

Rasch Modeling of Revised Token Test Performance: Validity and Sensitivity to Change

William D. Hula^{1,2}, Patrick J. Doyle^{1,2}, Malcolm R. McNeil^{1,2,3}, Joseph M. Mikolic¹

¹ Geriatric Research Education & Clinical Center, VA Pittsburgh Healthcare System

² Department of Communication Science & Disorders, University of Pittsburgh

³ Department of Otolaryngology, University of Pittsburgh Medical Center

INTRODUCTION

Purpose

To examine the utility of Rasch analysis in evaluating the validity and sensitivity to change of the Revised Token Test (RTT).

Background

The RTT (McNeil & Prescott, 1978) is a test of auditory language processing that has been well described and shown to differentiate persons with aphasia due to left brain damage from persons without brain damage or aphasia and from persons with right hemisphere lesions (McNeil et al. 1988). All psychometric analyses conducted to date on the RTT have employed classical test theory (CTT) (Nunnally & Bernstein, 1994).

- The 15-point RTT scoring system is an ordinal scale that may be inappropriate for averaging and parametric statistical analysis.
 - Like all other aphasia tests, the RTT treats a raw score behavior count as an interval-level measure, a practice that is viewed with increasing skepticism (Willmes, 2003; Wright & Linacre, 1989).
 - McNeil et al. (1989) derived an interval scale from RTT data using observer judgments of comprehension effectiveness.
 - There were differences in category order between the two scales, as well as a substantive difference in average overall score.
 - The derived interval scale was based on comprehension effectiveness; the intervality of the scoring system for measuring aphasia severity must be established separately.
- Potential negative consequences of treating ordinal-level data as interval measurement include:
 - invalid inferences about whether patient groups perform differently
 - insensitivity to change in performance over time

Rasch analysis (Rasch, 1960) is an alternative and well-established psychometric method that offers potential advantages over CTT. It has been widely applied in health status assessment, but has rarely been employed in aphasia assessment (Willmes, 1981; 1992).

- Rasch theory models the probability of a correct response to a given item as a function of item difficulty and person ability.

- Rasch model fit statistics inform questions of construct validity and identify unexpected patterns of item and person performance.
- Prominent features of Rasch theory and CTT are compared in Table 1.

[INSERT TABLE 1 ABOUT HERE]

Research Questions

1. Does Rasch analysis support the content and construct validity of the RTT?
 - a. Do particular subgroups of items demonstrate poor Rasch model fit?
 - b. Does principal components analysis (PCA) of Rasch model residuals suggest the presence of a single underlying construct (i.e., auditory language processing ability) in the data?
 - c. Do item difficulty estimates follow predicted patterns of intra-subtest homogeneity and inter-subtest differences?
2. Do Rasch-derived measures or traditionally computed scores provide more sensitive measurement of between-group differences and better sensitivity to change?

METHOD

Participants and Procedure

- The 55-item RTT (Arvedson et al., 1986) was administered to 53 right and 54 left-hemisphere stroke survivors at three MPO as a part of a larger protocol investigating patient-reported health status in stroke survivors.
- 32 participants had aphasia and 50 had cognitive-communication impairments, as determined by certified SLPs.
- A sub-sample of this group was re-tested at six MPO (n = 41 RH, 48 LH).
- Each RTT element within a command was scored traditionally and each response was also scored dichotomously, as correct/incorrect.
 - For dichotomous scoring, all responses containing a command element score of 10 (element reversal) or lower were scored as incorrect. All other responses were scored as correct.

Analysis: Item Calibration, Person Ability Estimation, & Content Validity

- Traditional RTT overall scores were calculated as the average of the element scores.
- Rasch item difficulty and person ability scores were derived from the dichotomously scored data, using Winsteps (Linacre & Wright, 2003).
 - Item difficulty and person ability are reported as logit scores, i.e. the natural logarithm of the odds of a correct response.
 - Item difficulty and person ability are modeled such that the probability of a correct response is 50% when item difficulty and person ability are equal.
 - Item difficulty is estimated first, scaled to mean = 0, SD = 1, and person ability is estimated on this scale.

- Items demonstrating >50% variation between observations and Rasch model expectations were excluded (Linacre & Wright, 1994).

Analysis: Construct Validity

- PCA of model residuals was used to calculate the proportion of variance in the data accounted for by the dimension of item difficulty-person ability.

Analysis: Known-Groups Validity & Sensitivity to Change

- Traditional and Rasch-derived RTT person scores were each submitted to a two-way ANOVA with group (right-hemisphere without aphasia vs. left-hemisphere with aphasia) and time post onset (3 MPO vs. 6 MPO) as factors ($\alpha=.05$).
 - Sixty-two subjects (n = 33 RH, 29 LH with aphasia) were included.
 - Subjects answering all items correctly were excluded because Rasch estimates of extreme scores are potentially biased.

