

An Optimal Estimator of True Mark under Double Blind Marking

Oliver Linton*
London School of Economics

June 23, 2004

1 The Issue

At this time of year when we are buried in marking and have to try to agree on the ‘right mark’ for candidates I thought I should supply a statistical solution to this problem, that is the problem of combining two independently produced pieces of evidences on the candidates ability.

We adopt the following statistical model for the generation of the marks. We observe the marks X_{ij} for candidate i by marker j , where typically the number of candidates n can be large [too large] and $j = 1, \dots, J$, where J is the number of markers typically $J = 2$. The model is

$$X_{ij} = \mu_i + \varepsilon_{ij}, \quad (1)$$

where μ_i is the true score of the candidate and ε_{ij} is the marker/candidate specific measurement error. The ε_{ij} are considered random variables whose realisation is different across experiments. We shall suppose that ε_{ij} is independent across i and across j . We shall further suppose that $\varepsilon_{ij} \sim N(0, \sigma_j^2)$, where σ_j^2 is the marker specific error scale. This model is consistent with Edgeworth’s (1888) definition of true mark as the average value that would occur if infinitely many independent markers evaluated the script. A special case is where $\sigma_1^2 = \sigma_2^2$. Allowing for $\sigma_1^2 \neq \sigma_2^2$ seems more reasonable since we do find in practice quite large variation in $\text{var}(X_{ij})$ across j when the marking is conducted blind.

When $J = 1$, i.e., there is a single marker, it is not possible to determine from the observed marks what is the relative contribution of true score and measurement error to the overall variance. It is not possible to therefore put a confidence interval on true score except the trivial ultra conservative one. With double blind marking it

*Department of Economics, London School of Economics, Houghton Street, London WC2A 2AE, United Kingdom. E-mail address: lintono@lse.ac.uk. Research was supported by the Economic and Social Science Research Council of the United Kingdom. Thanks to Andrew Chesher for pointing out some references new and old, and to Alan Manning and Danny Quah for helpful comments.

turns out that one can estimate consistently the relative contributions to the overall variance and put a confidence interval on true score. This is an important merit of double blind marking from a statistical point of view.

A second question we address in the context of double blind marking, is: what is the best estimate of μ_i for each candidate given X_{i1}, X_{i2} ?

Before embarking on the solution to this well-defined statistical problem we should comment on the context and related work. A recent paper by Brooks (2004) discusses the issue of double marking in the context of A-level marking and university marking. In that paper and its many references there is a large body of evidence on the actual outcome of exams. One issue is the shape of the distribution of the marker's marks, with many shapes being found from the gendarmes hat to the more exotic trimodals and bimodals. Another issue arises in double marking is the correlation between markers with varying opinions about what is 'right' and acceptable - a benchmark figure being 0.70 correlation. Many UK universities have produced policy statements in the public domain about their marking policy. These are typically lengthy legalistic documents and are more about process and public perception than best practice. Double marking is a very labour intensive activity and it seems reasonable that given this system is in place one should make the best possible use of the resulting data.¹ The purpose of this note is just to point out a statistical approach to doing that.

2 The Proposal

The statistical problem is well known in panel data and the special case where $\sigma_1^2 = \sigma_2^2$ was treated by Neyman and Scott (1948). The problem is that one cannot obtain consistent estimates (as $n \rightarrow \infty$) of μ_i although one can get consistent estimates of σ_j^2 as we show below. These can be used to obtain consistent confidence intervals for the true mark. Furthermore, even though there are not consistent estimates of μ_i there are still optimal ones according to standard criteria.

Suppose that σ_j^2 was known, then the Best Unbiased Estimator of μ_i is the GLS estimator

$$\tilde{\mu}_i = \frac{\sum_{j=1}^2 X_{ij} \sigma_j^{-2}}{\sum_{j=1}^2 \sigma_j^{-2}}, \quad (2)$$

which has variance

$$\text{var}(\tilde{\mu}_i) = \frac{1}{\sum_{j=1}^2 \sigma_j^{-2}}. \quad (3)$$

In particular, it will be more efficient than the simple average $\bar{\mu}_i = (X_{i1} + X_{i2})/2$, in the sense that $\text{var}(\tilde{\mu}_i) \leq \text{var}(\bar{\mu}_i) = (\sigma_1^2 + \sigma_2^2)/4$ with equality only when $\sigma_1^2 = \sigma_2^2$.² Although we do not know σ_j^2 it turns out we can estimate them consistently.

One is tempted to take as an estimate of σ_j^2 the sample variance of marker j 's scores, but this is not correct because it also includes the variance of the true under-

¹White (2000) provides an interesting critique of double marking and discusses alternatives. Our proposal leaves aside all questions of incentives for markers.

