EDUCATIONAL AND PSYCHOLOGICAL MEASUREMENT
1993, 53

SIMPLE AND HIERARCHICAL MODELS FOR STOCHASTIC TEST MISGRADING

JIANJUN WANG
Center for Science Education
Kansas State University

The test misgrading is treated in this paper as a stochastic process. Three characteristic values, the expected number of misgradings, the expected inter-occurrence time of misgradings, and the expected waiting time for the *n*th misgrading, are discussed based on a simple Poisson model and a hierarchical Beta-Poisson model. Regular tests for classroom teaching and the Test of Written English (TWE) in the Test of English as a Foreign Language (TOEFL) are illustrated as examples in the model construction.

TEST misgrading is a fundamental problem in educational and psychological measurement. Even for well designed tests such as the Test of Written English (TWE) in the Test of English as a Foreign Language (TOEFL) (Educational Testing Service, 1992), scoring errors might still be unavoidable in a long-time hand-grading process (Braun, 1988). The expected number of misgradings is an important index for a test score evaluation. The expected inter-occurrence time of misgradings is another reference index for administrators who need to set a time table for a test grading. Because there is no expected misgrading happening in the inter-occurrence time, a reasonable waiting time for a short break is desired to be longer than the inter-occurrence time. On the other hand, given a time table, a researcher might want to know how many misgradings occurred before the short break. In general, the answer should come from a comparison between the expected waiting time for the *n*th misgrading and the waiting time for the short break. Hence, the information of interest for a test grading is three-fold:

1. Expected number of misgradings;
2. Expected inter-occurrence time of misgradings;
3. Expected waiting time for the nth misgrading.

Purpose

There are many reasons for explaining a test misgrading. Nevertheless, none of them has a deterministic relationship with the three-fold information. In practice, the number of misgradings is counted from a probabilistic misgrading process. Thus, an appropriate stochastic model is required for extracting the information about the misgrading. The idea of the stochastic approach is perhaps not a new one. As no account of it has been found in the literature so far, the purpose of this paper was to present stochastic models based on a simple Poisson process and a hierarchical Beta-Poisson approximation.

Stochastic Models

Simple Poisson Model

A simple method to discuss a test misgrading is to model it as a coin-tossing process. The events of head and tail correspond to the result of correct grading and misgrading respectively.

For a test taken by n students, the result of the grading is equivalent to the result of tossing a coin n times. Since a single coin tossing is a Bernoulli trial, the result of tossing a coin n times has a binomial (n, p) distribution with p equal to the probability of a tail in each trial. For most well designed tests, the p is small. Hence, the tossed coin is unbalanced, and the binomial (n, p) distribution can be approximated by a Poisson distribution with parameter λ equals np.

Poisson process is an elementary stochastic process. The three-fold information has been readily solved in most stochastic textbooks (Bhat, 1984). The information is summarized in Table 1.

TABLE 1
The Three-fold Information from the Simple Poisson Model

1. Expected number of misgradings is λ with variance equal to λ;
2. The expected inter-occurrence time of misgradings is t/λ with variance equal to t^2/λ^2;
3. The expected waiting time to the nth Poisson occurrence is nt/with variance equal to nt^2/λ^2;

Where t = Total time spent in test grading.

The application of the Poisson model is straightforward. Under a regular grading environment, a grader may choose a well designed test and grade a number of students. Then, the grader might check the proportion of the misgradings and estimate the parameter λ in the Poisson model. The next time when the grader is grading a similar test, the estimated parameter can be used to identify a distribution from the Poisson family, and the three-fold information can be inferred approximately by using the results in Table 1.

Refinement of the Simple Poisson Model

According to the simple Poisson model, the test misgrading can be simulated by tossing an unbalanced coin many times. However, in regular classroom teaching, a test may contain several items. As the items are mutually different, the item misgradings for each student are essentially equivalent to the results of tossing a set of different coins. For a test taken by more than one student, the simulation becomes more complicated because of the increased coin sets. Fortunately, the complexity does not come into picture in most cases when the student achievement is evaluated by the total score of the whole test. At the whole test level, each student receives the same set of items. Thus, the test misgrading can be modeled as the result of tossing a single unbalanced coin many times. In other words, instead of dealing with the item misgrading, the simple Poisson approximation is a pertinent model for the misgrading at the whole test level.

Another limitation in the simple Poisson model is that only one grader is considered in the grading process. However, in a large scale test, such as the TWE in TOEFL, a group of graders is needed to evaluate the free-response tests independently. Because of the individual differences, the probability of misgrading will naturally vary from person to person. The total number of misgradings in such a case is a summation of the misgradings made by the group of graders.