RESULTS

Item Calibration, Content Validity, & Construct Validity

- Six items were excluded based on poor model fit (>50% variation from expectation): Subtest I command 1; Subtest III command 1; Subtest IX commands 2, 3, 5, & 8.
- PCA of model residuals revealed that the item difficulty-person ability dimension accounted for 70.6% of the variance; the next factor accounted for 1.4%.
- Individual item difficulty values are presented in Figure 1.
- The mean item difficulty values for each subtest are shown in Figure 2, along with the average of the right and left-brain damaged 50th %ile scores for each subtest reported by McNeil and Prescott (1978).

[INSERT FIGURES 1 AND 2 ABOUT HERE]

Known-Groups Validity and Sensitivity to Change

- Traditional RTT overall score
 - ANOVA revealed a significant main effect of group only ($p = 0.001$, effect size (ES, η^2) = 0.18).
 - Neither the main effect of time post onset ($p = 0.08$, ES = 0.05) nor the group X time interaction ($p = 0.16$, ES = 0.03) was significant.
- Rasch-derived RTT score
 - ANOVA revealed a significant main effect of group ($p = 0.001$, ES = 0.16) and time post onset ($p = 0.01$, ES = 0.10).
 - The group X time interaction was not significant ($p = 0.15$, ES = 0.04)
- Mean scores are presented in Figure 3.

[INSERT FIGURE 3 ABOUT HERE]

DISCUSSION

Content and Construct Validity

- Rasch analysis generally supported the content and construct validity of the RTT.
 - The pattern of average item difficulty for each subtest was consistent with item content and similar for Rasch and CTT-derived values.
 - PCA of model residuals suggested the presence of a unidimensional construct.
- Rasch analysis failed to support certain aspects of the instrument's validity.
 - Intra-subtest spread of item difficulty values suggests that the subtests are not necessarily homogenous with respect to item difficulty, as originally proposed.
 - Willmes (1981) found similar results with a German version of the Token Test (Orgass, 1976).
 - This finding may be the result of a relatively small and poorly targeted patient sample; a larger and more severely impaired sample might obtain less intra-subtest item variability.
 - Six items excluded due to poor model fit either came from subtest IX or were the initial items of subtests I and III.
 - Subtest IX differs from previous subtests by inclusion of adverb phrases that unimpaired subjects sometimes find confusing.
 - The initial item of subtest III differs from previous items in that it contains a two-part command; subtests III-XIII all follow the pattern of responding using two tokens.
 - Most excluded items were poorly targeted to the participant sample; misfit might result from the small number of incorrect responses to these items.

Known-Groups Validity and Sensitivity to Change

- Traditional and Rasch-derived RTT scores demonstrated similar sensitivity to differences between right-hemisphere stroke survivors without aphasia and left-hemisphere stroke survivors with aphasia.
- Rasch-derived scores were more sensitive to change over time, showing a small-moderate effect size that was statistically reliable in the current sample; the traditional RTT score showed a small effect size that failed to reach significance.
 - This finding is remarkable because Rasch scores were computed from dichotomized data, while the traditional scores were based on a 15-point multidimensional scale
 - Traditional scoring of the RTT weights each individual element equally, and treats ordinal category assignment as interval measurement.
 - In Rasch scoring, a correct response to a harder item results in a higher ability estimate than a correct response to an easier item; the logit transformation explicitly converts ordinal data into interval measurement.

Future Directions

- High person ability estimates relative to item difficulty values suggest that, like other aphasia tests, the RTT may be insensitive to subtle impairments; inclusion of more difficult items should be considered.
- Rasch models can be used to develop computer adaptive tests, which have the potential to reduce testing time and response burden while maintaining measurement precision.

Classical Test Theory		Rasch Theory	
Property	Consequence	Property	Consequence
Scores depend directly on items used to estimate them; test psychometrics depend on normative sample characteristics	Scores from different tests measuring same construct are not directly comparable; all items must be administered to derive an interpretable score	Abstracted measures are independent of items used to estimate them; test properties are sample-independent	Scores from different tests are more easily equated and directly comparable; ability estimates from partial data are robust
Raw score behavior counts/ordinal data are treated as interval measures	Distorts measurement, esp. at ability extremes; may be insensitive to change over time	Abstracts true interval-level measures from ordinal raw data	Linear measurement across ability range; more sensitive & valid measurement of change
Test reliability depends on the heterogeneity of the normative sample	Pearson r is predominant index; single SEM not valid across range of ability	Computes error variance as a function of ability and item difficulty	Provides valid indices of precision across the range of measurement
Will tolerate a range of normative sample sizes	May be used with smaller samples so long as they are representative	Requires large, well-targeted samples for stable item parameter estimates	Dichotomous items require $n \geq 100$, more for calibration of polychotomous items
Permits generalization of the model to fit observed data; emphasizes description over prediction	Leads to proliferation of models, which may obscure measurement constructs	Requires data to fit a priori model; emphasizes prediction over description	Leads to increasing specification of measurement constructs; enhances reliability and validity of measurement

