Applying Many-facet Rasch Measurement to Evaluate the Modified Clock Drawing Test

Roy Surya, Ananta Yudiarso, Marselius Sampe Tondok
University of Surabaya, Indonesia

The *Clock Drawing Test* is usually used to evaluate prospective patients with cognitive decline as a screening tool. This test is well-known for having rapid administration and flexibility across different cultures. The objective of this study was to explore the psychometric properties of the C*lock Drawing Test* using the Many-Facet Rasch Measurement approach. We modified the 18-point clock drawing test, specifically adjusting the scale into three levels: poor, fair, and good. This study also involved the *Global Deterioration Scale* (GDS) as a screening tool to classify participants' cognitive decline status. The Many-Facet Rasch model was applied to analyze 9,045 rating sequences, three clinical psychologist raters, 208 participants, and 15 items. We found the psychometrics information of this modified test was sufficient with the summary of Rasch statistics along with rating scale, item, and rater difficulties analysis. Based on the Wright map, the majority of the items grouped in the middle level, along with the rating scale of this test provided category measure of (-1.73), 0.0, to 1.74. Item 5 "number spacing equal" was the most challenging item for the participants, while item 3, "number all the same (Roman/Arabic)" was the most unchallenging item. Moreover, items 1, 15, and 6 were indicated as misfit items. Rater agreement yielded at 74%. According to ROC analysis, this test effectively predicted participants with mild cognitive impairment and mild dementia based on GDS criteria. Similar studies are recommended to improve the usage of the many facets of Rasch approach in the screening process under many raters' circumstances.

Keywords: Many Facet Rasch Measurement, Clock Drawing Test, Cognitive Screening

There are various types of tests in neuropsychological assessment that are often used to assist in both diagnostic processes and screening purposes. The *Clock Drawing Test* (CDT) is a well-known measurement tool recognized for its rapid administration process (2-5 minutes), (Agrell & Dehlin, 1998; Cullen et al., 2007; Pinto & Peters, 2009). This tool is favored by practitioners due to its flexibility for cross-cultural use (Borson et al., 1999; Storey et al., 2002). Shulman et al. (1986) found the

Author info: Correspondence should be sent to: Ananta Yudiarso, Faculty of Psychology, University of Surabaya, Indonesia ananta@staff.ubaya.ac.id

North American Journal of Psychology, 2024, Vol. 26, No. 1, 95-112.

administration of this tool to be non-threatening and comfortable for elderly patients. Previous studies found that CDT was an efficient screening tool to evaluate cognitive abilities, (Hazan et al., 2018). CDT is also valued for its broad capacity to assess various cognitive dysfunctions, including visual ability, memory, visuospatial skills, planning, abstraction, concentration, understanding, and response processes, (Ismail et al., 2010; Shulman, 2000). Moreover, this test encompasses quantitative and qualitative tests, each with diverse administration techniques, (Lam et al., 1998; Manos & Wu, 1994; Mendez et al., 1992; Sunderland et al., 1989; Wolf-Klein et al., 1989).

Some studies discovered the validity of this test through its correlation with other instruments, for instance, the *Mini-Mental State Examination* (MMSE; Folstein et al., 1975) or the *Montreal Cognitive Assessment* (MoCA; Nasreddine et al., 2005). Initially, Shulman et al. (1993), who developed the oldest scoring system, found a significant correlation between Shulman’s scoring system and MMSE. However, there was some disagreement with that finding because a recent study found that this test did not correlate significantly with MMSE. Ilardi et al. (2020) found that several scoring methods, including Shulman’s version, did not strongly correlate with MMSE. Carneiro (2015), on the other hand, discovered that some scoring versions had a significant correlation with MMSE and MoCA's just in the visuospatial task domain section. Cahn-Weiner et al. (2003) stated that CDT is not designed for diagnosis due to its inability to precisely specify cognitive dysfunction. Several studies categorized CDT as a screening tool for patients with suspected dementia symptoms, (Babins et al., 2008; Kørner et al., 2012; Manos & Wu, 1994; Tabari & Amini, 2021).

Numerous previous validation studies have been conducted, encompassing various approaches such as the construction of new items, modification of existing ones, and even altering the scale. Several noteworthy findings have garnered attention from both researchers and practitioners. For instance, in a study by Ricci et al. (2016), the *Clock Drawing Test* (CDT) items were developed using a Likert scale. Their research used principal component analysis with some additional analyses. While their study successfully elucidated the psychometric properties of the test, it fell short of providing a comprehensive understanding of each item's difficulty level or its interaction with the participants. Similarly, Emek-Savaş et al. (2018) directed their analysis towards evaluating three scoring systems for the CDT. While their findings yielded valuable insights, more comprehensive information could be discovered by analyzing all the items individually. Furthermore, they used the Intraclass Correlation Coefficient (ICC) for reliability measurement in scenarios involving multiple raters and it may offer insights into the consistency of their assessments. However, this approach may encounter challenges in explaining the severity of discrepancies among raters, leaving us with the question of 'how large are these differences?’

