Article

Perceptual and Motor Skills
2018, Vol. 126(1) 70–86
© The Author(s) 2018
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/0031512518812183
journals.sagepub.com/home/pms
SAGE

Is the d2 Test of Attention Rasch Scalable? Analysis With the Rasch Poisson Counts Model

Purya Baghaei[1], Hamdollah Ravand[2], and Mahsa Nadri[1]

Abstract

The d2 test is a cancellation test to measure attention, visual scanning, and processing speed. It is the most frequently used test of attention in Europe. Although it has been validated using factor analytic techniques and correlational analyses, its fit to item response theory models has not been examined. We evaluated the fit of the d2 test to the Rasch Poisson Counts Model (RPCM) by examining the fit of six different scoring techniques. Only two scoring techniques—concentration performance scores and total number of characters canceled—fit the RPCM. The individual items fit the RPCM, with negligible differential item functioning across sex. Graphical model check and likelihood ratio test confirmed the overall fit of the two scoring techniques to RPCM.

Keywords

sustained attention, selective attention, d2 test, Rasch Poisson Counts Model, validity

[1] English Department, Mashhad Branch, Islamic Azad University, Mashhad, Iran
[2] English Department, Vali-e-Asr University, Rafsanjan, Iran

Corresponding Author:
Purya Baghaei, English Department, Islamic Azad University, Ostad Yusofi St., 91871-Mashhad, Iran.
Email: pbaghaei@mshdiau.ac.ir

Introduction

Attention is a basic neurocognitive function that is a prerequisite for performance on more complex cognitive tasks (Cooley & Morris, 1990). Deficits in attention are symptoms of several disorders in individuals with acquired brain injury (Sohlberg & Mateer, 2001) and other disorders including schizophrenia and metabolic disturbances (Mirsky, Anthony, Duncan, Ahearn, & Kellam, 1991). Furthermore, affective and anxiety disorders such as depression and bipolar disorder coincide with impairments of attention.

Research on identifying the cognitive profiles of adults with ADHD demonstrates that they suffer from difficulties in attention, working memory, inhibition, and flexibility (Fuermaier et al., 2013). Inattention and problems of concentration are deemed to be the core symptoms of attention deficit hyperactivity disorder (ADHD). Therefore, measuring attention is part of the diagnosis for ADHD (American Psychiatric Association, 2013).

The d2 test of attention is a widely used paper and pencil measure of sustained and selective attention in which both components of speed and accuracy have been taken into consideration in its scoring system. It is a cancellation test in which respondents have to cross out target variables among similar nontarget stimuli (Brickenkamp & Zillmer, 1998).The target variables are ds' with two dashes above or below them. The targets are randomly interspersed among nontarget characters. Nontarget characters are d's with one, three, or four dashes above or below them and p's with one, two, three, or four dashes above or below them. The target and nontarget characters are presented in 14 consecutive lines. Separate time limits of 20 seconds are allotted for each line with no pause between the lines. Because the d2 test is timed and requires focus on the target stimuli among background irrelevant noise, it is also considered a measure of speed, scanning accuracy, and selective attention.

The construct validity and reliability of the test have been investigated with factor analysis and against criterion measures. Brickenkamp and Zillmer (1998) demonstrated that the d2 test correlates with symbol digit modalities test (Smith, 1973), Stroop (1935) color word test, and trail making test Parts A and B (Reitan &Wolfson, 1985), all of which are measures of attention, scanning, and mental flexibility. Brickenkamp and Zillmer (1998) also showed that the test has a weak correlation with performance and verbal subtests of the Wechsler Adult Intelligence ScaleRevised (Wechsler, 1981), which was deemed as evidence of divergent validity for the d2 test. Researchers have also reported high retest reliability estimates for the test (Brickenkamp & Zillmer, 1998; Lee, Lu, Liu, Lin, & Hsieh, 2017; Steinborn, Langner, Flehmig, & Huestegge, 2018). In fact, the d2 test is regarded as an extremely reliable test. Steinborn et al. (2018) recently, using an interpolated resampling technique (cumulative reliability function analysis), demonstrated that the d2 reliability is retained even when test length is reduced by 50%. One hypothesized reason for the exceptional test–retest reliability of the d2 test is the mode of administration. In particular, it is believed that the participant's task motivation is constantly refreshed by both the experimenter's presence and the time-critical instructions to shift to the next line, both of which seem to encourage participants to give their best performance (cf. Pieters, 1985; Steinborn, Langner, & Huestegge, 2017; Van Breukelen et al., 1995, for further theoretical considerations).

In an attempt to demonstrate the validity of the d2 test with an American sample and extend its scoring system, M. E. Bates and Lemay (2004) administrated it along with several other neuropsychological measures to a relatively large participant sample. The other tests employed in their study included measures of attention, processing speed, abstract reasoning, verbal ability, visual spatial ability, and working memory. Results demonstrated that the test is internally consistent as measured by Cronbach's alpha. Principal component analysis of all the measures along with different subscores of the d2 (as separate variables) revealed that five factors can be extracted from the data. The first factor was a speed factor with the total number of characters processed, the total number of characters correctly processed, concentration performance (CP; total number of correctly canceled minus total number incorrectly canceled), and the digit symbol substitution test loading on this factor. A second factor, named scanning accuracy, was comprised of total errors, errors of omission, and percent errors. An intelligence factor, a scanning deterioration or acceleration factor, and a memory factor also emerged. These findings were interpreted as convergent and discriminant validity evidence for the d2 test.

