GENERAL DISCUSSION OF SECTION I

Formal Discussant: Joseph Z bin

I'm afraid I'm here under false pretenses because to summarize these very highly developed methodological and statistical papers is beyond me. Of course it is true that by reputation I'm supposed to be a statistician, but statisticians think I'm a psychologist, and I guess I've pulled wool over both their eyes.

To make this discussion interesting, I will have to provide you with some kind of framework, some kind of a loom on which to weave my comments. The problem discussed this morning, the comparison of controlled vs. naturalistic studies, arises from the fact that most of our methodology, ANOVA, correlational analysis, etc., has been developed for studies with randomly generated samples, and the application of these methods to naturalistic studies has to be justified. This is really the essential problem before us.

It would have been nice of God to have so ordained the world that for every illness there would be a sufficiently large population of identical twins and non-identical twins with large proportions of discordant pairs who were readily accessible for examination. Then our troubles would be over. We could have both a naturalistic and a controlled experiment on natural groups. Since this is not the case, we have to examine the methods of controlled experimentation and see how they can be adapted.

My own inclination is to re-invoke a concept which has been dormant for a long while, actually since 1944 when Slater introduced it, namely the concept of vulnerability to mental disorder. A simple statement of this hypothesis is that the vulnerability becomes manifested depending upon the stressors (both internal and external) which impinge on the person. The person with low vulnerability would require a tremendous amount of stress before he would become ill; once he does so, he'll have an episode; that episode doesn't last forever and when he recovers, he's as well as he was before. The more vulnerable person requires less stress in order to get into an episode, but he, too, soon recovers and
returns to his baseline functioning. But you don't always know that the episode is over in his case because he was so poor to begin with; it looks as if he's still sick. Maybe many of our chronics belong in this category; not that they're still sick, they just were unable to cope before the illness got them, and there's no reason to believe that after the illness episode leaves them that they're any more capable of coping with life than before. This essentially is a framework in which some of the measures to determine vulnerability and life events stressors might best be viewed. A longitudinal prospective would determine how vulnerability develops, and how the sensitivity to life stressors can become greater or less depending upon the variety of experiences to which one is exposed.

Accepting this hypothesis, what implications do papers in this group have for a vulnerability hypothesis? In general, longitudinal research takes on the role of following the development and growth of vulnerability in individuals with different levels of risk, and of determining the forces in the ecological niches that they occupy, that either elicit or inhibit the actualization of the risk. The same ecological niche will give rise to different outcomes depending on the vulnerability of the individual in the niche, while the same degree of vulnerability will lead to different outcomes, depending on the niche's degree of noxiousness. Consequently, we must have measures of both the individual's vulnerability and the parameters of the niche if we are to do significant research. Let me remind you in passing that we have far better measures of vulnerability, such as consanguinity, psychophysiological, neurophysiological, and biochemical measures, than we have measures of the parameters in the ecological niche the person occupies, like stressors and life events.

However, given that we have such measures regardless of their reliability, what has Dr. Overall's panoramic view had to say about the treatment of data? In longitudinal studies difference scores over time with or without intervention loom very large, and the differences that are found between contrasting groups, concordant vs. disconcordant twins, for example, treated and untreated groups, etc., are an important criterion for measuring vulnerability, stressors, and their impact. Thus it might be important to compare the differences in growth and intelligence from age 4 to age 8 in brain damage, with The question raise differences between meaningful difference answer, paradoxical differences in gro a differential bet a maximum. It's h If you accept this have the following sense. If you havent scores for cor the reliable diff so they will give a differential bet

Dr. Joseph FJ with Biometrics Res one for a long til examined Overall's we believe we have assume that there time in their rep subject achieved except for errors is constant betwe measurement. No it's 50, then at it's 40. It's a interaction means Hence, if you cor each individual, supplemented by e dent, they corre late, the end res As a matter of fa scores many, many would be left wit

If you make emerge with Overs the more realistic action, that is, action in the diff
age 4 to age 8 in children who have experienced minimal brain damage, with the growth in comparable normal controls. The question raised by Dr. Overall is how reliably measurements are made, and the reliability of the answer paradoxically is that the reliabilities of the differences in growth are zero, and the line on the difference in the growth is a maximum. It is hard to believe, but that's what he said.