Table 1. Comparison of properties of classical test theory and Rasch theory

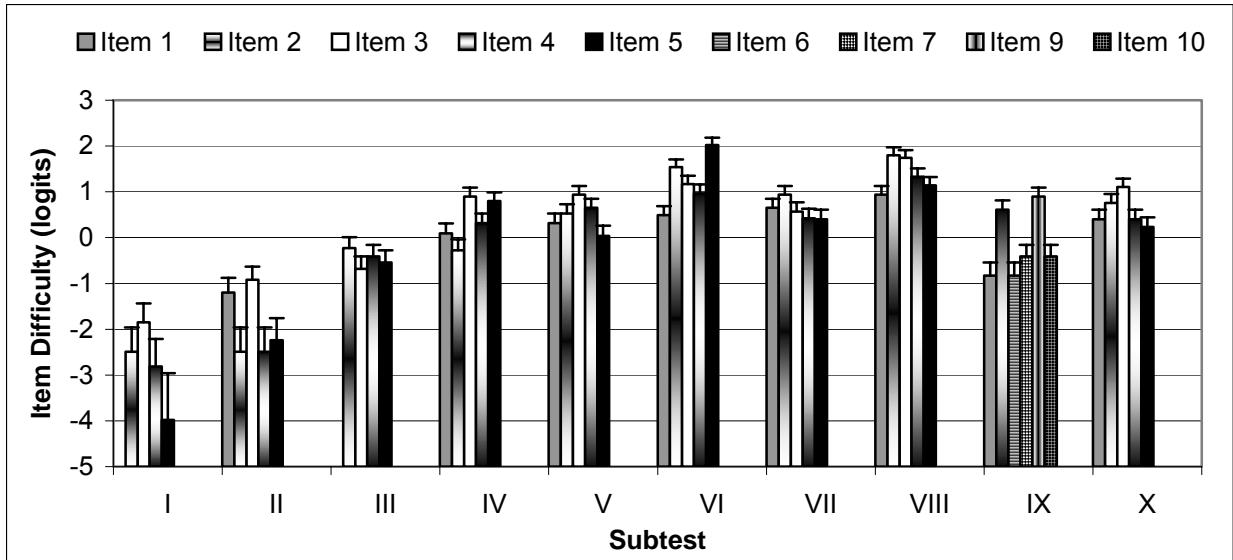


Figure 1. Individual estimated item difficulty values, reported in logits, i.e. the log odds of a correct response. Items excluded due to misfit (I.1, III.1, IX.2, 3, 5, & 8) are omitted from the graph. Error bars represent one standard error.