One may define x_i to be the number of misgradings made by the ith grader, λi to be the parameter in the simple Poisson model, and m to be the number of graders in the grading group. The simple model suggests that each of the x_i's has a Poisson (λ_i) distribution, $i = 1, \ldots, m$. Hence, the momentum generating function (mgf) for the total number of misgradings $(\sum_{i=1}^{m} X_i)$ follows (Casella and Berger, 1990):

$$M_{\sum_{i=1}^{m} X_i}(t) = \prod_{i=1}^{m} e^{\lambda_i (e^t - 1)} = e^{(\sum_{i=1}^{m} \lambda_i)(e^t - 1)} \quad (1)$$

In contrast, the mgf for a Poisson (λ) random variable x is:

$$M_x(t) = e^{\lambda(e^t - 1)} \quad (2)$$

A comparison of Equation (1) and (2) leads to the conclusion that the total number of misgradings $\sum_{i=1}^{m} X_i$ has a Poisson distribution with the parameter equal to $\sum_{i=1}^{m} \lambda_i$. Because the large scale test has a different Poisson parameter, the results in Table 1 need to be adjusted to reflect the parameter change. The three-fold information for the large scale test is listed in Table 2.

TABLE 2
The Three-fold Information for M-grader Test Grading

1. Expected number of misgradings is $\sum_{i=1}^{m} \lambda_i$ with variance equal to $\sum_{i=1}^{m} \lambda_i$;
2. Expected inter-occurrence time of misgradings is $t/(\sum_{i=1}^{m} \lambda_i)$ with variance equal to $t^2/(\sum_{i=1}^{m} \lambda_i)^2$;
3. Expected waiting time for the nth misgrading is $nt/(\sum_{i=1}^{m} \lambda_i)$ with variance equal to $nt^2/(\sum_{i=1}^{m} \lambda_i)^2$;

Where t = The total grading time for each of the graders.

A further problem in the Poisson model arises from the fluctuation on the probability of misgradings. The individual differences among graders may cause the probability variation for a large scale test. However, for a regular test graded by one person, the fluctuation could also be possible due to the changes of the individual's mental and physical conditions. For example, although the probability of misgrading for a well designed test tends to be small, it could be adjusted to a large value for the effects of distraction, fatigue, or lack of training. As many such changes are unpredictable, the probability of misgrading should be treated as a random variable in general. To account for the probability fluctuation in a one-grader test grading, a Bayesian hierarchical model is needed to merge the two random variables, the Poisson count of misgrading and the Poisson parameter fluctuation.

Hierarchical Beta-Poisson Model

The Poisson parameter λ for a regular classroom examination is decided by the probability of a grader's misgrading. The parameter, as a random variable, has a pattern of distribution described by a probability density function (pdf). Since the misgradings are made

by human beings, modern neurophysiology should be resorted to construct the pdf definition.

The operation of a neuron system is controlled by a firing threshold (Lawson and Staver, 1989). A neuron system is "on" if an incoming signal generates the level of the neuron activity to exceed the firing threshold; otherwise, the neuron system is "off." In the context of the test misgrading, the "on" of the grader's neuron system corresponds to a small possibility of misgrading, and results in a small parameter value for the simple Poisson model. The "off" state, on the other hand, corresponds to a large parameter value. Because a neuron system has only two states, the probability of finding the parameter with a medium value is close to zero. Based on the knowledge of a neuron system, the fluctuated Poisson parameter has a U-shaped probability density function.

Among the most commonly used statistical distributions, the Beta (α,β) family has a U-shaped distribution when $\alpha < 1$, and $\beta < 1$. Relevantly, the constrained Beta distribution can be chosen to model the parameter fluctuation. The Bayesian method is applicable at this point when the distributions have been identified for both the simple misgradings and the parameter fluctuation.

One may define x to be the number of misgradings, τ to be the inter-occurrence time of misgrading, w to be the waiting time for the nth misgrading, λ to be the Poisson parameter, t to be the total time spent in the test grading, and $\sim$ to represent the phrase "is distributed as". Then one may write:

$(x|\lambda) \sim$ Poisson (λ); $(\tau|\lambda) \sim$ Exponential (t/λ); $(w|\lambda) \sim$ Gamma $(\eta, t/\lambda)$; and $\lambda \sim$ Beta (α, β).

The expected values and variances which are of interest to researchers can be calculated as follows:

$$\mathrm{E}(x) = \mathrm{E}[\mathrm{E}(x|\lambda)] = \mathrm{E}(\lambda) = \frac{\alpha}{\alpha + \beta};$$

$$\mathrm{Var}(x) = \mathrm{Var}[\mathrm{E}(x|\lambda)] + \mathrm{E}[\mathrm{Var}(x|\lambda)] = \mathrm{Var}[\lambda] + \mathrm{E}[\lambda]$$

$$= \frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta - 1)} + \frac{\alpha}{\alpha + \beta};$$

$$\mathrm{E}(\tau) = \mathrm{E}[\mathrm{E}(\tau|\lambda)] = \mathrm{E}\left[\frac{t}{\lambda}\right] = \frac{t(\alpha + \beta - 1)}{\alpha - 1};$$