Dr. Joseph Plotnik, head of the Psychometrics section with Biometrics Research Unit, said it is a very important point. We believe there is no correlation of the differences in growth between the mean and the reliability of the differences in growth are zero, and the line on the differences in the growth is a maximum. If you accept this conclusion at its face value, you would have the following: for a given difference in the mean of two groups, the reliability of the difference of the mean of two groups would be limited to the maximum of those extremes. If you want to find...
difference produces the highest differentiation between groups. Thus, contrary to what Overall states, the reliability of difference scores has the highest relevance for comparison of treatment effect or other group differences.

The following is the situation with respect to Overall's view concerning the analysis of co-variance applied to unreliable covariates. The authors cited by Overall make a simple assumption that there is no treatment effect, that there is no differential from pre- to post-test between the groups, and proceed to demonstrate that because of the error of measurement that's present, there may appear a significant difference between groups which are in fact similar. Overall, on the other hand, does not deal with this assumption; he assumes that there is a difference between groups and proceeds to demonstrate that the presence of errors does not matter. It is clear that an observed differential between the groups can be due either to the error of measurement following Kahneman (Control of Spurious Association) and reliability of the the controlled variable (see Overall's bibliography) or to the factors beyond errors of measurement, following Overall, to reach a conclusion that the two approaches must be considered together, but unfortunately Overall is unilateral in this treatment, neglecting the Kahneman approach.

Another problem which Overall tackles is that of non-parallel regression lines in analysis of covariance, which is a problem we are often met with. Overall's suggested remedy for non-parallel regression lines in analysis of covariance, namely to stratify on the covariate, is most inefficient. Nonparallel lines necessarily imply treatment effects; but of different magnitude and different direction, depending on the value of the covariate. To stratify is to ignore the quantitative nature of the covariate, to give up too many degrees of freedom, and most seriously to fail to apply an analysis method especially developed to handle the situation, namely the Johnson-Neyman method. What this method provides for the case of two groups is a partition of the range of the covariate into three regions. In the lowest region, one of the groups may be said to have a significantly lower mean than the other; in the highest region, it may be said to have a definitely higher mean than the other; in the middle region, generally corresponding to
those values of the covariate near to where the two lines intersect, no significant difference between the groups can be declared. The algebra necessary to carry out the Johnson-Neyman method is more complicated than the algebra for Overall's suggested remedy, but the payoff is much more precise and usable.

The cautions which Overall points to are well taken and should be followed by all those who tilt with ANCOVA. The areas for legitimate use of ANCOVA to eliminate the initial differences between naturally formed groups and covary differences are still very debatable. As Meehl has pointed out in his nuisance variables and ex post facto design, the essential purpose of ANCOVA is to determine what the result would be if the groups were made comparable with respect to the uncontrolled variables, and as Anderson points out, one may well wonder what exactly it means to ask what the data would be like if they weren't what they are.

Instead of resorting to ANCOVA, the earlier workers turned to matching individuals in the two contrasting groups on the covariate; but this led to selected samples, as Dr. Overall pointed out, on the basis of the fortuitous presence of matchable pairs and to restrictions on generalization from unrepresentative samples as well as to differences due to regression effects in the two groups.

Fleiss and Tanur (1973) have provided another solution which seems to avoid most of the difficulties of matching and ANCOVA. They base their method not on the performance of individual subjects, but rather on the average performance of many different groups of subjects. Perhaps the classic example of this is an article on the elements of generalized or lawful relationships between height and weight by A.S.C. Ehrenberg in 1968 in the Journal of the Royal Statistical Society, in which he takes very distinct groups of children with different ages and different other variables and by taking the log weight against height, he points out that the means that all fall in a straight line, and thus describe a straight line which relates height and low weight across many different kinds of groups. The kind of groups he has are 13-year-olds from Birmingham, 7-year-olds from Ottawa, 5-year-olds from Birmingham again, and so on. If you were to find a group of children whose measures did not fall along
that line, then you can look further to see why they deviate significantly.

