Estimation of Misclassification Rates for Human and AI Reading Accuracy Measurements

Ngoc Vu Advisor: Dr. Nelis Potgieter
Department of Mathematics, Texas Christian University, Fort Worth, Texas

Introduction

Oral Reading Accuracy is an important measure of reading proficiency. Traditionally scored by human observers, recently AI systems integrating voice recognition software have also been utilized, see Nese et al (2015). This study develops two statistical models to evaluate the efficacy of technology in ORA assessment. The goal is to estimate the true positive and true negative rates associated with different scoring methods.

Definitions

Each observed WRC score can be characterized as the sum of a true positive and a true negative score. Here, we define:

→ **True Positive Rate (π_{tp})**: The proportion of correctly read words that are accurately classified as correctly read by the scoring system.

→ **True Negative Rate ($1 - \pi_{tn}$)**: The proportion of incorrectly read words that are mistakenly classified as correctly read by the scoring system.

We let $\boldsymbol{X}$ represent the **True Count Variable** – a gold-standard measurement of the number of correctly read words in a passage consisting of N words. The distribution of X is assumed unknown with mean and variance μ_x and σ_x^2.

To define the **Observed Count Variable**, we define **True Positive Score** Component

$$\boxed{[\mathrm{X}_1|X] \sim \mathrm{Bin}(X, \pi_{tp})}$$

and **True Negative Score** component

$$\boxed{[\mathrm{X}_2|X] \sim \mathrm{Bin}(N - X, 1 - \pi_{tn}).}$$

The **Error-prone Count** is then given by

$$\boxed{\mathrm{Y}{=}\mathrm{X}_1 + X_2.}$$

Two such error-prone counts can be realized. Let $\boldsymbol{Y_1}$ denote the **Score** observed by a **Human** assessor and let $\boldsymbol{Y_2}$ denote the **Score** recorded by the **AI** voice recognition software.

Using the properties of the binomial distribution, as well as the properties of conditional expectations, we can quantify the relationships between gold-standard counts and error-prone counts.

Data Description

The ORA data was collected from a sample of 507 elementary school students. Each student was assigned one of ten study passages of different length and difficulty. Both human and AI evaluations recorded the WRC score. Figure 1 and Table 1 below illustrate the observed scores (Y_1 and Y_2) against the gold-standard score X for each student.

