Using Item Scores and Response Times in Person-Fit Assessment

Kylie Gorney[1], Xiang Liu[2], and Sandip Sinharay[2]

[1]University of Wisconsin-Madison

[2]Educational Testing Service

2022 Ideas in Testing Research Seminar

Introduction

- Person-fit assessment is used to identify individuals displaying unusual response behavior
- Several person-fit statistics have been developed for item scores, but few have been developed for item RTs and even fewer have been developed for item scores and RTs

Introduction

Table 1. Existing Person-Fit Statistics.

	Data Source		
Approach	Item Scores	Item RTs	Item Scores & RTs
Frequentist	l_s^*	l_t^*	–
Bayesian	p_s	p_t	p_{st}

Method

Hierarchical framework (van der Linden, 2007)

- 2PL model for the item scores
- Lognormal model for the item RTs
- A bivariate normal distribution for the person parameters, ability (θ) and speed (τ)

Method

Purpose

Develop two frequentist methods for assessing person-fit in item scores and RTs.

1. Combining individual person-fit statistics
2. Joint model person-fit statistic

Objective

Compute two individual person-fit statistics (one for the item scores, and one for the item RTs), and then combine them to form a single statistic.

- Item scores: l_s^* (Snijders, 2001)
- Item RTs: l_t^* (Sinharay, 2018)

Method

Combining Individual Person-Fit Statistics

- Problem: l_s^* and l_t^* exist on two different metrics
  - l_s^* has an asymptotic $\mathcal{N}(0,1)$ null distribution
  - l_t^* has a χ^2_{n-1} null distribution
- Transform using the inverse CDF method
  - q_s^* has an asymptotic χ^2_1 null distribution
  - q_t^* has a χ^2_1 null distribution
- Their sum has an asymptotic χ^2_2 null distribution

$$q_{st}^* = q_s^* + q_t^* \tag{1}$$

Objective

Compute a single person-fit statistic using the likelihood function of the joint model for item scores and RTs.

- Standardized log-likelihood statistic (to be used with θ and τ)

$$l_{st} = \frac{l - E[l]}{\sqrt{\text{Var}(l)}} = \frac{W_n}{\sqrt{n}\sigma_n} \quad (2)$$

- Asymptotically correct version (to be used with $\hat{\theta}$ and $\hat{\tau}$)

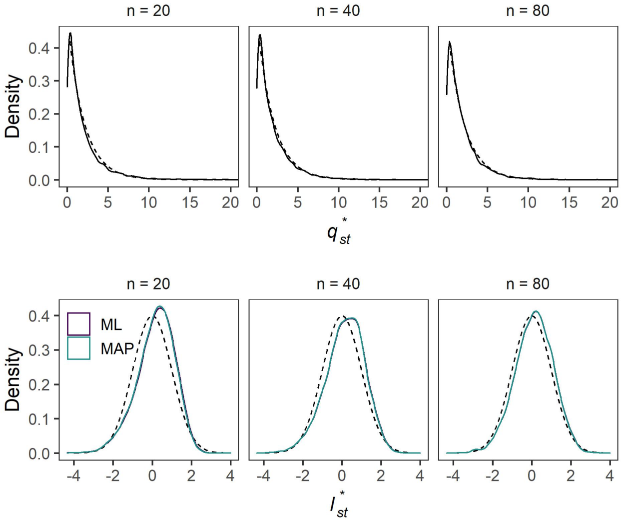
$$l^*_{st} = \frac{W_n + c_n s_0}{\sqrt{n}\tilde{\sigma}_n} \quad (3)$$

Simulation Studies

- Study 1: The Null Distributions of q^*_{st} and l^*_{st}
- Study 2: Performance of the Person-Fit Statistics

Simulation Studies

Study 1: The Null Distributions of q^*_{st} and l^*_{st}



- 1,000 examinees
  - 90% non-aberrant
  - 10% aberrant
- 100 replications
- `LNIRT` package in R

- Test length
  - 20
  - 40
  - 80
- Percentage of contaminated items
  - 10
  - 20
  - 40
- Correlation between θ and τ
  - 0.2
  - 0.5
  - 0.8

- Type I error rates decreased and power increased as...
  - test length increased
  - the percentage of contaminated items increased
- Across all conditions, q^*_{st} and l^*_{st} displayed satisfactory Type I error rates and larger power than the existing person-fit statistics

Table 2. Power (40-Item Test, $\alpha = 0.05$).

	Existing		New	
Aberrance	l_s^*	l_t^*	q_{st}^*	l_{st}^*
Preknowledge	.176	.309	.344	.350
Random responding	.314	.882	.896	.899

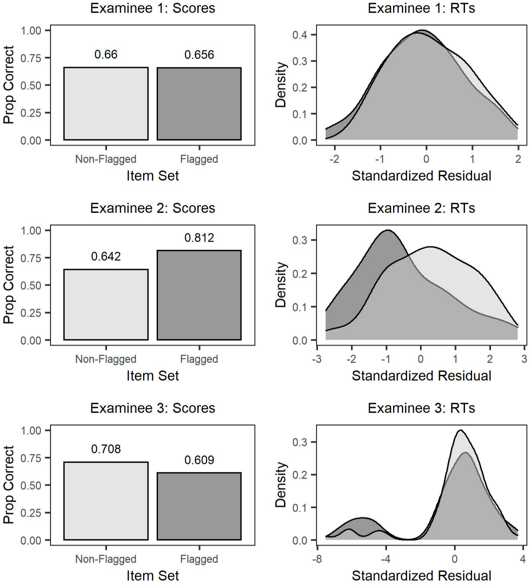
Real Data Example

- Form 1 of the credentialing data set of Cizek and Wollack (2017)
- 1,624 examinees (41 flagged), 170 items (64 flagged)

Table 3. Proportions of Statistically Significant Values ($\alpha = .05$).

Examinee Group	q^*_{st}	l^*_{st}
Non-Flagged	.196	.184
Flagged	.317	.268

Real Data Example



Conclusion

- We developed two frequentist person-fit statistics for item scores and RTs
- Appear to be promising tools for detecting aberrant behavior
- Future directions
  - Additional simulation conditions and real data sets
  - Investigate differences between q^*_{st} and l^*_{st}
  - Extensions that utilize additional process data

References

- Cizek, G. J., & Wollack, J. A. (Eds.). (2017). *Handbook of quantitative methods for detecting cheating on tests*. Routledge. https://doi.org/10.4324/9781315743097
- Sinharay, S. (2018). A new person-fit statistic for the lognormal model for response times. *Journal of Educational Measurement*, *55*(4), 457–476. https://doi.org/10.1111/jedm.12188
- Snijders, T. A. B. (2001). Asymptotic null distribution of person fit statistics with estimated person parameter. *Psychometrika*, *66*(3), 331–342. https://doi.org/10.1007/BF02294437
- van der Linden, W. J. (2007). A hierarchical framework for modeling speed and accuracy on test items. *Psychometrika*, *72*(3), 287–308. https://doi.org/10.1007/s11336-006-1478-z