What we have done in a similar situation is to apply this method to behavioral data. It has been known for a long time that schizophrenics have a slower reaction time than normals, partly as a function of lack in motivation and so on. Some work has actually been able to shorten the reaction time in schizophrenics to normal, if they use methods for increasing motivation. We presented in a group of schizophrenics and normals, light and sound in a variety of random sequences, and then compared the response to light following a light (ipsi modal) to the response to light following sound (cross modal). In this way we randomized the effect of motivation. We discovered that the response to light following a sound was longer than the response to light following a light. Because of problems in the original diagnostic assignments, we reviewed the diagnostic status of the entire group and discovered that there were some schizophrenics who were not schizophrenic, and there were some normals who weren't normal; so we had four groups: "pure normals," "impure normals," "pure schizophrenics," and "impure schizophrenics". We did a covariance analysis and there was no difference among the groups. Then we applied the Fleiss-Tanur method. We discovered with this method that the schizophrenics were distinctly different. This is essentially not a new, but a different technique for getting away from errors of measurement and getting away from intra-group regression lines. We're not worried about intra-group relationships; all we're worried about is do the mean relationships between groups fall along a straight line?

This matter of reliability introduces a very interesting old classic problem, the placebo problem. Why is it that you always get about 30% of the placebo group improved, and furthermore, why are they not always the same people on repeat trials? If you assume that the original reliability of the classification measure is of the order of 57%, then it follows mathematically that 30% will show improvement, no matter what the treatment. And furthermore, if you look at outcome data over the centuries, you usually find that one-third of the patients get better, one-third get worse, one-third remain unchanged. If you assume that the reliability of the diagnosis of normal/abnormal and for outcome is approximately .57, again you will come up with one-third, one result no matter how this unreliable 1962).

Let me say, I think it's a table method with field. It's true and for first-used it in the is it so important for one reason, at a particular the dead, the Dr. Fleiss and use in follow-up 1976).

Open Discussion

Overall: the concept of magnitude of individual difference in measure of significance to minimize the the magnitude of moment models, ex a true component individuals in th component. The inherent in get a more precise measuring with the components between subjects, liability does a simple subjects standable to me difference score change. If sub entirely the tr
DISCUSSION OF SECTION I

one-third, one-third, one-third as a necessary statistical result no matter what the treatment amounts to. So we have this unreliability problem with us most of the time (Windle, 1962).

Let me say one or two words about Myrna Weissman's paper. I think it's an excellent idea to bring back the old life table method which has been neglected for so long in our field. It's true, we have used it for mortality analysis and for first-admission analysis, but until now we haven't used it in the follow-up studies that have been done. Why is it so important to have the life table approaches? Well, for one reason, it includes all data. You don't have to stop at a particular point in time, you don't have to throw away the dead, the drop-outs - they're all included in the picture. Dr. Fleiss and his colleagues have recently illustrated its use in follow-up studies of manic-depressives (Fleiss, et al., 1976).

Open Discussion

Overall: In response to your points about reliability, the concept of reliability is concerned with the relative magnitude of the components of variance due to true individual differences and the components of variance due to random measurement error. With regard to the power of tests of significance of change, one way to maximize that power is to minimize the experimental error, so we're concerned with the magnitude of the experimental error. In terms of measurement models, experimental error is composed of two components: a true component due to the true differences between the individuals in the sample, and in addition, a measurement error component. The measurement error component is essentially inherent in the measuring instrument, and it always pays to get a more precise measuring instrument. But given a particular measuring instrument, then we should still be concerned with the component of variance due to true differences between subjects, or what I'll call sampling error. Now reliability does depend on the magnitude of interaction in a simple subjects by time design; however, it is understandable to me how you could say that the more reliable your difference scores, the more useful they are in assessing change. If subtracting one score from another would remove entirely the true component, the true differences between
subjects, and leave only the measurement error as a source of variability of the difference scores, you will have reduced your experimental error to a minimum insofar as it is possible to do so with that measuring instrument. But at the same time, if you achieve that you have zero reliability. Thus, the closer our difference scores approach zero reliability for a given measuring instrument, the better they are in terms of measuring significance of change. I won't comment on the analysis of covariance problems, since we do have a rather extensive treatment of that in another manuscript. We have also in press a more technical reply to Dr. Fleiss in Psychological Bulletin, 1976, where any interested person can see your criticisms of the change-score reliability issue debunked.