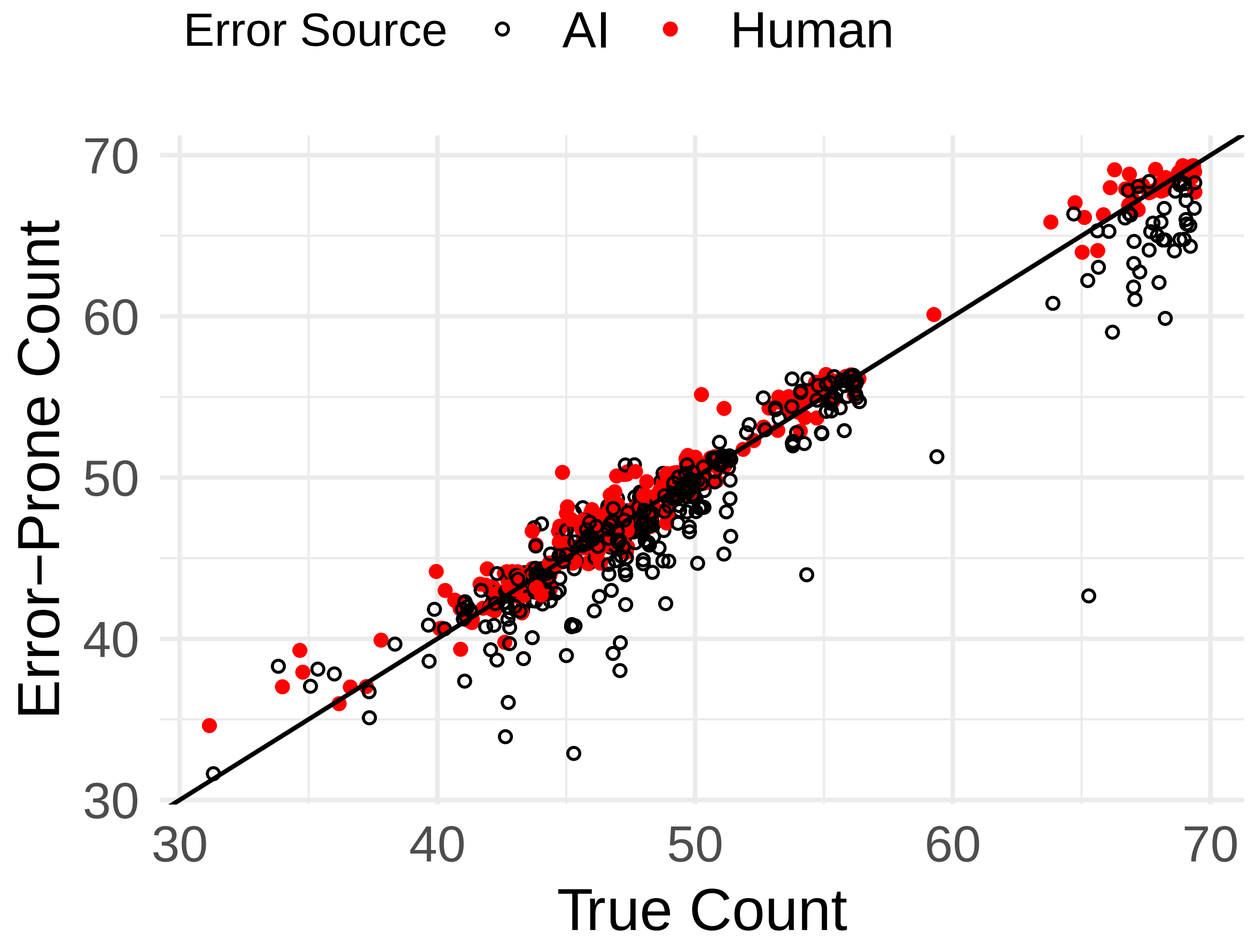

Figure 1: Comparison of True and Error-Prone Counts in ORA Assessment: Human vs. AI.

In Figure 1, human scores (red) cluster more closely to the reference line, indicating greater rating consistency. Furthermore, AI scores (black) are much more likely to fall below the reference line, indicating bias introduced by automatic scoring.

	Human	AI
Correct Score	.643	.422
1 Point Off	.264	.324
>1 Point Off	.093	.254

Table 1: Accuracy Proportions: True vs Error-Prone WRC

Although human evaluators show evidence of greater consistency, this rating methods require more time and effort. Therefore, we are interested in understanding and addressing the errors of AI systems for further improvements in WRC measurement.

Model Illustration

Here are two scenarios illustrating the impact of incorrect counts. With $\boldsymbol{(\pi_{tp}, \pi_{tn}) = (0.95, 0.65)}$, the distribution shifts leftward, yielding a lower mean and inflated standard deviation. Parameters $\boldsymbol{(\pi_{tp}, \pi_{tn}) = (0.99, 0.05)}$ result in a rightward shift, producing a higher mean and reduced standard deviation.

