POSTER#220

Enhancing Non-Experts' Understanding of Uncertainty in Psychometric Results Through Effective Visual Probabilistic Representations

Authors: Ashley Willis[1], Adrian Jewett[1], Brandon Foster[1]

Affiliations: 1. Lumanity – Patient-Centered Outcomes, Boston, MA, USA

INTRODUCTION

- Psychometric analysis is crucial in drug development for obtaining reliable and valid clinical outcomes assessments (COAs) to evaluate drug effects, understand patient experiences, aid diagnosis, and support clinical decision-making
- For clinical decision-makers who are not psychometricians (hereafter referred to as "non-experts"), evaluating whether a COA is valid and reliable for an intended use can be challenging. As a result, "rules of thumb" are often used to interpret psychometric results
- Measures that appear psychometrically acceptable based on these "rules of thumb" may in fact have a higher risk of poor performance than is currently understood. Such heuristics, while convenient, may lead to overestimation of a COA's true psychometric strengths for a given use – resulting in underpowered trials, misleading results, or more generally exacerbating the replication problem in science[1]
- Frequency framing is a data visualization technique that converts continuous probability distributions for an outcome to discrete outcomes proportionate to their likelihood. These visualization techniques can make probabilities more tangible and understandable to general audience
- Quantile dot plots are one such visualization.[2] In these plots, a dot represents an equal increment of probability, discretizing an underlying probability distribution for some value of interest
- This visualization technique has been shown in other contexts to enhance memory of distributional information and can promote consistent decision-making in the presence of risk[2]

OBJECTIVES

Our aims were to:

1. Conduct a psychometric evaluation of simulated COA data using Bayesian methods to generate posterior distributions for statistics of interest
2. Introduce quantile dot plots as a frequency framing tool to summarize and interpret the Bayesian posteriors for the statistics of interest
3. Demonstrate how quantile dot plots enhance understanding of uncertainty and enable non-psychometric experts to ask questions about the data that they otherwise would not have been able to

METHODS

Data

- A COA was simulated with 20 items. The data-generating model sampled factor loadings for 21 items. The first 20 items represented the items for the COA and sampled factor loadings from a uniform distribution with loadings ranging from 0.30 to 0.90. The final factor loading for the 21st item was set to 0.95 to generate a Patient Global Impression of Severity (PGIS) measure. Item thresholds for response option endorsement were set to ensure evenness
- Items were simulated for 500 subjects using a graded response model with the R package *mirt*.[3,4] The Total Score for the measure was computed by summing the responses from the 20 items
- Five continuous co-validators with scores ranging from 0 to 100 were simulated in relation to the Total Score, with correlation magnitudes of 0.08 (Discriminant), 0.18 (Small), 0.25 (Small/Medium), 0.35 (Medium), and 0.50 (Large)

Analyses

A suite of basic psychometric analyses was pursued to evaluate different aspects of the reliability and validity of the COA. These included:

- Structural validity through model fit of a unit-weighted unidimensional confirmatory factor analysis (CFA) model (Figure 1)
- Internal consistency with Coefficient Omega (Figure 2)
- Convergent validity with Pearson correlational analyses and five co-validating measures (Figure 3)
- Known-groups validity using regression modelling to evaluate sequential mean differences in COA scores across PGIS strata for severity (Figure 4)

For each set of results above, Bayesian modelling was carried out using the R package *blavaan* using non-informative priors.[4,5] Posterior probability distributions were generated for all results. The posterior probability distributions were then converted to quantile dot plots using the *ggdist* package, and plots were generated using *ggplot2*.[4,6,7] Dots within the plot were color coded for each analysis according to oft-cited "rules of thumb" for interpretation of results, using the following criteria:

- **Structural validity:** Bayesian comparative fit index (BCFI) and Bayesian Tucker–Lewis index (BTLI) ≥ 0.95[8]
- **Internal consistency:** Internal consistency was characterized as poor (< 0.50), moderate (0.50–0.75), good (0.75–0.90), and acceptable for clinical measures (> 0.90)[9]
- **Convergent validity:** Pearson correlation coefficients were evaluated by performance criteria as discriminant (< 0.10), small (0.11–0.30), medium (0.31–0.50), and acceptable (> 0.50)[10]
- **Known-groups validity:** Estimated mean Total Score differences were characterized for direction and magnitude of change in the expected direction[11]

Finally, results generated from a conventional psychometric workup utilizing frequentist methods were generated for each analysis to compare the results from each method. These results are represented by the mode for each plot.