Zubin: Well, I'll hide behind my excuse of being a psychologist. I think that if you have a yardstick that contracts and expands, it is unreliable, as compared to the yardstick that stands up rigidly and gives you the same measure each time. Just on the basis of common sense, I wouldn't take that rubber yardstick, I'd take the rigid yardstick.

Overall: It is the reliability of difference scores, though, Dr. Zubin, not the reliability of the rigid measuring instrument.

Zubin: Dr. Overall should be aware that we are not stuck with the measurement error component of variance for a given measuring instrument; we can make it as small as we like by obtaining as many repeat determinations on each subject, as necessary. We are, however, stuck with the interaction component of variance. On every psychologically meaningful variable different subjects change by different amounts. I would be suspicious of a measuring instrument for which this was not the case.

Formal Discussant: Fini Schulsinger

I'll start out by congratulating the people here in Rochester with the Monroe County Register. That's a very good thing and can be used for many, many things. As you know, we have a register in Denmark and we have utilized it for many purposes. I'll give you a little list of things we have used this old register for. As you know, we have done
family studies using adopted schizophrenics. We have screened the adoptive population, and we have screened their parents in the psychiatric register. We have been doing studies of adoptees who have a biological parent who is alcoholic. One other question we are studying as well is an issue that has been discussed for years; namely, whether the hallucinogenic psychoses from LSD and amphetamines are psychosis per se, or if they only happen in people who have a family disposition for schizophrenia. We are doing a study on that using the register where we can easily pick out 500 LSD and amphetamine psychoses over the last few years and do a family study. So that's very nice and to further tell how good it is, I'll mention the study by Tsuang and Winokur because that is a good illustration of how difficult it is to do studies when you have not got a psychiatric register. It's a big, big, big work to do studies like their study. If they had had a psychiatric register in Iowa City, they wouldn't have had to do all that work.

I conceive of the Iowa 500 study as a technically well-thought out repetition of many, many earlier studies using follow-up and family histories, and it might very well yield a relatively exact version of the natural history of schizophrenia and of affective disorder in a changing era during which ECT, insulin treatment, and neuroleptics were introduced. It would help out, though, I think, if the investigators structured their research to test a known hypothesis in this field and also some sociological hypothesis; I think the outcome would be more precise if the analyses were designed to test certain hypotheses over this long span of changing years.

What we are doing in Denmark is an aggregate of studies, and there is one common paradigm used. That is the paradigm of an old Danish geneticist, Johansson: the genotype plus the environment results in the phenotype. And most of the research which is going to be discussed later in this meeting can be viewed with this model as a background. Let me give you just a small example of how flexible it is and how nice it is. For instance, we did a family study using psychopathic children who were adopted away in early life, and we wanted to test a hypothesis that brain damage as a result of perinatal complication might be an etiological factor. We had the psychopaths, we had the birth histories, and the types and prevalence of mental illness in the biological
relatives. The hypothesis then was that if such complications are an etiological factor, you would expect less character disorder in the biological relatives of those psychopaths who had had complicated births. You have the genotype (a family score of mental illness) as the dependent variable. You have the environment, which is well defined because it is a perinatal score. And you have the phenotype that is diagnosis, so it is under control, everyone is a reliably-diagnosed psychopath. If you try this model out on other types of studies as well, you will find that any of the three parts of the model can be either the dependent variable, or deferred or be under control.