To our knowledge, no study so far has examined the fit of the d2 test to item response theory (IRT) models (Birnbaum, 1968). IRT models are a class of psychometric theories which model the relationship between an examinee's performance on an item and the examinee's overall location on the latent trait. The probability of a correct response to an item is assumed to be a function of a persons' ability and some item parameters. With a higher ability parameter, the probability of a correct response is expected to increase. The relationship between ability and probability of a correct response is depicted graphically in a set of graphs called item characteristic curves which are the main tool for assessing item quality. Therefore, one straightforward way to examine the psychometric quality of an item is to check whether examinees with higher locations on the latent trait have higher probabilities of endorsing an item.

Although intelligence tests are commonly analyzed using IRT models, processing speed tests are still evaluated using classical test theory because the structure of speed tests does not match the requirements of most IRT models (Doebler & Holling, 2016). Usually in speed tests, there are several simple tasks, and the unit of analysis is a count of correct answers to these tasks. This is unlike individual right or wrong or Likert-type items that are commonly fed into IRT models. IRT models are very flexible and allow for detailed item analysis, provision of standard errors of measurement for different ability levels, computerized adaptive testing, optimal planning of test designs, test equating, and differential item functioning (DIF; Doebler & Holling, 2016).

In this study, we examine the fit of the d2 test to the Rasch Poisson Counts Model (RPCM, Rasch, 1960/1980). The structure of the test, that is, a combination of 14 lines of stimuli each with a separate time limit, makes it an ideal candidate for RPCM scaling. Thus, in this study, the overall fit of the d2 test to RPCM, the fit of the individual items (lines), and the reliability of the test are all examined.

The RPCM Measurement Model

The RPCM (Rasch, 1960/1980) is a unidimensional member of the family of Rasch models (RMs). It is used for timed tests where counts of correct replies or errors, within each task, are modeled instead of replies to individual items (Doebler, Doebler, & Holling, 2014; Jansen, 1997). Such testing conditions arise in speeded neuropsychological or psychomotor tests in which respondents must tick off an unlimited number of items within a fixed time period (Spray, 1990). The RPCM is expressed as follows:

$$p(Y_{vi} = y_{vi}) = \frac{\exp(-\mu_{vi})\mu_{vi}^{y_{vi}}}{y_{vi}!}$$

where y_{vi} is the total raw score or counts of errors for person v on part i of the test.The scores on the tasks are assumed to be independent conditional on the parameters $\mu_{vi'} v = 1, \ldots N, i = 1, \ldots, l$ and Poisson distributed. μ_{vi} is the mean of the raw scores or errors on part i of the test for person v. This average is assumed to have a multiplicative composition and is the product of person's ability θ_v and item's easinessσ_i:

$$\mu_{vi} = \theta_v \sigma_i$$

It is possible to estimate θ_v and σ_i independent of each other; hence, separability of parameters holds (for derivation see Rasch, 1960/1980). RPCM has been applied in psychomotor testing (Spray, 1990), the testing of attention or processing speed (Doebler & Holling, 2015), oral reading errors (Jansen, 1997; Rasch, 1960/1980; Verhelst & Kamphuis, 2009), reading comprehension (Verhelst & Kamphuis, 2009), and divergent thinking (Forthmann et al., 2016).

The fit of data to a latent trait model, such as the RM, is evidence that the covariation among the test items is caused by an underlying latent factor which could be the intended construct and is, therefore, considered validity evidence (Baghaei & Tabatabaee-Yazdi, 2016; Borsboom, 2008). When the RM fits, the homogeneity of the latent variable is supported. "More" or "less" of a latent trait only makes sense if the trait is homogeneous. Needless to say, when the RM does not hold, adding raw scores to compute an overall score is not warranted, as we are then adding components of a heterogeneous latent variable. If a set of items measure a single latent variable, "then the Rasch model is the necessary and sufficient conceptualization. If they do not, then the set of items contains a mixture of variables and there is no simple, efficient, or unique way to know their utility for measuring anything" (Wright, 1977, p. 224). The fit of the RM is evidence that the latent variable is quantitative, and items and the latent variable can be measured on an interval scale with a common unit of measurement (Wright, 1988).

Method

Participants and Instrument

We administered the d2 test of attention according to its standard procedures (Brickenkamp & Zillmer, 1998) to 138 nonclinical Iranian university students (68% female). The age range was 19 to 52 years ($M = 24.26$, standard deviation [SD]=5.64). The data were collected as part of a project on the cognitive correlates of listening comprehension in English as a foreign language (Nadri, 2018). As mentioned earlier, the test consists of 14 lines of characters where respondents should cross out d's with two dots in 20 seconds. This time limit is allotted for each line separately. This structure makes the test optimal for RPCM analysis. Participation in the study was voluntary and no financial compensation was made. On conclusion, participants were thanked for their cooperation and time and were provided with the profiles of their cognitive and English language skills. The institutional review board approved the study and waived the need for participants' written consent (IRB decision # 145/د).