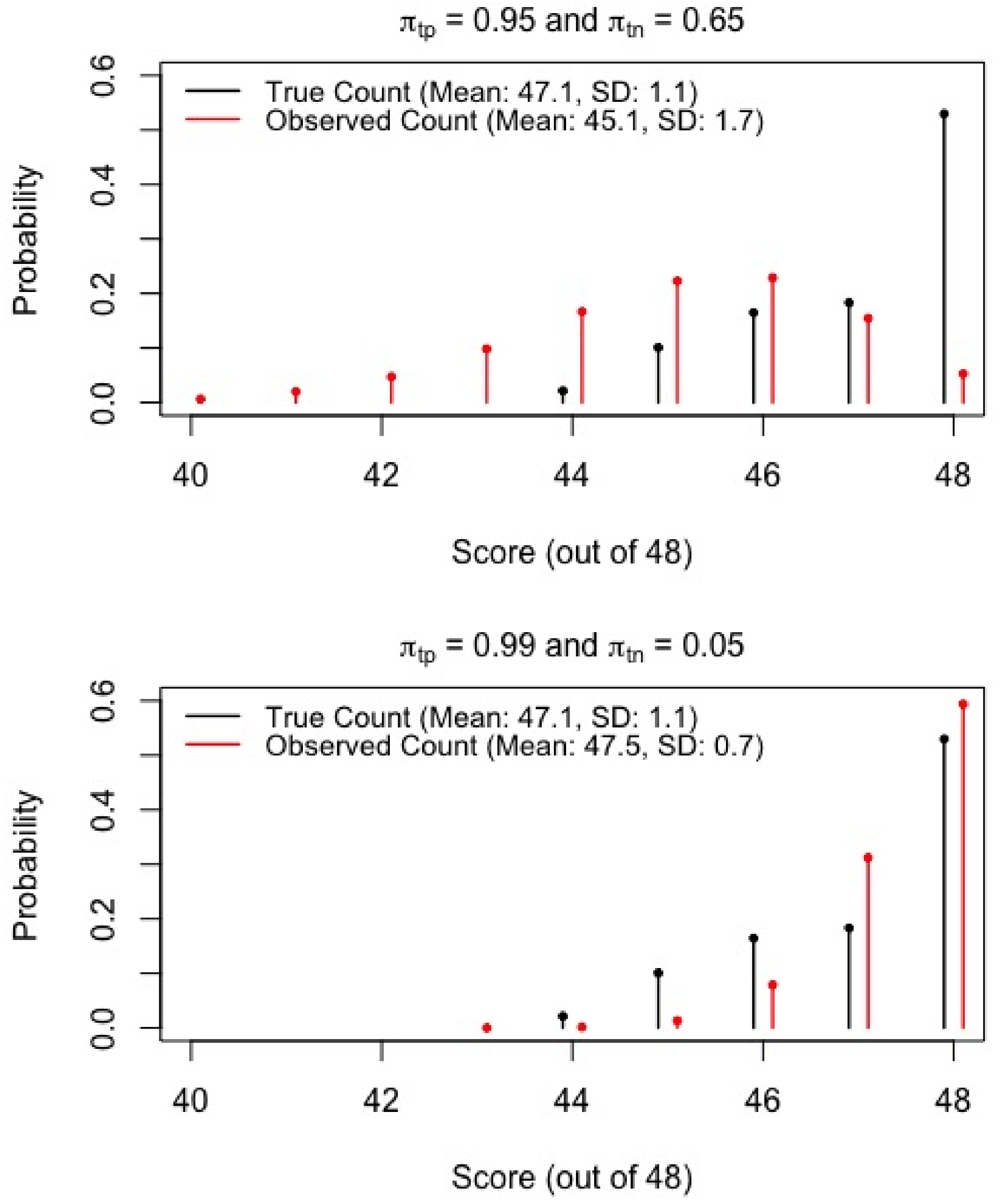

Figure 2: Illustrating the Error Effect

Complete Data Solution

Our solution assumes access to gold-standard measurements X alongside error-prone data Y. From binomial distribution properties, we have moment relationships

$$\boxed{\mu_y = E[Y] = \mu_x \cdot \pi_{tp} + (N - \mu_x) \cdot (1 - \pi_{tn})}$$

and

$$\boxed{\sigma_{xy}^2 = Cov[X, Y] = \sigma_x^2(\pi_{tp} + \pi_{tn} - 1).}$$

Using methods of moment estimators, replacing $(\mu_x, \mu_y, \sigma_x^2, \sigma_{xy})$ by $(\bar{x}, \bar{y}, s_x^2, s_{xy})$, we have estimated accuracy rates

$$\boxed{\begin{aligned} \widehat{\pi}_{tn} &= 1 - \tfrac{\bar{y}}{N} + \tfrac{\bar{x}}{N} \cdot \tfrac{s_{xy}}{s_x^2}, \\ \widehat{\pi}_{tp} &= \tfrac{\bar{y}}{N} + \tfrac{s_{xy}}{s_x^2} \cdot \left(1 - \tfrac{\bar{x}}{N}\right). \end{aligned}}$$

Data Application

The respective values of π_{tn} and π_{tp} were estimated from our data. The typical sample size is around 50 measurements per passage; estimates are summarized in Table 2. Due to the restricted range of accuracy rates, we report truncated values, $\tilde{\pi}_{tp} = \min(\hat{\pi}_{tp}, 1)$ and $\tilde{\pi}_{tn} = \min(\hat{\pi}_{tn}, 1)$.

	True Positive		True Negative	
Passage	Human	AI	Human	AI
1	0.9990	0.9951	0.8328	1.0000
2	0.9975	0.9939	0.6648	0.9383
3	0.9975	0.9619	0.7579	1.0000
4	0.9972	0.9922	0.6997	1.0000
5	1.0000	0.9886	0.7603	0.8574
6	0.9997	0.9859	0.7212	0.6529
7	0.9975	0.9822	0.7164	0.6582
8	1.0000	0.9753	0.7265	0.6734
9	1.0000	0.9854	0.7189	1.0000
10	1.0000	0.9782	0.8727	0.9585
Summary Measure	0.9989	0.9839	0.7471	0.8739

Table 2: Comparison of Human and AI Estimations

Conclusion & Future Works

Both human recorders and the AI systems demonstrate overall strong performance. Human recorders show higher consistency in True Positive values across passages while AI systems outperform human recorders in True Negative rates.

In practice, gold-standard ORA measurements are unavailable. To this end, our next step will be to explore solutions assuming a statistical distribution for X. Then, defining $\hat{m} = (\bar{y}_1, \bar{y}_2, s_1^2, s_2^2, s_{12})^\top$,

$$\boldsymbol{\theta} = (\mu_x, \pi_{tp1}, \pi_{tp2}, \pi_{tn1}, \pi_{tn2})^\top, \text{ and}$$

$$m(\boldsymbol{\theta}) = (\mu_1(\boldsymbol{\theta}), \mu_2(\boldsymbol{\theta}), \sigma_1(\boldsymbol{\theta})^2, \sigma_2(\boldsymbol{\theta})^2, \sigma_{12}(\boldsymbol{\theta}))^\top,$$

we will estimate accurate rates by minimizing $D(\boldsymbol{\theta}) = (\hat{m} - m(\boldsymbol{\theta}))^\top \boldsymbol{\Sigma}^{-1}(\hat{m} - m(\boldsymbol{\theta}))$.

References

Nese, J. F. T., Kamata, A., & Alonzo, J. (2015). Exploring the evidence of speech recognition and shorter passage length in computerized oral reading fluency. *Paper presented at the Society for the Scientific Study of Reading meeting*, Kona, HI.