Rasch measurement theory has been utilized across many fields for measurement tool development purposes (Alexandrowicz et al., 2018; Batchelder et al., 2020; Camargo & Henson, 2015; Franchignoni et al., 2011; Han & Li, 2015; Natanael, 2021; Park et al., 2021; Petrillo et al., 2015; Zahirah & Susanto, 2021). This method was selected for its capacity to conduct linear and objective measurements, as it treats the Likert scale as ordinal data (Boone, 2016; Sumintono, 2018). Essentially, this analysis aims to test the compatibility of the empirical data with the model fit (de Ayala, 2009). Beyond its logit transformations, this approach also emphasizes various features, including item-person difficulties, fit statistics, standard error, and point-measure correlation. Construct validity in this model can be assessed through Rasch residual principal component analysis or items’ fit statistics, (Linacre, 2011). Moreover, researchers can determine the validity of the rating scale using Andrich threshold analysis (Chong et al., 2022).

The many-facets Rasch model (MFRM) was developed as an advanced technique in the measurement process, taking into consideration various facets (Linacre, 1994; Tavakol & Pinner, 2019). The primary focus of the MFRM is to establish fair measurement with bias estimation that may occur among facets during the evaluation process. Facets can encompass various elements within the evaluation setting, such as raters, places, and task variability (Bond & Fox, 2015). Previous studies have demonstrated that MFRM is commonly employed for evaluation in educational testing (Farlie et al., 2021; Gordon et al., 2021; Huebner & Skar, 2021; Uto, 2021). Nevertheless, this study seeks to extend the application of this method to clinical settings due to its capability to handle evaluations involving multiple raters, providing additional features compared to classical test theory.

To the best of the researchers’ knowledge, no studies have endeavored to examine the psychometric qualities of the clock drawing test using the many-facet Rasch measurement. Consequently, in this study, researchers propose~~d~~ employing a multifaceted Rasch measurement to thoroughly investigate and delineate the psychometric features of this test.

METHOD

Participants

This study received approval from the ethics committee of the University of Surabaya. The data collection process occurred in multiple elderly care facilities in Surabaya. All participants gave informed consent after receiving comprehensive details about the study, and individuals with sensory or motor impairments that could hinder their participation in data collection were screened. No initial diagnoses were made in this research; however, the researcher utilized the *Global Deterioration Scale* (GDS) to offer an initial overview of the participants' conditions.

A total of 208 volunteers, comprising 40.3% males and 59.6% females, participated in this study. Participants' ages ranged from 48 to 99 years, with a mean (*SD*) of 64.798 (8.815). All participants in this study had varying levels of education, ranging from 0 to 22 years, with a mean (*SD*) of 9.029 (4.937). GDS measurements were obtained through brief screenings conducted during clinical interviews with participants, assisted by caregivers. Subsequently, the researchers categorized the participants into five characteristics based on GDS criteria.

Table 1. Demographic Characteristics of Participants Based on GDS Criteria

Characteristic	Frequency (%)
No Cognitive Decline	33 (15.8%)
Age Associated Memory Impairment	119 (57.2%)
Mild Cognitive Impairment	40 (19.2%)
Mild Dementia	13 (6.2%)
Moderate Dementia	3 (1.4%)

Clock Drawing Test

We utilized the 18-point scoring system of the *Clock Drawing Test* developed by Babins et al. (2008). This version represents an updated version of the scoring system by Rouleau et al., (1992), incorporating five items with combined response options (binary and ternary). Participants were instructed to draw a clock face along with the numbers, and the hands should point to 11:10. In this study, we modified the scoring technique into a continuous rating scale ranging from 1 to 3, with response options: 1 = poor, 2 = fair, 3 = good. The modified 18-point scoring technique in this study did not change either the content of the items or the instruction. The researchers transformed all of the sub-items from items 3 and 4 into individual items, standardizing the response options into a three-point rating scale. Consequently, the modified CDT using an 18-point scoring technique in this study comprised 15 items with equal response options.

Global Deterioration Scale

The *Global Deterioration Scale* (GDS) comprises a set of measurement tools developed by Reisberg et al. (1982), and designed for the rapid and efficient screening of patients suspected to have dementia. Additionally, this test has proven valuable for evaluating and monitoring the progression of patients diagnosed with Alzheimer's Disease (Eisdorfer et al., 1992). According to Choi et al. (2016), GDS is not a perfect measurement tool for further evaluating cognitive decline in patients, as compared to MMSE or CDR. However, in this study, the researchers employed the GDS as a screening tool to classify participants' cognitive decline status.