²The relative efficiency is $(1+s)^2/4s$, where $s = \sigma_1^2/\sigma_2^2$, which is unbounded at $s = 0$ and $s = \infty$ with a single global minimum at $s = 1$.

lying scores. One may also get the impression that σ_j^2 are pretty similar by looking at the sample variances, so that the sample average estimator is efficient, but this is misleading for the same reason - if the true variance is a large part of this, then the true ratio of σ_2^2/σ_1^2 could be quite different from the ratio of the sample variances.

In the sequel we treat the latent scores μ_i as random variables with certain population moments existing. This is just for notational convenience [one could also treat them as deterministic quantities and work with sums instead of moments]. We note that

$$\text{cov}(X_{i1}, X_{i2}) = \sigma_\mu^2 \quad (4)$$

$$\text{var}(X_{i1}) = \sigma_\mu^2 + \sigma_1^2 \quad (5)$$

$$\text{var}(X_{i2}) = \sigma_\mu^2 + \sigma_2^2, \quad (6)$$

where $\sigma_\mu^2 = \text{var}(\mu_i)$. It follows that

$$\sigma_1^2 = \text{var}(X_{i1}) - \text{cov}(X_{i1}, X_{i2})$$

$$\sigma_2^2 = \text{var}(X_{i2}) - \text{cov}(X_{i1}, X_{i2}),$$

which suggests the sample estimators

$$\hat{\sigma}_\mu^2 = \frac{1}{n} \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)$$

$$\hat{\sigma}_1^2 = \frac{1}{n} \sum_{i=1}^n (X_{i1} - \bar{X}_1)^2 - \frac{1}{n} \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2)$$

$$\hat{\sigma}_2^2 = \frac{1}{n} \sum_{i=1}^n (X_{i2} - \bar{X}_2)^2 - \frac{1}{n} \sum_{i=1}^n (X_{i1} - \bar{X}_1)(X_{i2} - \bar{X}_2),$$

where $\bar{X}_j = n^{-1} \sum_{i=1}^n X_{ij}$, $j = 1, 2$. It can be shown that

$$\hat{\sigma}_j^2 \xrightarrow{P} \sigma_j^2$$

as $n \rightarrow \infty$. Our proposal for estimating μ_i is then

$$\hat{\mu}_i = \frac{\sum_{j=1}^2 X_{ij} \hat{\sigma}_j^{-2}}{\sum_{j=1}^2 \hat{\sigma}_j^{-2}}, \quad (7)$$

which can be computed from any spreadsheet. It can be shown that

$$\hat{\mu}_i - \tilde{\mu}_i \xrightarrow{P} 0$$

as $n \rightarrow \infty$, so that the estimation error in $\hat{\sigma}_j^2$ is of smaller magnitude than the error in $\tilde{\mu}_i$. For large n , the estimator $\hat{\mu}_i$ should be approximately efficient.

In addition, one has consistent estimates of the variance components $\hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{\sigma}_\mu^2$, which can be used to identify erratic markers and to form confidence intervals. Define the interval

$$\hat{I}_i = \hat{\mu}_i \pm z_{\alpha/2} \frac{1}{\sqrt{\sum_{j=1}^2 \hat{\sigma}_j^{-2}}},$$

where $z_{\alpha/2}$ is the normal critical value. It follows that

$$\Pr[\mu_i \in \widehat{I}_i] \rightarrow 1 - \alpha \quad (8)$$

as $n \rightarrow \infty$. This interval gives some idea of the intrinsic accuracy of the outcome.

One can also compute the missclassification probability based on this, that is, what is the probability that on the true score you would get grade k when on the observed score you received x marks.

Note that the correlation between marker 1 and 2 is

$$\rho_{12} = \frac{\sigma_\mu^2}{\sqrt{(\sigma_\mu^2 + \sigma_1^2)(\sigma_\mu^2 + \sigma_2^2)}}$$

and this is equal to $\sigma_\mu^2/(\sigma_\mu^2 + \sigma^2)$ in the special case that $\sigma_1^2 = \sigma_2^2 = \sigma^2$. A correlation of 0.8 corresponds to a signal noise ratio (σ_μ^2/σ^2) of 4 i.e., the ratio of standard errors is 2 in that case.

3 Some Extensions

The normality assumption is not crucial: the GLS is (asymptotically) BLUE although not (asymptotically) BUE when ε_{ij} are not normally distributed, and our estimator is still more efficient than the midpoint except in the homogenous case when they are equally efficient. Furthermore, the estimation of σ_j^2 does not require the distributional assumptions at all.