Open Discussion

Thomas: With regard to ethical issues in longitudinal research, a point made by Dr. Weissman, I think that the nature of longitudinal studies, at least certain of them, is such that ethical issues of great importance inevitably arise. Even if you have a study such as we have in which you gather data on a population over time, and then a certain number of the youngsters develop behavior problems, you're faced with the question, do you intervene and advise the parents on the basis of the information you have? If you don't intercede, what are the ethical issues concerned? Many such ethical issues come up in long-term longitudinal studies that don't come up in short-term studies.

Regarding the difference between naturalistic and controlled studies, in many long-term longitudinal studies the differences in care that some children get may provide a blend of the naturalistic and experimental. You may start with a naturalistic study, in which you gather behavioral data without controls, yet you may get differential outcomes with differential life experiences occurring which provide you essentially with the possibility of an experimental situation occurring naturally. For the converse, in longitudinal experimental or controlled studies, over time changes do occur aside from the changes we think we're introducing by our own variables that we control, so that these studies become more naturalistic.

Weissman: Regarding ethical issues in longitudinal research, I think that in psychiatry we're not at the point...
DISCUSSION OF SECTION I

for example, where the Framingham longitudinal study of heart disease is. As a result of the Framingham study, we know what the risk factors are for heart disease, and we can define who has a high probability of developing heart disease and who might be treated with preventive measures. We don't quite know this in psychiatry. Hopefully in the future we will be able to define which portion of the well population is at high risk for developing a psychiatric disorder. In a longitudinal study in psychiatry, if you have a child who develops a disease, then you would recommend treatment for him; but we don't know enough about the risk factors for psychiatric disorders and say who among the well persons we should treat as a preventive measure.

Zubin: There's one more point, I believe that needs to be said here. It isn't so much a question of whether naturalistic studies are good or not good. The real issue methodologically is do we have the techniques for examining the data that we obtain on naturally formed groups? For example, the application of the covariance method to naturalistic groups has been questioned. I would not be surprised if some of the designs using analysis of variance might also be criticized. In other words, the question is how can we translate from the randomization approach in experimentation to observational studies, and I think that in itself is worthy of a session sometime.

Strauss: I think it works the other way, too, that sometimes the controlled study may eliminate a lot of the variables that you don't yet know are important. Carrying out controlled studies without having a good mapping of a large area using the naturalistic approach first can lead to a lot of error and misunderstanding. There may be variables that are crucial determinants of results that you have no idea of. I think there is a reciprocity between the naturalistic and controlled designs. In the naturalistic study, you should collect a wide range of data, too. Deciding to carry out controlled studies should usually imply that you have an idea what the key variables are, and that they are the ones you're looking at.

Klein: There are two kinds of study - a hypothesis-generating study, and a hypothesis-testing study. What John Strauss has said is that you have to have some good hypothesis generating studies before you get to the hypothesis testing stage, and I think that's scientific common
sense. The question is, to what degree can you test hypothe-
ses on the basis of non-experimental designs. I think that
the approach that John Overall has taken, is, that he has
moved away from the idea of testing hypotheses to model
fitting. What he's saying is that if he can generate models
and show you that his model fits the naturalistic data, it
may not be the right model, but at least it fits the data
pretty well, so we ought to pay attention to that model. I
think that's what he's doing, but that's not my idea of a way
to test hypotheses, but rather of a way to generate them.

Fleiss, J.L., and Tanur, J.M. The analysis of covariance
in psychopathology. In Hanuver, M., Salzinger, K., and
Sutton, S. (Eds.), Psychopathology, contributions
from the social, behavioral and biological sciences.

Windle, C. Prognosis of mental subnormals: a critical re-
view of research. American Journal of Mental Disease,

Fleiss, J.L., Dunner, D.L., Stallone, F., and Fieve, R.R.
The life table: a method for analyzing longitudinal
studies. Archives of General Psychiatry, 1976,
33, 107-112.