Many Facets Rasch Measurement

The data were analyzed using the many-facet Rasch measurement. the origins of this model can be traced back to Verlhust's differential equation, stemming from the Rasch model (Bock, 1997). The many-facets Rasch model is a transformation logistics-based equation, slightly distinct from Rasch measurement. This method considers facets as factors that may influence the evaluation process, encompassing various elements such as raters, places, or even task methods (Linacre, 1994). We employed the model represented by Equation 1, wherein θ_n represents the examinee's ability, β_i describes the category difficulty of the test, α_j is the rater severity, and τ_k is the category coefficient (Eckes, 2014). For instance, if a rater is included as a facet, one of the main advantages of this model is its ability to identify bias facet in logit measurement.

Equation 1. Many Facet Rach Model (3 Facets)

$$In\left[\frac{p_{nijk}}{p_{nijk-1}}\right] = \theta_n - \beta_i - \alpha_j - \tau_k$$

We utilized FACET version 3.86.0 for data analysis. This software employs the unconditional joint maximum likelihood (UCON) approach, chosen for its independence, flexibility in handling missing data, and ability to analyze extensive datasets (Linacre, 2011). The iterative process in this analysis will adhere to convergence criteria standards, with a PROX iteration (< 0.5), while JMLE iteration will follow the standards by stopping at (≤ 0.001) and a maximum residual score (≤ .1) (Linacre, 2011).

Procedure

The procedure in this study was divided into two phases: data collection and rater assessment. The researchers involved three psychologists and research assistants throughout the process. After ensuring that all participants comprehended the informed consent, we provided them with blank A4 sheets of paper and writing tools. Subsequently, brief interviews were conducted with caregivers and participants to complete the GDS. Following this, participants were remotely instructed through digital devices, with the instructions displayed on the screen. The researchers then duplicated the participants' work, which was later subjected to the assessment phase by three raters, each conducting their evaluation independently.

RESULTS

Summary Rasch Statistics

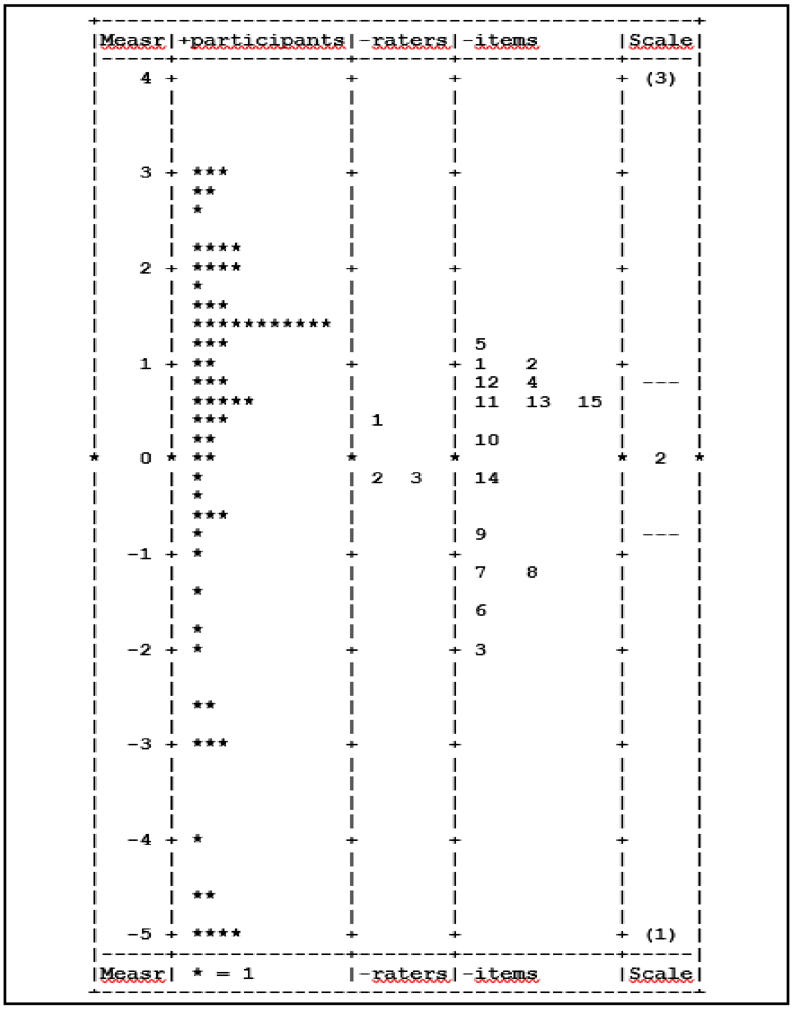
A total of 3 raters x 15 items x 208 participants were involved in this study, resulting in 9,360 response sequences. Table 2 shows the summary statistics of the Rasch measurement. The global chi-square value of this test was $\chi 2$ (*d.f*) = 10,247.22553 (3.798) with a significance level of 0.6520. The chi-square analysis for all facets yielded significant results ($p \geq 0.00$), and the degree of freedom (*d.f*) was set as N-1, where N represents the total number of observations for each facet. Based on the unidimensionality test, this measurement tool exhibits a variance explained by the measure of 59.67%. In contrast to the Cronbach α reliability test, the separation and strata reliability measures indicate the length of the test and data variations for difficulty levels. This test particularly demonstrated high score separation (5.80), showcasing its capability to distinguish participants' abilities effectively. In some cases where an instrument fails to meet this criteria (< 2), proposed by Linacre (2011) additional relevant items may be necessary to extend the test. Moreover, with item separation (19.01), this test had a sufficient number of participants with various abilities.

