ON MORE POWERFUL TESTS OF JUDGE AGREEMENT WITH A KNOWN STANDARD

DAVID H. ROBINSON
ST. CLOUD STATE UNIVERSITY

To establish the existence of his abilities, a judge is given the task of classifying each of \( N = rs \) subjects into one of \( r \) known categories, each containing \( s \) of the subjects. An incomplete design is proposed whereby the judge is presented with \( b \) groups, each one containing \( n = rs/b < r \) subjects. The \( n \) different categories corresponding to members of the group are known. Using the total number of correct classifications, this method of grouping is compared to that in which the group size is equal to the number of categories. The incomplete grouping is shown to yield a more powerful test for discriminating between the null hypothesis that the judge is guessing the classifications and the alternative hypothesis that he has some definite abilities. The incomplete design is found to be most effective (powerful) when the number of subjects in a group is limited to two or three.

Key words: nominal data, power comparisons, incomplete design.

1. Introduction

Measuring judge agreement with a known standard has been a problem appearing in statistical journals and literature for some time. In the typical experimental setting, a judge is presented with a group of \( N \) objects, hereafter referred to as “subjects,” each of which is to be classified into one of \( r \) categories. For example, Fisher’s tea-tasting lady (1971) was presented with eight cups of tea and asked to identify each as belonging to one of two different groups, according to the tea-making process. (See Tocher, 1950, and Gridgeman, 1959 for discussions of this experiment.)

More recently, Wackerly, McClave, and Rao (1978) and Wackerly and Robinson (1983) dealt with the question of designing a more powerful experiment to test for nominal scale agreement between a judge and a known standard. The original experiment motivating their research (Blume, 1977) was one wherein 16 subjects at each of three anxiety levels were asked to describe their dreams on tape. A clinical psychologist (judge) was given the task of classifying the 48 tapes into three groups of 16, according to judged anxiety levels. Wackerly and Robinson (1983) advised using a design in which the judge is presented with the tapes in groups of three, consisting of one representative from each of the three categories. Such a design yields a more powerful test for disproving the “guessing” hypothesis, that the judge makes random assignment of subjects to each category.

In this paper, evidence is presented in favor of using small groups of two or three subjects to be classified at a time, even when the number of categories in the experiment is greater than three. This design shall be called an incomplete design, since not every category will be represented within a group. The design presented in Wackerly and Robinson (1983) will be called a complete design. In the incomplete design, a judge is presented with a small group of subjects, for instance two at a time, and is told which two categories correspond to members of the group. The judge’s task is to identify which of the two subjects belongs to each of the two given categories. In such a case, either both of the

The author is grateful for the suggestions of the referees and the editor, which greatly improved the paper.
Requests for reprints should be sent to David H. Robinson, Department of Mathematics, St. Cloud State University, St. Cloud, MN 32611.

0033-3123/85/0900-7039$00.75/0
© 1985 The Psychometric Society
judge's classifications would be correct or both incorrect, as exactly one subject must be classified into each of the given categories.

In section 2, the mathematical model and resulting probabilities are introduced. Section 3 contains power comparisons for the incomplete and complete designs. Conclusions are found in section 4.

2. The Mathematical Model

Throughout the paper, the same notation is utilized as in Wackerly and Robinson (1983) for the complete design, denoted Design C. Assume there are r categories with s subjects belonging to each category. In Design C, the judge is presented with s distinct groups, each containing r subjects, one from each of the r categories. The judge is to consider each group separately and classify the r subjects in each group into the r categories.

Let \( \theta_{ij} \) represent the probability that, under the complete design, a subject from category \( i \) will be judged as being of type \( j \). Note that

\[
\sum_{j=1}^{r} \theta_{ij} = 1 \quad \text{for } i = 1, 2, \ldots, r.
\]

Design C employs the following assumptions:

(i) Within each group, the judgments are made at random subject to the values of \( \theta_{ij} \) and the constraint that exactly one item of each type appears in each group.

(ii) The number of correct judgments in any single group is independent of the number of correct judgments in any other group.

The incomplete design shall be denoted by Design I. In this design, it is assumed the subjects are presented in groups of size \( n \) (\( n < r \)) for classification by the judge. For purposes of comparison with Design C, it is assumed there are \( s \) subjects of each type, although this is not an inherent requirement of Design I. Thus, the \( N = rs \) subjects are divided into groups of \( n \), yielding \( b = N/n \) such groups. All \( n \) subjects within a group are of distinct types. If possible, the grouping may be done in a balanced fashion, so that each pair of types appears together in the same number of groups. (Such an arrangement is analogous to a balanced incomplete block design, with a group of \( n \) subjects contained in a block. Design C is then analogous to a complete block design.)

The judge, having been presented with a group of \( n \) subjects, is told which \( n \) categories correspond to the group, and is asked to classify the individual subjects accordingly. Let \( p_{ij} \) denote the probability, under Design I, that a subject of type \( i \) will be judged to be in category \( j \). Then for the \( k \)th group,

\[
\sum_{j \in G_k} p_{ij} = 1 \quad \text{for } i \in G_k \text{ and } k = 1, 2, \ldots, b,
\]

where \( G_k \) is the subset of \( \{1, 2, \ldots, r\} \) corresponding to the categories present in the \( k \)th group.

Some simplifying assumptions are needed. Let us assume that (i) and (ii), stated above for Design C, also hold for Design I. In addition, assume

(iii) For a given \( n \) and for any \( i \) and \( j \), the probability \( p_{ij} \) is the conditional probability of classifying a subject of type \( i \) as being type \( j \), given the \( n \) categories known to be represented in the group; i.e.,

\[
p_{ij} = \frac{\theta_{ij}}{\sum_{l \in G_k} \theta_{il}}.
\]