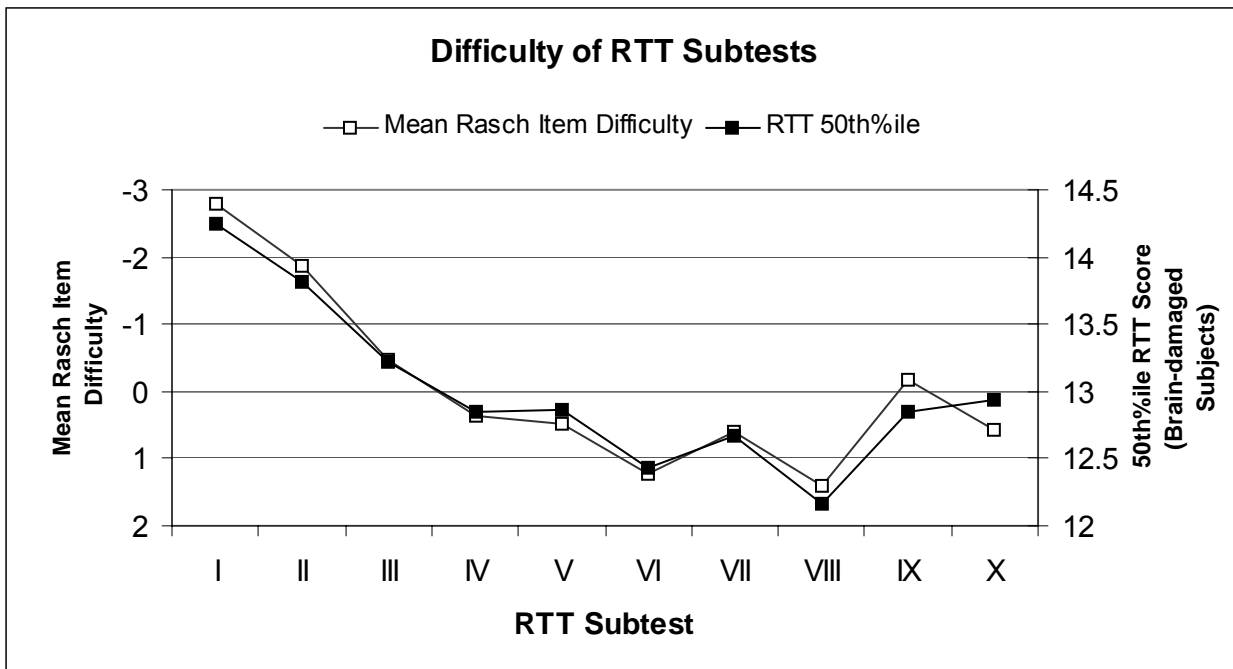


Figure 2. Comparison of Rasch and CTT-derived measures of RTT subtest difficulty. The value axis for Rasch item difficulty is reversed to make the two sets of values directly comparable.

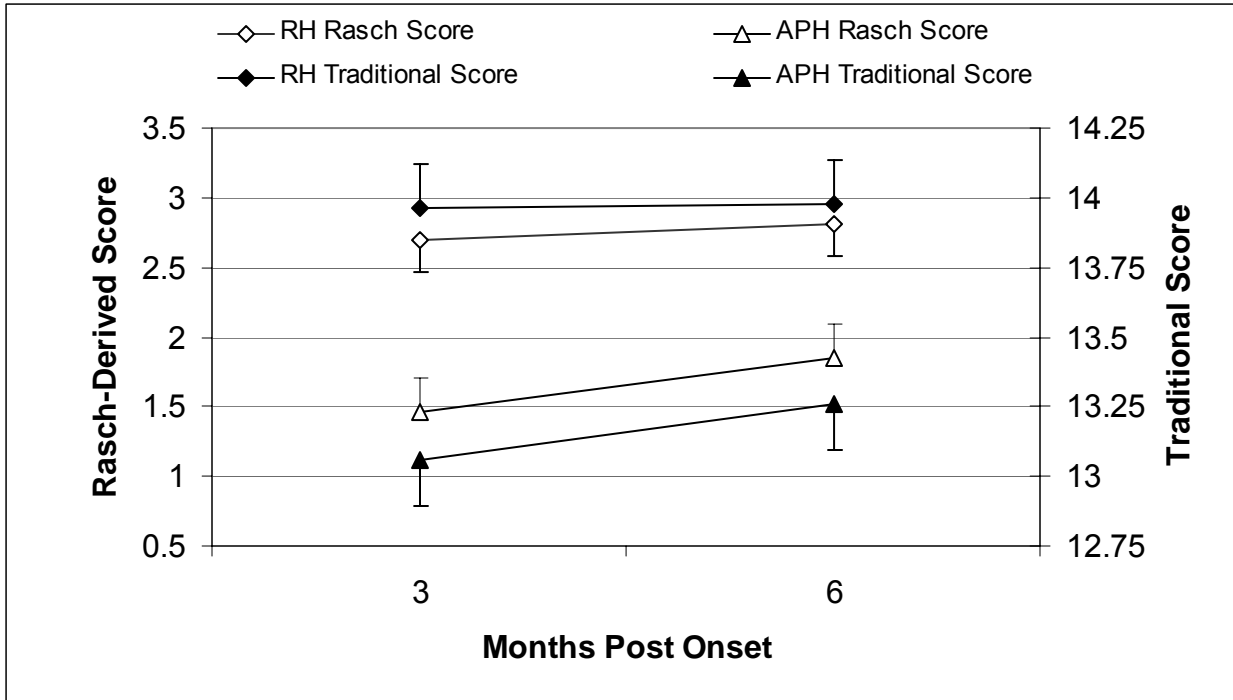


Figure 3. Mean traditional RTT overall and Rasch-derived RTT scores for right-hemisphere stroke survivors without aphasia (RH, n = 33) and left-hemisphere stroke survivors with aphasia (APH, n =29) at 3 and 6 months post onset (MPO). Error bars represent one standard error.