Table 2. Rasch Summary Statistics

Statistics	Participants	Raters	Items
M (measure)	0.95	0.00	0.00
S.D. (measure)	1.93	0.36	1.21
M (S.E.)	0.35	0.03	0.08
Adj. True *SD*	1.87	0.36	1.21
$\chi 2$	3280.3	225.1	1976.2
df	207	2	14
Strata	5.80	14.51	19.01
Separation	0.94	0.99	0.99

The Wright map (Figure 1) consisted of all facets of this study (participants, raters, and items) in logit rulers with the same linearity. This visualization enables us to analyze the relationship between facets. Briefly, the majority of our participants are located on the upper side of the map, whereas the items tend to cluster in the middle area of the map.

Figure 1. Wright Map of All Facets



```
+----------------------------------------------+
|Measr|+participants|-raters|-items      |Scale|
|-----+-------------+-------+------------+-----|
|   4 +             +       +            + (3) |
|     |             |       |            |     |
|     |             |       |            |     |
|     |             |       |            |     |
|     |             |       |            |     |
|   3 + ***         +       +            +     |
|     | **          |       |            |     |
|     | *           |       |            |     |
|     |             |       |            |     |
|     | ****        |       |            |     |
|   2 + ****        +       +            +     |
|     | *           |       |            |     |
|     | ***         |       |            |     |
|     | ************|       |            |     |
|     | ***         |       | 5          |     |
|   1 + **          +       + 1   2      +     |
|     | ***         |       | 12  4      | --- |
|     | *****       |       | 11  13  15 |     |
|     | ***         | 1     |            |     |
|     | **          |       | 10         |     |
*   0 * **          *       *            *  2  *
|     | *           | 2  3  | 14         |     |
|     | *           |       |            |     |
|     | ***         |       |            |     |
|     | *           |       | 9          | --- |
|  -1 + *           +       +            +     |
|     |             |       | 7   8      |     |
|     | *           |       |            |     |
|     |             |       | 6          |     |
|     | *           |       |            |     |
|  -2 + *           +       + 3          +     |
|     |             |       |            |     |
|     |             |       |            |     |
|     | **          |       |            |     |
|     |             |       |            |     |
|  -3 + ***         +       +            +     |
|     |             |       |            |     |
|     |             |       |            |     |
|     |             |       |            |     |
|     |             |       |            |     |
|  -4 + *           +       +            +     |
|     |             |       |            |     |
|     |             |       |            |     |
|     | **          |       |            |     |
|     |             |       |            |     |
|  -5 + ****        +       +            + (1) |
|-----+-------------+-------+------------+-----|
|Measr| * = 1       |-raters|-items      |Scale|
+----------------------------------------------+
```

All raters in this study appear to have similar logit values, without many discrepancies among them. Further analysis of this map will be statistically presented in the next sections. Overall, from this figure the modified CDT 18-point items demonstrate the capability to cover participants at a moderate level of difficulty.

Clock Drawing Test Psychometrics Properties

All items in this study adhered to the standard fit test for Rasch modeling. However, researchers discovered that a few items exhibited poor infit-outfit statistics, exceeding the optimal threshold. Bond and Fox (2015) recommended both infit and outfit for MNSQ thresholds ideally should fall within the range of (0.5 - 1.5), while ZSTD should be within