Results

There are a number of scores that are computed on the basis of d2 test performance, including the total number of characters canceled (TN, total number—an index of processing speed), errors of commission (C, nontarget characters canceled), errors of omission (O, target characters respondents failed to cancel), total errors (TE, sum of the errors of commission and errors of omission), the number of characters correctly canceled minus the number of errors of commission (CP), and the number of characters correctly canceled minus the sum of the errors (TN-TE) (Brickenkamp & Zillmer, 1998). These scores are separately calculated for each of the 14 lines of the test. In this analysis, each line of the test is considered an item and is a unit of analysis.

We ran six separate RPCM analyses on the six scoring techniques mentioned earlier. We used "lme4" package (D. Bates et al., 2017) in R (R Development Core Team, 2016) to estimate the models.[1] We evaluated global model fit by

Table 1. LRTs With Median of Raw Scores as a Partitioning Criterion for Overall Model Check.

Scoring technique	Chi square	*p*	*df*
O	35.75	.00	13
C	53.85	.00	13
TE	37.34	.00	13
TN	12.82	.46	13
CP	14.98	.31	13
TN-TE	36.57	.00	13

TN = total number; C = nontarget characters canceled; O = errors of omission; CP = concentration performance; TE = total number of errors.

examining differences in item parameters across two subsamples of the data using likelihood ratio tests (LRTs), similar to Andersen's (1973) test for the binary RM. The approach is based on the invariance requirement of the RM which states that the item parameters should remain constant in different subsamples of the data (Baghaei, Yanagida, & Heene, 2017). In this approach, in the first step, the sample is partitioned according to a criterion such as sex or raw score median. Two new models are then estimated. In the first model, a group-specific intercept is added. The item parameters are assumed to be equal in this model for both groups. In a subsequent model, ''group'' by "item" interaction terms which represent the group-specific deviations in item difficulty are added. In other words, the items are allowed to have different difficulty parameters across the groups. The interaction of group and the individual items, that is, the difference in item difficulty for respondents in one of the groups relative to those in the other group, is accounted for in this model. The two models are then compared with LRT. If the model with the interaction term fits better, it means that this model accounts for more variability in the data and the item parameters are not constant across the groups; hence, the RPCM does not hold (Baghaei & Doebler, 2018).

To evaluate the overall fit of the d2 test, we partitioned the data on the basis of test takers' median raw scores for each scoring technique and performed LRT for the low scorers versus high scorers. Table 1 shows that the LRT was nonsignificant when the CP and TN scores are computed. Therefore, the d2 test fits the RPCM only when these two scores are computed, while other scoring techniques result in misfit to the RM.

We also conducted graphical model checks to examine which scoring type is a better fit to the model. In the graphical check, predicted values for each person are plotted against their standardized Pearson residuals. Roughly symmetrical Pearson residuals with few outliers confirm acceptable fit (Baghaei & Doebler, 2018). Furthermore, standardized (Pearson) residuals are normally distributed

