Rethinking Visualization of Known-Groups Validity

Nikola Markiewicz[1], **Piper Fromy**[1]

[1] SeeingTheta

Introduction

The default plot

Known-groups validity is typically visualised as a boxplot with jittered points.

That default works best with a "normal" sample: a few hundred patients, symmetric scores and a clean anchor.

Methods

What we simulated

Measure. Graded Response Model, 10 items, total score 0–40.

Groups. PGIS at baseline: Mild, Moderate, Severe.

Sample sizes. Small: n = 20, normal: n = 150, large: N = 2000.

Stress test. Floor, ceiling, zero inflation, a deliberately weak anchor.

Compared. Boxplot vs eCDF, bootstrap CDF, ridge, raincloud, waterfall.

Results

Only the eCDF held up

The boxplot only worked at n = 150: unstable at 20, overplotted at 2000.

The eCDF was interpretable at every size.

All four data pathologies broke the boxplot. The weak anchor was worst: tidy boxes, no real separation (fig 2)

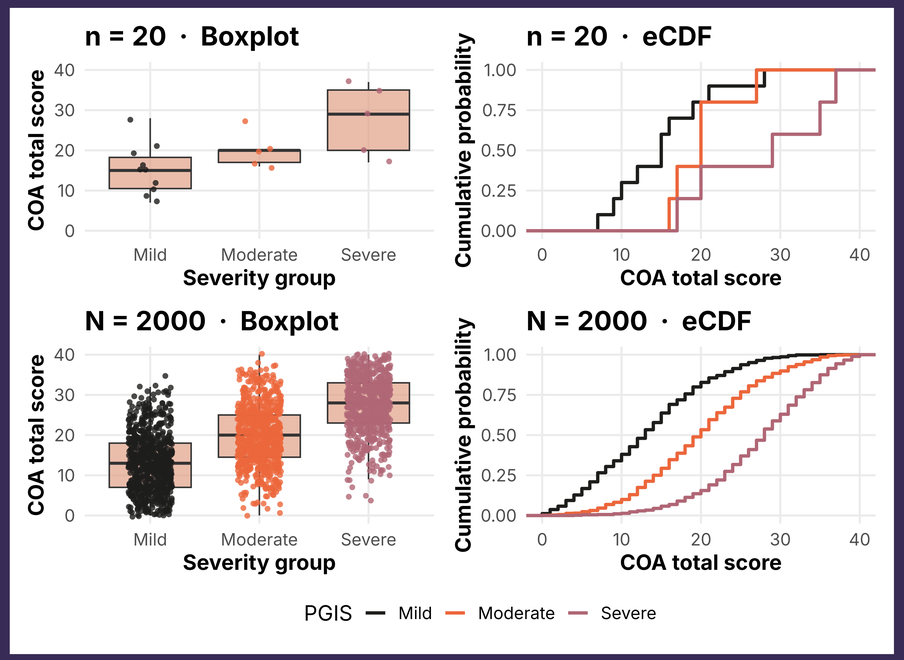
Discussion

What to do instead

Default to the eCDF instead of using the boxplots, especially when working with "non-normal" samples. Add a bootstrap CDF at small n so uncertainty is visible rather than assumed.

In very small samples and very large samples, boxplots hide whether the groups differ.

The eCDF curve shows it.



Want to help grow this into a full guide? Scan the QR code to get involved.

SeeingTheta

Discrimination using AUC is useful at large sample sizes

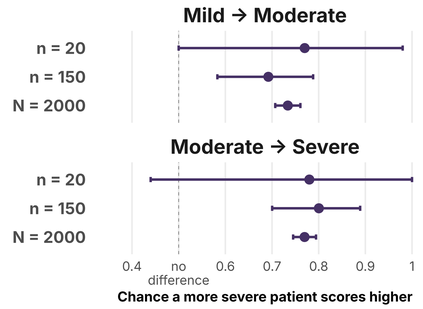


Figure 1: How reliably a more severe patient outscores a milder one, with 95% confidence intervals. The estimate holds; only the interval widens.

At small samples, the confidence interval crosses the chance threshold, but for adequately powered samples this display is useful.

Four data pathologies, four boxplot failures

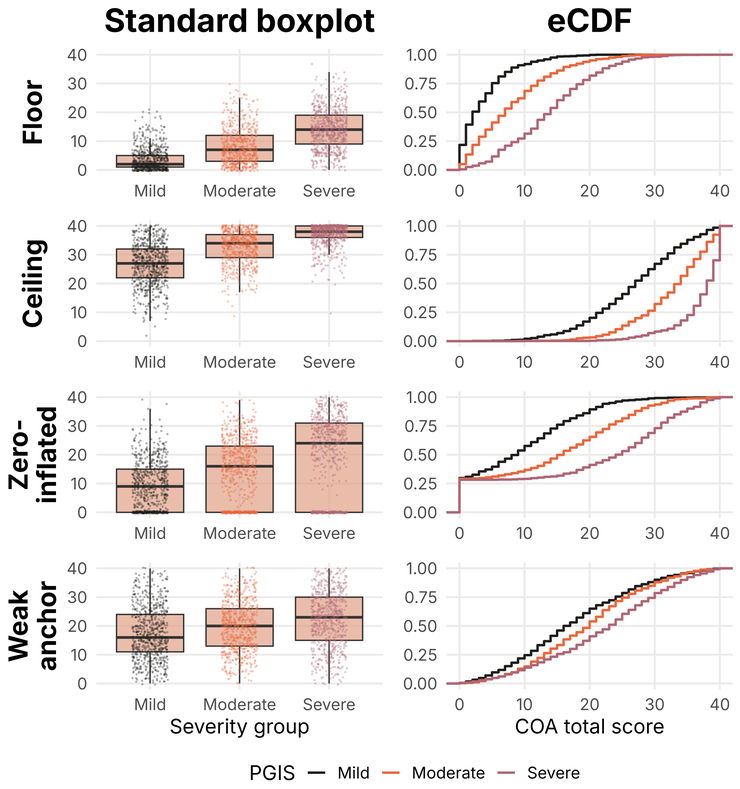


Figure 2: Floor, ceiling, zero inflation and a weak anchor at N = 2000, drawn both ways. Read down each column.

Contact
Nikola Markiewicz
nikola@seeingtheta.com
seeingtheta.com

References
Figures built in R with ggplot2 [1] and ggridges [2].
1. Wickham H, Chang W, Henry L, et al. ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics. R package version 4.0.3. 2026.
2. Wilke CO. ggridges: Ridgeline Plots in ggplot2. R package version 0.5.7. 2025.