Table 3. Items Analysis

Code	Item	*M*	Infit MS	Infit *Z*	Outfit MS	Outfit *Z*
5	Number spacing equal (1,2,4,5,7,8,10,11)	1.66	0.68	-7.3	0.93	-0.49
4	Number spacing equal (3,6,9, 12)	0.97	0.86	-2.81	1.04	0.36
2	Center	0.95	0.84	-3.32	1.02	0.23
1	Contour integrity of the clock face	0.9	0.8	-4.09	2.62	9
12	Size difference of the hands is respected (minute and longer)	0.88	1.35	6.11	1.23	2.04
11	Minute hand is towards correct number	0.68	1.21	3.68	1.01	0.09
10	Hour hand is towards correct number	0.53	1.2	3.29	1.05	0.48
15	Gestalt	0.37	0.55	-9	0.54	-5.61
13	Arrows are drawn	0.2	1.1	1.53	0.96	-0.35
14	Hands are joined or within 12 mm (1/2") of joining	-0.01	0.86	-2.25	0.83	-1.73
9	Clock has two recognizable hands	-0.47	0.98	-0.27	0.79	-1.97
7	No missing or added numbers	-0.62	1.48	5.32	1.17	1.4
8	Numbers clockwise and correct sequence	-1.42	1.16	1.51	0.7	-1.94
6	Number inside circle	-1.92	1.92	6.25	2.2	4.25
3	Numbers all the same (Roman/Arabic)	-2.68	1.31	2	0.93	-0.13

the range of +2 to – 2. Item number 1 (infit Z = -4.92, outfit MS = 2.62, outfit Z = 9), 15 (infit Z = -9, outfit Z = -5.61), and 6 (infit MS = 1.92, infit Z = 6.25, outfit MS = 2.2, outfit Z = 4.25) were identified as the most unfit items of this test, as they exceeded the recommended thresholds. In contrast, item numbers 3, 8, and 13 demonstrated good fit with the unidimensional model.

The specific standard error of this test ranged from .06 - .15, with point-measure correlation ranging from 0.58 to 0.72. According to Linacre (2011), a point-measure correlation of ≥ .4 is recommended as the ideal standard for discriminating between participants with high and low abilities. Specifically, item numbers 15, 8, and 9 had the highest point-measure correlation in this test, with values of 0.75 (15), 0.74 (8), and 0.73 (9) respectively.

Item numbers 5 (1.66), 4 (0.97), and 2 (0.95) posed the greatest challenges in this test, suggesting that many participants struggled to meet the assessment standards set by the three raters for these items. Conversely, researchers found that item numbers 8 (-1.42), 6 (-1.92), and 3 (-2.68) were the easiest for participants.

Table 4. Rating Scale Analysis

Response	Quality Control			Rasch-Andrich Thresholds		Expectation	
	Avg. Meas.	Exp. Meas.	Outfit MNSQ	Measure	*S.E.*	Cat.	-0.5
1 (poor)	-1.17	-1.30	1.6		-	-1.73	-
2 (fair)	0.38	0.60	0.8	-0.42	0.04	0.00	-0.97
3 (good)	2.13	2.08	1.1	0.42	0.03	1.74	0.99

As researchers modified the scoring system into Likert scale, the validity test for the rating scale demonstrated that all the responses in this test were normally functioning. Participants or raters in this study did not experience confusion in recognizing the response. Based on Table 4, the Andrich thresholds and category expectations respectively exhibited consistent steps: none, -0.42, 0.42, and -1.73, 0.00, 1.74.

Rater Analysis

As part of the evaluation process, researchers also conducted analysis for the raters. All raters in this study exhibited a standard error ranging from 0.03 to 0.04 and a point-measure correlation ranging from 0.68 to 0.70, with both infit and outfit ZSTD values failing to meet the fit model criteria. Rater 1 demonstrated the most precise fit statistics, while Rater 2 appeared to be less fitted or underfit, and Rater 3 exhibited a higher degree of overfit compared to others. According to Table 5, researchers observed that rater 3 was the most stringent and selective in assigning scores, while rater 1 was more lenient compared to others in evaluating participants' test results. The inter-rater agreement value was 74%, indicating that these raters had a 26% disagreement rate with each other.

Table 5. Rater Analysis

Raters	Measure	Infit MS	Infit *Z*	Outfit MS	Outfit *Z*
Rater 1	-0.37	1.04	1.23	1.19	2.57
Rater 2	0.02	1.06	2.29	1.26	3.85
Rater 3	0.35	0.93	-2.67	0.94	-0.91

Prediction Toward GDS Criteria

Additionally, researchers conducted ROC analysis to assess the predictive ability of the modified 18-point CDT logit scores concerning the *Global Deterioration Scale* (GDS) score. Table 6 indicates that this test version was highly optimal in predicting cognitive impairment and mild dementia among participants based on GDS criteria, with sensitivity and specificity ranging from 0.80 to 0.89 and 0.87 to 0.92, respectively.