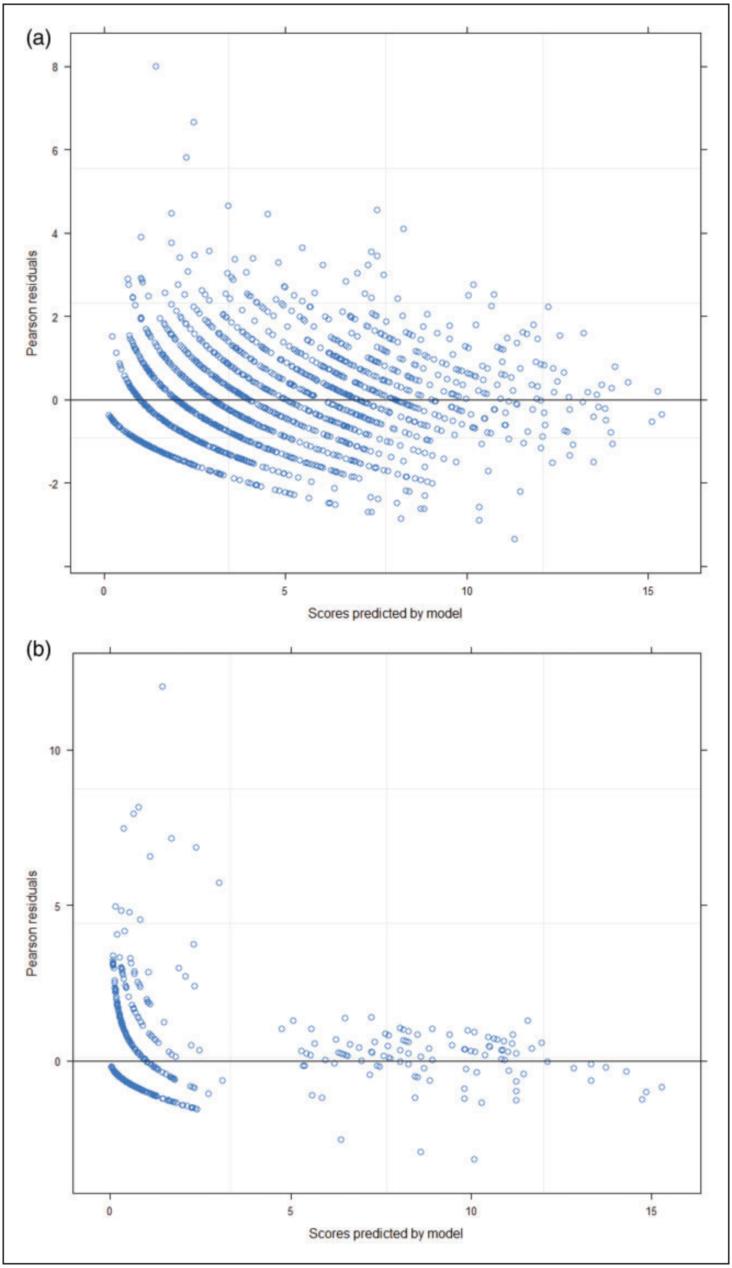

Figure 1. Graphical model check for different scoring techniques: (a) Errors of omission, (b) errors of commission, (c) total errors, (d) total characters canceled, (e) concentration performance, and (f) characters correctly canceled minus the sum of the errors.

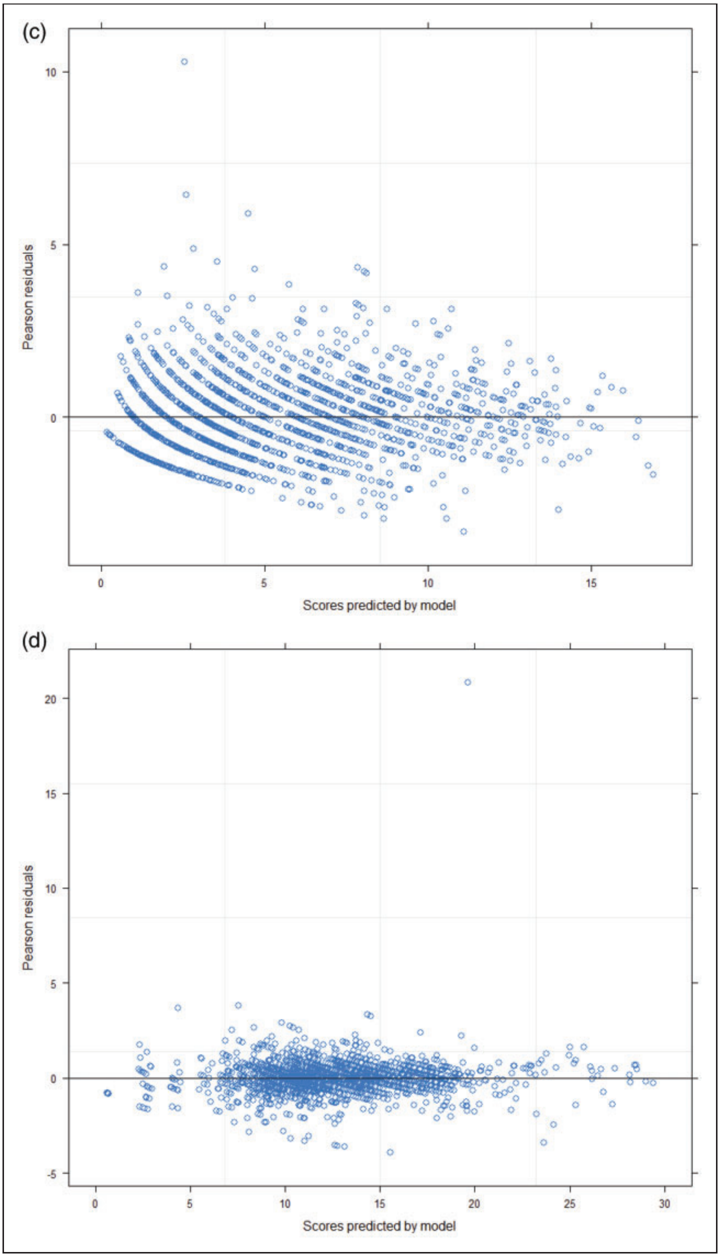

Figure 1. Continued.