RESULTS

Structural validity (Figure 1)

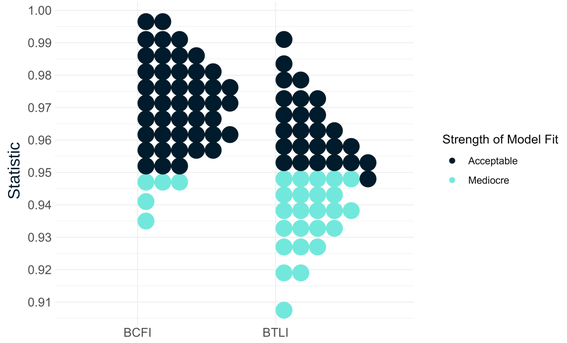
Conventional analysis results suggest acceptable structural validity (BCFI and BTLI ≥ 0.95)

- The probability that the BCFI and BTLI exceed the acceptable success threshold of ≥ 0.95 are 90% and 52%, respectively
- The BTLI has a relatively high likelihood (specifically 48%) of not meeting acceptable performance criteria. Observing this outcome might prompt a non-expert to reassess the model's fit or prompt them to seek input from psychometricians to understand the reasons why the BTLI might be lower

Each plot contains 50 dots. Each dot represents 2% probability. To determine the probability of each outcome, simply count the dots for each color and multiply by 2 to achieve a percentage.

Results from a typical workup utilizing frequentist methods are represented by the mode for each plot (highest number of dots).

Figure 1: Posterior draws of model fit statistics



Key: BCFI, Bayesian comparative fit index; BTLI, Bayesian Tucker–Lewis index.

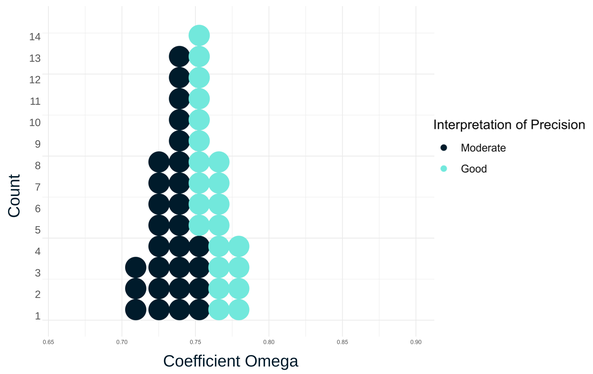
Note: The quantile dot plot is based on 50 quantiles. Each dot represents 2% of the data distribution. Interpretive criteria adapted from Hu & Bentler (1988)[8]

Internal consistency (Figure 2)

Conventional analysis results suggest good reliability (~0.75)

- Moderate reliability (0.50 – 0.75): 56% probability
- Good reliability (0.76 – 0.90): 44% probability
- The mode suggests good reliability. The variability around the conventional performance criteria is minimal (range is roughly 0.70 – 0.78) leading to more certainty about internal consistency of this COA

Figure 2: Posterior draws of reliability estimates



Note: The quantile dot plot is based on 50 quantiles. Each dot represents 2% of the data distribution. Interpretive criteria adapted from Portney & Watkins (2009), with values less than 0.50 as poor, 0.50 to 0.75 as moderate, 0.75 to 0.90 as good, and above 0.90 as acceptable for clinical measures[9]

Convergent validity (Figure 3)

Co-validator 1:

- *Conventional analysis results suggest a small magnitude correlation*
- Probabilities: Discriminant: 38%; Small: 60%; Medium: 2%
- If it was hypothesized that this co-validator was closely aligned to the construct measured by COA, the probability of the correlation being potentially discriminant might be unacceptable

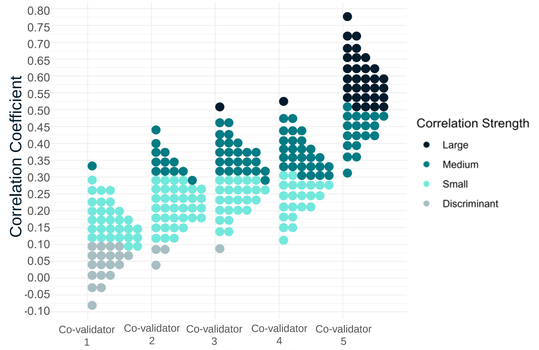
Co-validator 3:

- *Conventional analysis results suggest a medium magnitude correlation*
- Probabilities: Discriminant: 2%; Small: 48%; Medium: 54%; Large: 2%
- Unlikely to be discriminant, but still not a large difference between likelihood of small versus medium correlation

Co-validator 5:

- *Conventional analysis results suggest a large magnitude correlation*
- Probabilities: Medium: 38%; Large: 62%
- Can more confidently report a large correlation

Figure 3: Posterior draws of correlation coefficients



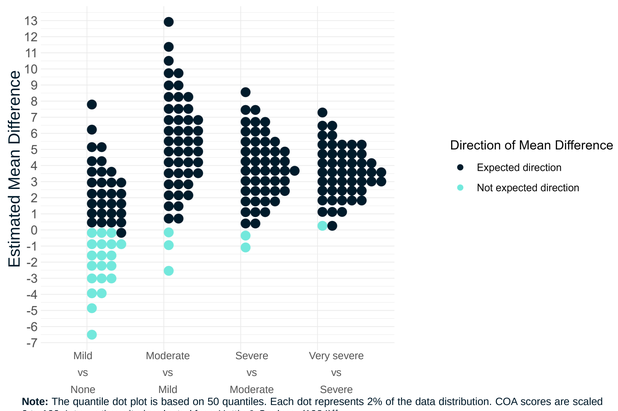
Note: The quantile dot plot is based on 50 quantiles. Each dot represents 2% of the data distribution. Interpretive criteria adapted from Cohen (1988)[10]

Known-groups validity (Figure 4)

Conventional analysis results suggest positive monotonic ordering of estimated mean differences across PGIS strata

- Positive mean differences between adjacent strata of the PGIS order in the expected direction
- However, 40% of the mean score differences between the *Mild* and *None* response categories were negative. This COA likely cannot differentiate between these two severity strata

Figure 4: Posterior draws of estimated mean differences



Note: The quantile dot plot is based on 50 quantiles. Each dot represents 2% of the data distribution. COA scores are scaled 0 to 100. Interpretive criteria adapted from Hattie & Cooksey (1984)[11]

CONCLUSIONS

- Understanding the psychometric performance of COAs is critical for drawing correct inferences. However, understanding the psychometric performance of COAs is a complex task that can be challenging for clinical decision makers who do not have psychometric training. In this regard, quantile dot plots offer a powerful tool for frequency framing the uncertainty or relative probability of results
- We demonstrated that Bayesian methods can be utilized when evaluating the psychometric properties of COAs to generate probabilistic statements about psychometric properties. Further, presenting these statements via intuitive visualization methods, such as quantile dot plots, allows non-experts to further enhance their understanding of COA psychometric properties and make more accurate and reliable assessments
- Overall, implementing quantile dot plots and utilizing Bayesian statistical methods together offer a more comprehensive understanding of psychometric results than frequentist methods alone. This improved understanding may empower non-experts to make more informed decisions regarding COA selection and use in both clinical trials and practice

REFERENCES

1. Loken, E., & Gelman, A. *Science (1979)* 2017; 355: 584–585. **2**. Kay M, Kola T, Hullman JR, et al. In: *Conference on Human Factors in Computing Systems - Proceedings*. Association for Computing Machinery, 2016, pp. 5092–5103. **3**. Chalmers RP. *J Stat Softw* 2012; 48: 1–29. **4**. R Core Team. 2020; https://www.R--project.org. **5**. Merkle, E., & Rosseel Y. *J Stat Softw*; 85. Epub ahead of print 2018. DOI: 10.18637/jss.v085.i04. **6**. Kay, M., Wiernik, B. *American Statistician* 1999; 53: 276–281. **7**. Wickham, H., Springer-Verlag New York, https://ggplot2.tidyverse.org (2016). **8**. Hu, L., Bentler, P. *Struct Equ Modeling* 1999; 6: 1–55. McDonald RP. 1st ed. Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 1999. **9**. Portney, LG., Watkins, MP., others. Philadelphia: F. A. Davis Company, 2015. **10**. Cohen J. *Curr Dir Psychol Sci* 1992; 1: 98–101. **11**. Hattie J, Cooksey RW. *Appl Psychol Meas* 1984; 8: 295–305.

An electronic version of the poster can be viewed by scanning the QR code.