Table 6. ROC analysis

Category	Age Associated Memory Impairment	Mild Cognitive Impairment	Mild Dementia	Moderate Dementia
Significance	0.003	< 0.001	< 0.001	0.919
Accuracy	0.783	0.808	0.891	0.917
AUC	0.701	0.879	0.929	0.525
Sensitivity	0.992	0.800	0.970	0.000
Specificity	0.030	0.818	0.692	1.000

DISCUSSION

We employed an 18-point scoring system developed by Babins et al. (2008) with modifications to the response using a continuous rating scale. In this study, Rasch measurement explained 59.67% of the variance. However, Linacre (2011) emphasized the necessity of conducting Rasch residual principal component analysis to investigate new dimension probability. Additionally, the researchers did not solely rely on this indication to assess the validity of this test, as fit statistics are also valuable in confirming unidimensionality. Item numbers 1, 15, and 6 were identified as the most misfit items of this test. Furthermore, participants with unexpected responses were more likely to play a significant role in the item's outfit measurement, as it is more sensitive to outliers and extreme values (Brentari & Golia, 2007; Engelhard, 1992). Nevertheless, this study did not eliminate any items. According to Abdaziz et al. (2014), the elimination of items requires other fulfilled conditions, such as point-measure correlation (≥ 0.4), in addition to infit and outfit tests.

The most challenging item to reach agreement on was item 5, which required participants to estimate the distance between numbers on the clock (1, 2, 4, 5, 7, 8, 10, 11) evenly. According to Tranel et al. (2008), participants who failed in tasks involved spatial ability were found to have lesions in their parietal lobe, specifically in the supramarginal gyrus. Talwar et al. (2019) also found that decreased activity in the parietal and bilateral occipital lobes is associated with poor performance in drawing a clock on the CDT in general. The findings of this study revealed that the item measuring the overall gestalt of a clock (item 15) was the most reliable for discriminating between high and low-ability participants, with a point-measure correlation of 0.75. This finding contradicted those of Bennasar et al. (2013), who found that the length of the clock's hands (item 12) was the most reliable criterion for differentiating between participants with cognitive issues and those without the spatial ability to analyze and plan while drawing a clock.

On this modified CDT-18 points, we modified the scoring system into a rating scale model. The previous scoring system included combined responses ranging from 0 to 1 and 0 to 2. In some studies outside the neurocognitive field, a continuous scale has been proven to yield robust and valid results compared to the binary response (Markon et al., 2011; Munson et al., 2017). Additionally, some items in this test might be better suited to binary responses, such as number 7 (“no missing or added number”), where the answer is either ‘yes’ or ‘no’. Nevertheless, we researchers aimed to elicit an in-between response by introducing the middle option for participants to choose. This addition is intended to facilitate a gradient of responses to a question as simple as “How many numbers are missing/added compared to others?” Furthermore, the rating scale analysis performed well, indicating that this scoring system did not cause any confusion among the raters. Previous studies have affirmed that a simpler scoring system is preferable because a rigid system may introduce limitations to the test's ability to capture subtle errors (Borson et al., 1999). Additionally, it has been found to decrease the test’s ability to perform rapid screening (Mainland et al., 2014). Therefore, maintaining the simplicity and ease of execution of this test, without compromising its quality, is essential.

Linacre (2022) emphasized the importance of declaring the rater type in an evaluation process, as this is associated with the assumption of rater agreement. In this study, it was assumed that raters followed screening processes with rigid agreement. However, the inter-rater agreement results in this study (74%) did not meet the researchers' expectations, as the anticipated agreement was less than 90%. Notably, Rater 3 was identified as the most stringent examiner, exhibiting a 0.33 logit difference from Rater 2. While these discrepancies may be relatively small, for practical purposes, this information is valuable for evaluating raters' behavior during the screening process. For instance, it provides insights into whether some raters display extremely high strictness or leniency.

Lastly, in predicting the test using participants' logit scores from CDT, we observed that this test demonstrated better accuracy, yielding significant results in correctly identifying participants with normal cognitive function and those with mild cognitive impairment and mild dementia based on GDS criteria. On the other hand, Chiu et al. (2008) discovered that the CDT-18 points scoring method was optimum for discriminating between normal participants and those with mild cognitive dementia. Despite its inability to provide specific information about participants' cognitive decline, it proved effective in swiftly distinguishing between normal and suspected patients with cognitive decline as a prior information before further assessment.

We acknowledge the limitations of this study, as it identified 603 empirical bias terms in the rater-participant interaction out of a total of 9,045 terms. Additionally, the sample size of raters in this study was relatively small. Therefore, it is recommended that this study be replicated with a larger number of raters, and efforts be made to reduce empirical bias in the rater-participant interaction. Furthermore, this study presents notable implications that encourage researchers to apply MFRM to other screening tests. It is hoped that this will provide a more robust understanding of test construction and its application in clinical settings.