In practice we find not just differences in variances but also differences in means. Therefore, consider instead the model

$$X_{ij} = \mu_i + \alpha_j + \varepsilon_{ij} \quad (9)$$

with the same assumptions regarding μ_i, ε_{ij} , but now there are systematic differences in the level of marks assigned by the markers. In this case there is an identification issue - one can't allow all μ_i and all α_j to be unspecified. The natural normalization is to assume that $\alpha_1 + \alpha_2 = 0$. Letting $\alpha = \alpha_1$ for simplicity we have the model

$$X_{ij} = \mu_i + \alpha d_{ij} + \varepsilon_{ij},$$

where $d_{i1} = 1$ and $d_{i2} = -1$ for all i . This model is a standard panel data regression model with fixed effects μ_i . Suppose that both α and σ_j^2 are known, then the optimal estimator of μ_i is

$$\tilde{\mu}_i = \frac{\sum_{j=1}^2 (X_{ij} - \alpha d_{ij}) \sigma_j^{-2}}{\sum_{j=1}^2 \sigma_j^{-2}}$$

as before. To estimate α one can use the standard fixed effects estimator [Hsiao (2003, p32)], which in this case is

$$\hat{\alpha} = \frac{1}{2n} \sum_{i=1}^n (X_{i1} - X_{i2}).$$

To estimate the variances we use

$$\begin{aligned}\hat{\sigma}_\mu^2 &= \frac{1}{n} \sum_{i=1}^n (X_{i1}^* - \bar{X}_{*1}^*) (X_{i2}^* - \bar{X}_{*2}^*) \\ \hat{\sigma}_1^2 &= \frac{1}{n} \sum_{i=1}^n (X_{i1}^* - \bar{X}_{*1}^*)^2 - \frac{1}{n} \sum_{i=1}^n (X_{i1}^* - \bar{X}_{*1}^*) (X_{i2}^* - \bar{X}_{*2}^*) \\ \hat{\sigma}_2^2 &= \frac{1}{n} \sum_{i=1}^n (X_{i2}^* - \bar{X}_{*2}^*)^2 - \frac{1}{n} \sum_{i=1}^n (X_{i1}^* - \bar{X}_{*1}^*) (X_{i2}^* - \bar{X}_{*2}^*),\end{aligned}$$

where $X_{ij}^* = X_{ij} - \hat{\alpha}d_{ij}$ and $\bar{X}_{*j}^* = n^{-1} \sum_{i=1}^n X_{ij}^*$, $j = 1, 2$. It can be shown that $\hat{\sigma}_j^2 \xrightarrow{P} \sigma_j^2$ and $\hat{\alpha} \xrightarrow{P} \alpha$ as $n \rightarrow \infty$. We compute

$$\hat{\mu}_i = \frac{\sum_{j=1}^2 (X_{ij} - \hat{\alpha}d_{ij}) \hat{\sigma}_j^{-2}}{\sum_{j=1}^2 \hat{\sigma}_j^{-2}},$$

and the confidence interval \hat{I}_i as before.

4 An Example

We analyze the results of an exam taken at the LSE in 2003. A subset of size $n = 50$ of that exam was double blind marked by examiners 1 and examiners 2. The sample moments were $\bar{X}_1 = 51.38$ and $\bar{X}_2 = 59.46$ with sample variances respectively 155.02 and 159.62 and covariance 136.90. It follows that

$$\hat{\sigma}_1^2 = 18.12 \quad ; \quad \hat{\sigma}_2^2 = 22.72.$$

In this case the OLS and GLS estimators are similar [$\hat{\sigma}_1^2/(\hat{\sigma}_1^2 + \hat{\sigma}_2^2) \simeq 0.44$ compared with the OLS weights of 0.50] and the relative inefficiency of OLS is small. Note however that for the OLS estimator the 95% confidence interval around true mark is ± 6.26 . This should be cause for concern because it says that of every 20 students who fail an exam with mark 44 against pass mark of 50, one of them has true mark of 50 or more, i.e., passing.

References

- [1] Brooks, V., (2004): "Double Marking Revisited," British Journal of Educational Studies 52, 29-46.
- [2] Edgeworth, F.Y. (1888): "The statistics of Examinations," Journal of the Royal Statistical Society 51, 599-635.
- [3] Hsiao, C., (2003): "Analysis of Panel Data," Econometric Society Monograph, 2nd edition, Cambridge, UK.

- [4] Neyman, J., and E.L. Scott (1948): “Consistent Estimates based on partially consistent observations,” *Econometrica* 16, 1-32.
- [5] White, R., (2000): “Double marking versus monitoring of examination,” Discussion article, <http://www.prs-ltsn.leeds.ac.uk/philosophy/discussions/white.html>