$$\mathrm{Var}(\tau) = \mathrm{Var}[\mathrm{E}(\tau|\lambda)] + \mathrm{E}[\mathrm{Var}(\tau|\lambda)] = \mathrm{Var}\left[\frac{t}{\lambda}\right] + \mathrm{E}\left[\frac{t^2}{\lambda^2}\right]$$

$$= 2E\left[\frac{t^2}{\lambda^2}\right] - \left\{E\left[\frac{t}{\lambda}\right]\right\}^2$$

$$= \frac{2t^2(\alpha+\beta-1)(\alpha+\beta-2)}{(\alpha-1)(\alpha-2)} - t^2\left(\frac{\alpha+\beta-1}{\alpha-1}\right)^2;$$

$$E(w) = E[E(w|\lambda)] = E\left[\frac{nt}{\lambda}\right] = \frac{nt(\alpha+\beta-1)}{\alpha-1};$$

$$\mathrm{Var}(w) = \mathrm{Var}[E(w|\lambda)] + E[\mathrm{Var}(w|\lambda)] = \mathrm{Var}\left[\frac{nt}{\lambda}\right] + E\left[\frac{nt^2}{\lambda^2}\right]$$

$$= n^2t^2\left\{E\left(\frac{1}{\lambda}\right)^2 - \left[E\left(\frac{1}{\lambda}\right)\right]^2\right\} + nt^2E\left(\frac{1}{\lambda^2}\right)$$

$$= \frac{n^2t^2(\alpha+\beta-1)(\alpha+\beta-2)}{(\alpha-1)(\alpha-2)} - \frac{n^2t^2(\alpha+\beta-1)^2}{(\alpha-1)^2}$$

$$+ \frac{nt^2(\alpha+\beta-1)(\alpha+\beta-2)}{(\alpha-1)(\alpha-2)}$$

The three-fold information is summarized in Table 3.

Discussion

It has been assumed in this paper that there existed no purposeful misgradings. The number of misgradings was treated as a random variable, and the stochastic approach was employed to model the misgrading for a well-designed test. For mathematical simplicity, the Poisson model considered in this paper illustrated the model construction for a regular classroom test and the TWE test without acknowledgement of the parameter fluctuation. The hierarchical Beta-Poisson model postulated that the parameter fluctuation follows a symmetric or non-symmetric U-shaped distribution. However, the three-fold information has not been solved for the hierarchical model in most stochastic textbooks. The results in this paper were extracted by a Bayesian approach, and two parameters, α and β, remained to be estimated from a specific situation in practice.

The test misgrading could have been modeled as a binomial distribution. The reason for choosing the Poisson approximation is that the critical values for a Poisson distribution have already been tabulated. Normal distribution is another way to approximate the

TABLE 3
The Three-fold Information from the Hierarchical Model

1. Expected number of misgradings is $\frac{\alpha}{\alpha + \beta}$ with variance equal to

$$\frac{\alpha\beta}{(\alpha + \beta)^2 (\alpha + \beta - 1)} + \frac{\alpha}{\alpha + \beta};$$

2. Expected inter-occurrence time of misgradings is $\frac{t(\alpha + \beta - 1)}{\alpha - 1}$ with variance equal to $\frac{2t^2 (\alpha + \beta - 1)(\alpha + \beta - 2)}{(\alpha - 1)(\alpha - 2)} - t^2\left(\frac{\alpha + \beta - 1}{\alpha - 1}\right)^2;$

3. Expected waiting time for the nth misgrading is $\frac{nt(\alpha + \beta - 1)}{\alpha - 1}$ with variance equal to $\frac{n^2t^2(\alpha + \beta - 1)(\alpha + \beta - 2)}{(\alpha - 1)(\alpha - 2)} - \frac{n^2t^2(\alpha + \beta - 1)^2}{(\alpha - 1)^2} + \frac{nt^2(\alpha + \beta - 1)(\alpha + \beta - 2)}{(\alpha - 1)(\alpha - 2)};$

Where t = Total time spent in test grading.

binomial distribution. Nevertheless, based on modern neurophysiology, the probability of misgrading, which is either small or large in most cases, rarely takes a value around 0.5. Therefore, compared with normal distributions, the Poisson model is a more precise approximation to the binomial distribution.

REFERENCES

Bhat, N. U. (1984). *Elements of applied stochastic processes* (2nd ed). New York: Wiley.

Braun, H. I. (1988, Spring). Understanding scoring reliability: Experiments in calibrating essay readers. *Journal of Educational Statistics*, pp. 1-18.

Casella, G. and Berger, R. L. (1990). *Statistical inference*. Pacific Grove, CA: Brooks/Cole Publishing Company.

Educational Testing Service (1992). *Bulletin of information for TOEFL/TWE and TSE*. Princeton, NJ: Author.

Lawson, A. E. and Staver, J. R. (1989). Toward a solution of the learning paradox: Emergent properties and neurological principles of constructivism. *Instructional Science*, 18, 169-177.