Conclusion

The *Clock Drawing Test-18-point* revised version demonstrated overall decent validity and reliability test based on the Many-Facets Rasch model perspective. However, it exhibited some misfits in specific items, namely, items 1, 15, and 6. While items 5, 4, and 2 proved to be highly challenging, items 8, 6, and 3 were more accessible. The researchers also identified items 15, 8, and 9 as the best items for distinguishing between lower and higher-ability participants. The strata reliability indicated that this test effectively covered the diversity of participants' abilities in this study. According to the ROC analysis, the test is accurate as a screener to distinguish between healthy participants and those with mild cognitive impairment and mild dementia based on GDS criteria as preliminary data before the actual diagnostic test. However, all raters in this test exhibited inter-rater agreement below the reference standard (< 90%).

Funding: This research was supported by the Research and Community Service Institution of University of Surabaya, Indonesia.

Research Interest: There was no conflict of interest during our study.

REFERENCES

Abdaziz, A., Jusoh, M.S., Amlus, M.H., & Tuan Salwani, S. (2014). Construct validity: a Rasch measurement model approaches. *Journal of Applied Science and Agriculture*, *9*(11), 104–108.

Agrell, B., & Dehlin, O. (1998). The clock-drawing test. *Age and Ageing*, *27*, 399–403.

Alexandrowicz, R. W., Jahn, R., & Wancata, J. (2018). Assessing the dimensionality of the CES-D using multi-dimensional multi-level Rasch models. *PLoS ONE*, *13*(5), 1–19. https://doi.org/10.1371/journal.pone.0197908

Babins, L., Slater, M. E., Whitehead, V., & Chertkow, H. (2008). Can an 18-point clock-drawing scoring system predict dementia in elderly individuals with mild cognitive impairment? *Journal of Clinical and Experimental Neuropsychology*, *30*(2), 173–186. https://doi.org/10.1080/13803390701336411

Batchelder, L., Fox, D., Potter, C. M., Peters, M., Jones, K., Forder, J. E., & Fitzpatrick, R. (2020). Rasch analysis of the long-term conditions questionnaire (LTCQ) and development of a short-form (LTCQ-8). *Health and Quality of Life Outcomes*, *18*(1), 1–12. https://doi.org/10.1186/s12955-020-01626-3

Bennasar, M., Setchi, R., Bayer, A., & Hicks, Y. (2013). Feature selection based on information theory in the clock drawing test. *Procedia Computer Science*, *22*(April 2015), 902–911. https://doi.org/10.1016/j.procs.2013.09.173

Bock, R. D. (1997). Theory: A brief history of item response theory. *Educational Measurement: Issues and Practice*, *Winter*, 21–33. http://brainimaging.waisman.wisc.edu/~perlman/Kristin/bock-educ-meas-1997.pdf

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch Model* (Routledge (ed.); III).

Boone, W. J. (2016). Rasch analysis for instrument development: Why,when,and how? *CBE Life Sciences Education*, *15*(4), 1–7. https://doi.org/10.1187/cbe.16-04-0148

Borson, S., Brush, M., Gil, E., Scanlan, J., Vitaliano, P., Chen, J., Cashman, J., Sta Maria, M. M., Barnhart, R., & Roques, J. (1999). The clock drawing test: Utility for dementia detection in multiethnic elders. *Journals of Gerontology -*

Series A Biological Sciences and Medical Sciences, *54*(11), 534–540. https://doi.org/10.1093/gerona/54.11.M534

Brentari, E., & Golia, S. (2007). Unidimensionality in the Rasch model: how to detect and interpret. *Statistica*, *67*(3), 253–261. https://doi.org/10.6092/issn.1973-2201/3508

Cahn-Weiner, D. A., Williams, K., Grace, J., Tremont, G., Westervelt, H., & Stern, R. A. (2003). Discrimination of Dementia with Lewy bodies from Alzheimer disease and Parkinson disease using the Clock Drawing Test. *Cognitive and Behavioral Neurology*, *16*(2), 85–92. https://doi.org/10.1097/00146965-200306000-00001

Cahn, D. A., Salmon, D. P., Monsch, A. U., Butters, N., Wiederholt, W. C., Corey-Bloom, J., & Barrett-Connor, E. (1996). Screening for dementia of the Alzheimer type in the community: The utility of the Clock Drawing Test. *Archives of Clinical Neuropsychology*, *11*(6), 529–539. https://doi.org/10.1016/0887-6177(95)00041-0

Camargo, F. R., & Henson, B. (2015). Beyond usability: designing for consumers' product experience using the Rasch model. *Journal of Engineering Design*, *26*(4–6), 121–139. https://doi.org/10.1080/09544828.2015.1034254

Carneiro, D. F. R. (2015). Validation studies of the Clock Drawing Test in mild cognitive impairment. *European Journal of Neurology*, *22*, 495. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed13&NEWS=N&AN=71933926