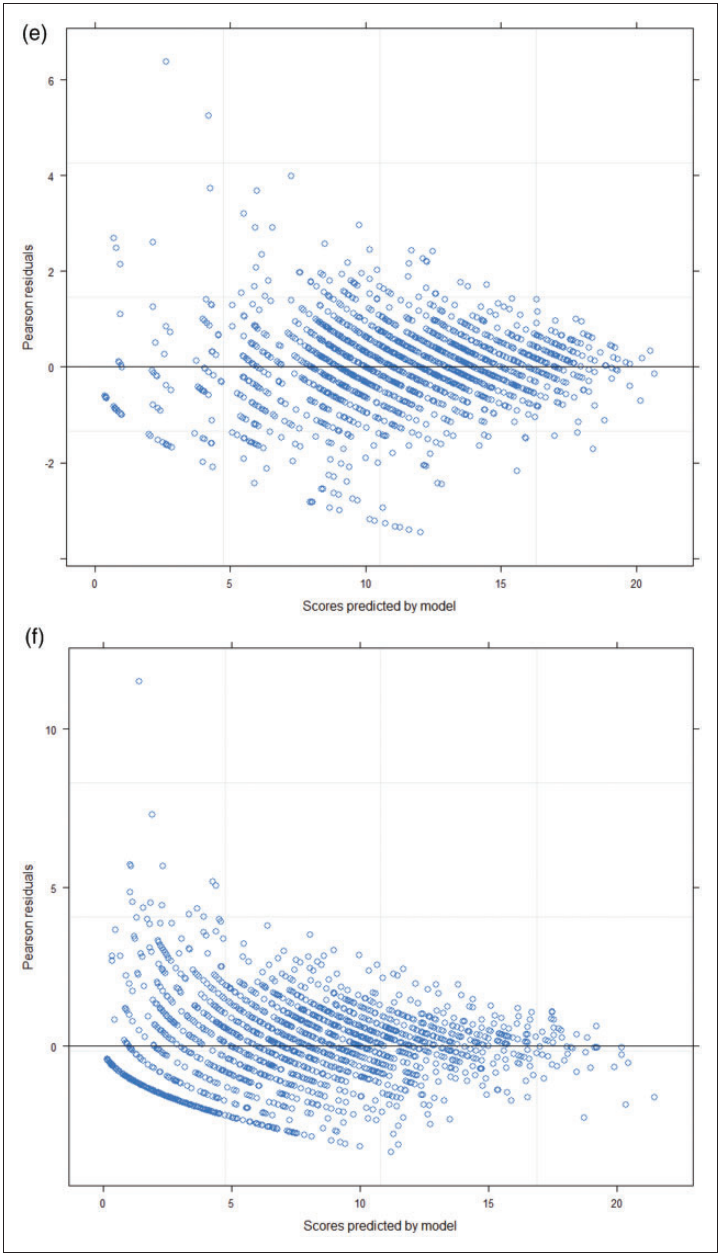

Figure 1. Continued.

Table 2. Item Easiness Parameters, their Standard Errors, and Chi-Square Fit Values for the CP and TN Scores.

	CP			TN		
Item	Estimate	*SE*	Fit	Estimate	*SE*	Fit
1	2.27	.068	.91	2.44	.053	3.90
2	2.21	.068	1.64	2.43	.053	2.41
3	2.29	.068	1.80	2.47	.053	1.48
4	2.25	.068	1.13	2.43	.053	.84
5	2.28	.068	1.05	2.46	.053	.89
6	2.22	.068	1.75	2.40	.053	1.08
7	2.21	.068	1.42	2.44	.053	20.58
8	2.27	.068	6.39	2.44	.053	2.82
9	2.21	.068	1.54	2.38	.053	.31
10	2.19	.068	.39	2.35	.053	.43
11	2.21	.068	3.43	2.37	.053	3.39
12	2.13	.068	4.33	2.37	.053	4.70
13	2.17	.068	.73	2.36	.053	5.50
14	2.21	.068	2.65	2.40	.053	1.62
SD (latent ability)	.73		.55			

Note. CP = concentration performance; TN = total number canceled.

with mean 0 and *SD* of 1.0 for good model fit. The graphs in Figure 1 also indicate that the CP scores (i.e., the number of characters correctly canceled minus the number of characters incorrectly canceled) and the TN scores (i.e., the total number of characters canceled) have the best fit. Because the C, O, E, and TN-E scores did not have a good fit, further analyses were run only on the CP and TN scores.

Table 2 shows the item easiness parameters, their standard errors, and their chi-square fit values for the two types of score. A chi-square type item fit statistic based on binning observed and predicted values showed that none of the item misfits in the two scoring techniques ($df = 5$, the number of subsets specified in the person scores). The *SD* for the latent ability parameters in the CP score was .73 and in the TN score was 55.

In both scoring procedures, item easiness parameters were very close to each other, suggesting that they are not significantly different. Two other analyses were run assuming that all the 14 items (lines) were equally difficult. For the CP scores, the information criteria and the deviance statistic showed that this model had a worse fit compared with the model where item difficulties were assumed to differ, χ^2 (13) = 37, $p < .01$.The same result was observed for the TE scores,

χ^2 (13) = 33.87, $p < .01$. Therefore, the difficulties of the liens significantly varied, regardless of the scoring method.

Gender DIF

CP scores. We examined gender DIF and global goodness of fit by investigating differences in item parameters across sexes with an LRT. The interaction of sex and the individual items (i.e., the difference in item difficulty on log-scale for males relative to females) was significant only for Item 2 (contrast = .26, $p < .01$). The LRT, however, showed that the model with the interaction term did not explain more variability in the data, χ^2 (13) = 18.84, $p = .12$. In other words, the item parameters remained constant across the two partitions of the sample, and the data fit the RPCM.

