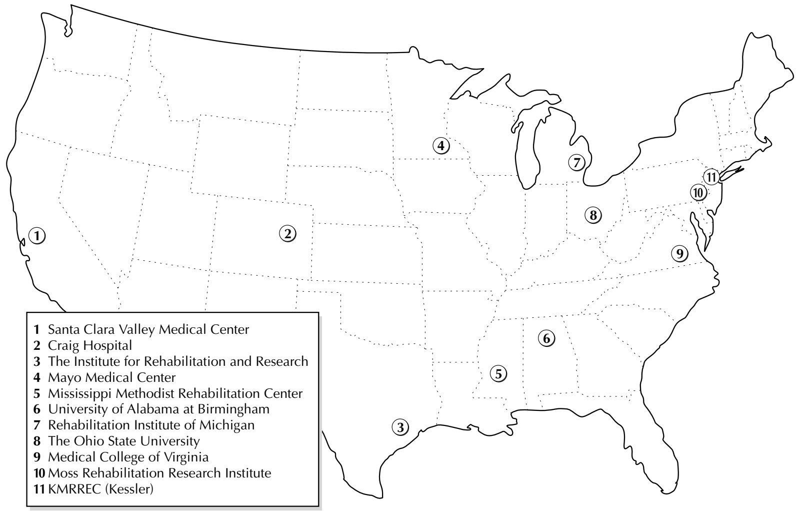
Please email us at <combi@tbi-sci.org> with your thoughts and suggestions. Let us know how we measure up! Thank you for allowing us to be your brain injury outcome measure resource! ☑

Outcome Oriented *is a project of the Center for Outcome Measurement in Brain Injury (COMBI) which is funded by the U.S. Department of Education, Office of Special Education and Rehabilitative Services, National Institute on Disability and Rehabilitation Research.*

Address inquiries to
***Jerry Wright**, Editor. PHONE (408) 295-9896 ext 20; FAX (408) 295-9913; EMAIL combi@tbi-sci.org*

This document is available online at:
<www.tbims.org/combi/combinews.html>

Credit to Our Collaborators


1 Santa Clara Valley Medical Center
2 Craig Hospital
3 The Institute for Rehabilitation and Research
4 Mayo Medical Center
5 Mississippi Methodist Rehabilitation Center
6 University of Alabama at Birmingham
7 Rehabilitation Institute of Michigan
8 The Ohio State University
9 Medical College of Virginia
10 Moss Rehabilitation Research Institute
11 KMRREC (Kessler)

The COMBI is a collaborative project of eleven brain injury centers located across the US. Without the expertise of these centers this project would not be possible. We would like to offer special recognition to the individuals at these facilities who have taken the time to prepare materials for the COMBI and act as contacts:

Tamara Bushnik, PhD, Jerry Wright, BA, Maurice Rappaport, MD, PhD, & Mary Lou Kohlmiller, RN, BSN at Santa Clara Valley Medical Center (Lead Center)

Dave Mellick, MA and Cindy Harrison-Felix, MS at Craig Hospital

Corwin Boake, PhD and Angelle Sander, PhD at The Institute for Rehabilitation Research

James F. Malec, PhD, LP at the Mayo Medical Center

Mark Sherer, PhD, ABPP-Cn at the Mississippi Methodist Rehabilitation Center

Tom Novack, PhD at University of Alabama at Birmingham

Marcel Dijkers, PhD at Mount Sinai School of Medicine (Formerly at the Rehabilitation Institute of Michigan)

Jennifer Bogner, PhD & John D. Corrigan, PhD at the Ohio State University

Jeffrey Kreutzer, PhD and Jenny Marwitz, MA at Medical College of Virginia

Tessa Hart, PhD at Moss Rehabilitation Research Institute

Scott Millis, PhD at Kessler Medical Rehabilitation Research and Education Corporation ☑

SANTA CLARA
VALLEY
MEDICAL CENTER

Rehabilitation Research Center for TBI & SCI
Santa Clara Valley Medical Center
950 South Bascom Avenue, #2011
San Jose, CA 95128

UPDATE
Center for Outcome Measurement
in Brain Injury (COMBI)
<www.tbims.org/combi>

Summer 2002