TE scores. We examined DIF and global fit across sex for the TE scores. DIF analysis across sex showed that Items, 5, 6, and 8 manifested negligible DIF with contrasts equal to .22, .20, and .24, respectively ($p < .01$). Nevertheless, an LRT showed that the model with sex as the interaction term did not explain more variability in the data, χ^2 (13) = 15.86, $p = .25$. This means that the item parameters remained invariant when sex was a partitioning criterion. Therefore, the magnitude of observed DIF for the three items across sex is harmless.

Reliability analysis. We estimated the ability-specific reliability of the measures (Baghaei & Doebler, 2018). Figure 2 shows the reliability estimates for different locations on the ability continuum for the CP and TE scores. The index of reliability was more than .80 at the lowest level for the CP scores and augmented to above .96 and was above .90 for a large portion of the ability continuum. For the TN scores, reliability was slightly lower than .80 at its lowest level but augmented to above .95 at highest levels and was very high for a large section of the ability scale.

Discussion

The d2 test of attention is a short and easy-to-administer measure of sustained attention. It is based on a well-grounded theoretical framework and is relatively well researched. The test has sound psychometric properties, and its validity has been demonstrated against other criterion measures by providing divergent and discriminant evidence. The aim of this study was to contribute to the validity literature of the test by examining its fit to a unidimensional RM. We examined the psychometric functioning of the d2 test of attention by investigating the fit of the individual items as well as the overall fit of the test to the unidimensional RPCM. The RPCM was chosen as the psychometric model to fit to the d2 test, as time limits are imposed on the individual lines and the counts of correct

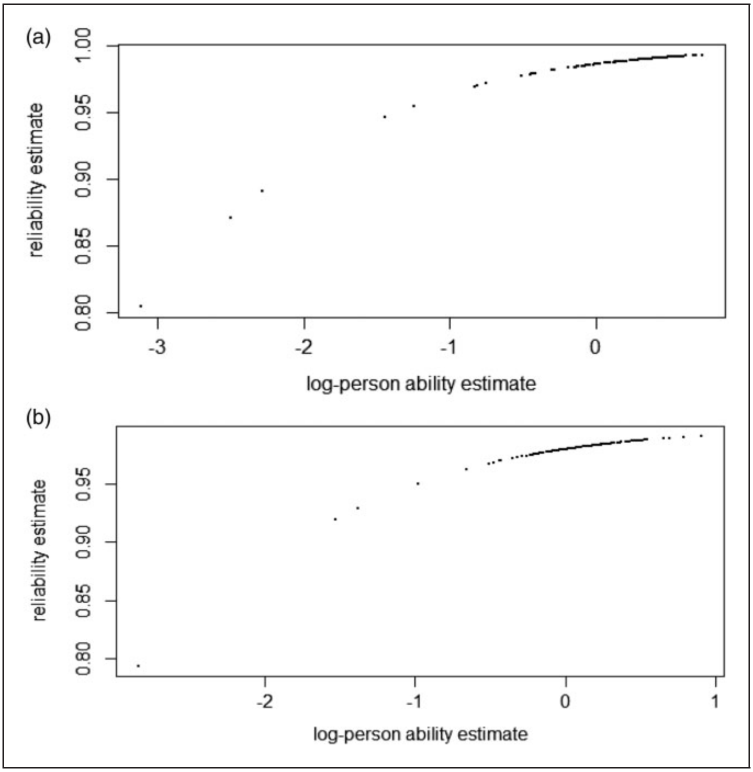

Figure 2. Ability-specific reliability graph: (a) Concentration performance scores and (b) total characters canceled scores.

replies or errors are modeled. Graphical inspection and LRT with sex and median of scores as partitioning criteria were utilized to evaluate the fit of the data to the model.

Our findings suggest that the test fits the RPCM best when the CP scores (total number of characters processed minus errors of commission) and total number of characters canceled are modelled. The model did not fit when it was applied to the errors of omission, errors of commission, total errors, and total correctly processed. This is consistent with previous research on the psychometric quality of the d2 test. In a recent study, Steinborn et al. (2018) examined the reliability of the d2 test and concluded that "...(a) only the speed score (and error-corrected speed score) is eligible for highly reliable assessment, [and] that (b) error scores might be used as a secondary measure (e.g., to check for aberrant behavior)..."(p. 339).

For both favored scoring types, none of the 14 items misfits the model according to a chi square test that compared the observed score for the items with the score predicted by the Poisson function. A few items showed negligible DIF across sex for both scoring types. However, the LRT with sex as a partitioning criterion was nonsignificant, supporting unidimensionality. When the sample was divided according to the median of CP and TE scores, all items were invariant and the LRT was nonsignificant.

The item parameters had a limited range in both scoring procedures. Nevertheless, the better fit of a model with different item parameters compared with a model where item difficulties were assumed to be identical indicated that item difficulties significantly vary. Ability-specific reliability measures showed that test reliability ranges between .80 and .97 for different sections of the ability scale for both CP and TE scores.

The results of the RPCM analysis demonstrated that the d2 test is an internally valid and accurate measure of attention. Fit of data to the RM is evidence that total raw score is a valid estimator of ability and justifies the use of sum scores as an indication of respondents' latent trait (Wright, 1989).

IRT models in general and RPCM in particular can be used as powerful psychometric models to evaluate neuropsychological tests where respondents have to perform simple tasks speedily within limited response time allotments. Under such testing conditions where test takers are supposed to identify target stimuli, every target is a potential item and it is rather difficult to envisage clear cut individual items. In such situations RPCM can be employed, as in this model, the total number of correct replies or errors within each time allotment is the unit of analysis instead of individual stimuli. Another advantage of the RPCM is that it is less complex than polytomous IRT models that might seem to be viable alternatives in such situations and, therefore, requires smaller sample sizes. Furthermore, commonly used polytomous IRT models such as the partial credit model (Masters, 1982) and the rating scale model (Andrich, 1978) do not incorporate the time limit into item difficulty estimation, and it is not possible to untangle the impact of task complexity from the impact of time constraint on item parameters.

A limitation of this study is that we examined overall fit using LRT only across sex and score because we had a relatively small sample size, and other classifications we had at our disposal (e.g., handedness or whether respondents wear glasses) were extremely small. Future research should examine fit across other partitions of the test takers. The RPCM analysis demonstrated that the d2 test is an efficient and precise instrument for measuring attention.

Summary and Conclusions

The findings of this study can be summarized as follows: (a) The d2 test fitted the RPCM best when the CP scores and the total number of characters canceled were modeled; (b) none of the 14 items (lines) misfits, according to a chi square test in the CP and TN scoring techniques, supporting unidimensionality; (c) few items showed negligible DIF across sex for the two scoring techniques; (d) LRTs with sex and the median of raw scores as partitioning criteria were nonsignificant for both scoring techniques, supporting model fit; (e) the test is highly reliable across a wide range of the ability continuum; and (f) the d2 test is an internally valid and accurate measure of attention.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Purya Baghaei http://orcid.org/0000-0002-5765-0413

Note

1. For details on RPCM estimation and R codes, see Baghaei and Doebler (2018).

References

American Psychiatric Association (2013). *Diagnostic and statistical manual of mental disorders*. Washington, DC: American Psychiatric Press.

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*, 123–140.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika*, *43*, 561–573.

Baghaei, P., & Doebler, P. (2018). Introduction to the Rasch Poisson Counts Model: An R tutorial. *Psychological Reports*. Advanced online publication. doi: 10.1177/0033294118797577

Baghaei, P., Yanagida, T., & Heene, M. (2017). Development of a descriptive fit statistic for the Rasch model. *North American Journal of Psychology*, *19*, 155–168.

Baghaei, P., & Tabatabaee-Yazdi, M. (2016). The logic of latent variable analysis as validity evidence in psychological measurement. *The Open Psychology Journal*, *9*, 168–175.

Bates, D., Maechler, M., Bolker, B., Walker, S., Christensen, R., Singmann, H., et al. (2017). *lme4: Linear mixed-effects models using 'Eigen' and S4* (R package, version 1.1-14). Retrieved from https://cran.r-project.org/web/packages/lme4/index.html

Bates, M. E., & Lemay, E. P. Jr. (2004). The d2 test of attention: Construct validity and extensions in scoring techniques. *Journal of International Neuropsychological Society*, *10*, 392–400.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 395–479). Reading, MA: Addison-Wesley.

Borsboom, D. (2008). Latent variable theory. *Measurement*, *6*, 25–53.

Brickenkamp, R., & Zillmer, E. (1998). *The d2 test of attention*. Seattle, WA: Hogrefe & Huber Publishers.

Cooley, E. L., & Morris, R. D. (1990). Attention in children: A neuropsychologically based model for assessment. *Developmental Neuropsychology*, *6*, 239–247.

Doebler, A., Doebler, P., & Holling, H. (2014). A latent ability model for count data and application to processing speed. *Applied Psychological Measurement*, *38*, 587–598.

Doebler, A., & Holling, H. (2016). A processing speed test based on rule-based item generation: An analysis with the Rasch Poisson Counts Model. *Learning and Individual Differences*, *52*, 121–128. doi: 10.1016/j.lindif.2015.01.013

Forthmann, B., Gerwig, A., Holling, H., Çelik, P., Storme, M., & Lubart, T. (2016). The be-creative effect in divergent thinking: The interplay of instruction and object frequency. *Intelligence*, *57*, 25–32.

Fuermaier, A. B. M., Tucha, L., Koerts, J., Aschenbrenner, S., Westermann, C., Weisbrod, M., Lange, K.W., Tucha, O. (2013). Complex prospective memory in adults with attention deficit hyperactivity disorder. *PLoS One*, *8*, e58338. doi:10.1371/journal.pone.0058338

Jansen, M. G. H. (1997). Applications of Rasch's Poisson counts model to longitudinal count data. In R. Langeheine & J. Rost (Eds.), *Applications of latent trait and latent class models in the social sciences* (pp. 380–388). Münster, Germany: Waxmann.

Lee, P., Lu, W.-S., Liu, C.-H., Lin, H.-Y., & Hsieh, C.-L. (2017). Test–retest reliability and minimal detectable change of the D2 test of attention in patients with schizophrenia. *Archives of Clinical Neuropsychology*. Advance online publication. doi: 10.1093/arclin/acx123

Manly, T., Anderson, V., Nimmo-Smith, I., Turner, A., Watson, P., & Robertson, I. H. (2001). The differential assessment of children's attention: The Test of Everyday Attention for Children (TEA-Ch), normative sample and ADHD performance. *Journal of Child Psychology and Psychiatry*, *42*, 1065–1081.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika*, *47*, 149–174.

Mirsky, A. F., Anthony, B. J., Duncan, C. C., Ahearn, M. B., & Kellam, S. G. (1991). Analysis of the elements of attention: A neuropsychological approach. *Neuropsychology Review*, *2*, 109–145.

Nadri, M. (2018). *The Contribution of attention, processing speed, and fluid intelligence to predicting Iranian EFL learners' listening comprehension: Development and validation of two measures of auditory attention*. Unpublished master's thesis. Mashhad: Islamic Azad University.

Pieters, J. P. M. (1985). Reaction time analysis of simple mental tasks: A general approach. *Acta Psychologica*, *59*, 227–269.

Rasch, G. (1980). Probabilistic models for some intelligence and attainment tests *(expanded edition)*. Copenhagen: Danish Institute for Educational Research. (Original work published 1960).

R Development Core Team. (2016). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http: //www.R-project.org/

Reitan, R., & Wolfson, D. (1985). *The Halstead-Reitan neuropsychological battery: Theory and clinical implications*. Tucson, Arizona: Neuropsychology Press.

Smith, A. (1973). *Symbol digit modalities test*. Los Angeles, CA: Western Psychological Services.

Sohlberg, M. M., & Mateer, C. A. (2001). Improving attention and managing attentional problems. Adapting rehabilitation techniques to adults with ADD. *Annals of the New York Academy Sciences*, *931*, 359–375.

Spray, J. A. (1990). One-parameter item response theory models for psychomotor tests involving repeated independent attempts. *Research Quarterly for Exercise and Sport*, *61*, 162–168.

Steinborn, M. B., Langner, R., Flehmig, H. C., & Huestegge, L. (2018). Methodology of performance scoring in the d2 sustained-attention test: Cumulative-reliability functions and practical guidelines. *Psychological Assessment*, *30*, 339–357.

Steinborn, M. B., Langner, R., & Huestegge, L. (2017). Mobilizing cognition for speeded action: Try-harder instructions promote motivated readiness in the constant-foreperiod paradigm. *Psychological Research*, *81*, 1135–1151.

Stroop, J. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*, 643–662.

Van Breukelen, G. J. P., Roskam, E. E. C. I., Eling, P. A. T. M., Jansen, R. W. T. L., Souren, D. A. P. B., & Ickenroth, J. G. M. (1995). A model and diagnostic measures for response-time series on tests of concentration: Historical background, conceptual framework, and some applications. *Brain and Cognition*, *27*, 147–179.

Verhelst, N. D., & Kamphuis, F. H. (2009). *A Poisson-gamma model for speed tests*. Measurement and Research Department Reports 2009-2. Arnhem, The Netherlands: Cito.

Wechsler, D. (1981). *WAIS-R : Wechsler adult intelligence scale-revised*. New York, NY: Psychological Corporation.

Wright, B. D. (1977). Misunderstanding the Rasch model. *Journal of Educational. Measurement*, *14*, 219–225.

Wright, B. D. (1988). Rasch model derived from Campbell concatenation: Additivity, interval scaling. *Rasch Measurement Transactions*, *2*, 16. Retrieved from https://www.rasch.org/rmt/rmt21b.htm

Wright, B. D. (1989). Dichotomous Rasch model derived from counting right answers: Raw scores as sufficient statistics. *Rasch Measurement Transactions*, *3*, 62. Retrieved from https://www.rasch.org/rmt/rmt32e.htm

Zillmer, E. A., & Kennedy, C. H. (1999). Preliminary United States norms for the d2 test of attention. *Archives of Clinical Neuropsychology*, *14*, 727–728. doi: 10.1093/arclin/14.8.727.

Author Biographies

Purya Baghaei is an associate professor in the English Department of Islamic Azad University, Mashhad, Iran. His major research interest is foreign language proficiency testing with a focus on the applications of item response theory models in test validation and scaling. He has also conducted research on the role of cognition in second language acquisition.

Hamdollah Ravand is an assistant professor at Vali-e-Asr University of Rafsanjan. His major areas of interest are Language Testing Assessment, Cognitive Diagnostic Modeling, Structural Equation Modeling, Multilevel Modeling, and Item Response Theory.

Mahsa Nadri holds an MA in TEFL (teaching English as a foreign language) from Islamic Azad University, Mashhad, Iran. Her research interests are the cognitive and neuropsychological correlates of foreign language ability.