Chiu, Y. C., Li, C. L., Lin, K. N., Chiu, Y. F., & Liu, H. C. (2008). Sensitivity and specificity of the clock drawing test, incorporating Rouleau scoring system, as a screening instrument for questionable and mild dementia: Scale development. *International Journal of Nursing Studies*, *45*(1), 75–84. https://doi.org/10.1016/j.ijnurstu.2006.09.005

Choi, Y. J., Won, C. W., Kim, S., Choi, H. R., Kim, B. S., Jeon, S. Y., Kim, S. Y., & Park, K. W. (2016). Five items differentiate mild to severe dementia from normal to minimal cognitive impairment - Using the Global Deterioration Scale. *Journal of Clinical Gerontology and Geriatrics*, *7*(1), 1–5. https://doi.org/10.1016/j.jcgg.2015.05.004

Chong, J., Mokshein, S. E., & Mustapha, R. (2022). Applying the Rasch Rating Scale Model (RSM) to investigate the rating scales function in survey research instrument. *Cakrawala Pendidikan*, *41*(1), 97–111. https://doi.org/10.21831/cp.v41i1.39130

Cullen, B., O'Neill, B., Evans, J. J., Coen, R. F., & Lawlor, B. A. (2007). A review of screening tests for cognitive impairment. *Journal of Neurology, Neurosurgery and Psychiatry*, *78*(8), 790–799. https://doi.org/10.1136/jnnp.2006.095414

de Ayala, R. J. (2009). *The Theory and Practice of Item Response Theory*. Guilford Press.

Eckes, T. (2014). *Many-facet Rasch measurement*. 1–52.

Eisdorfer, C., Cohen, D., Paveza, G. J., Ashford, J. W., Luchins, D. J., Gorelick, P. B., Hirschman, R. S., Freels, S. A., Levy, P. S., Semla, T. P., & Shaw, H. A. (1992). An empirical evaluation of the global deterioration scale for staging Alzheimer’s disease. *American Journal of Psychiatry*, *149*(2), 190–194. https://doi.org/10.1176/ajp.149.2.190

Emek-Savaş, D. D., Yerlikaya, D., & Yener, G. G. (2018). Validity, reliability and turkish norm values of the clock drawing test for two different scoring systems. *Turk Noroloji Dergisi*, *24*(2), 143–152. https://doi.org/10.4274/tnd.26504

Engelhard, G. (1992). Applied Measurement in Education The Measurement of Writing Ability With a Many-Faceted Rasch Model. *Applied Measurement in Education*, 171–191. https://doi.org/10.1207/s15324818ame0503

Farlie, M., Johnson, C., Wilkinson, T., & Keating, J. (2021). Refining assessment: Rasch analysis in health professional education and research. *Focus on Health Professional Education: A Multi-Professional Journal*, *22*(2), 88–104. https://doi.org/10.11157/fohpe.v22i2.569

Folstein, M. f., Folstein, S. E., & McHugh, P. R. (1975). A practical method for grading the cognitive state of patients for the clinican. *Journal of Psychiatric Research*, *12*, 189–198. https://doi.org/10.3744/snak.2003.40.2.021

Franchignoni, F., Ferriero, G., Giordano, A., Sartorio, F., Vercelli, S., & Brigatti, E. (2011). Psychometric properties of QuickDASH - A classical test theory and Rasch analysis study. *Manual Therapy*, *16*(2), 177–182. https://doi.org/10.1016/j.math.2010.10.004

Gordon, R. A., Peng, F., Curby, T. W., & Zinsser, K. M. (2021). An introduction to the many-facet Rasch model as a method to improve observational quality measures with an application to measuring the teaching of emotion skills. *Early Childhood Research Quarterly*, *55*, 149–164. https://doi.org/10.1016/j.ecresq.2020.11.005

Han, T. C., & Li, T. H. (2015). Applying the rasch model to construct the shipping industry employability indicators. *Journal of Marine Science and Technology (Taiwan)*, *23*(5), 741–747. https://doi.org/10.6119/JMST-015-0609-2

Hazan, E., Frankenburg, F., Brenkel, M., & Shulman, K. (2018). The test of time: a history of clock drawing. *International Journal of Geriatric Psychiatry*, *33*(1), e22–e30. https://doi.org/10.1002/gps.4731

Huebner, A., & Skar, G. B. (2021). Conditional standard error of measurement: classical test theory, generalizability theory and many facet Rasch measurement with applications to writing assessment. *Practical Assessment, Research and Evaluation*, *26*, 1–20. https://doi.org/10.7275/vzmm-0z68

Ilardi, C. R., Garofalo, E., Chieffi, S., Gamboz, N., La Marra, M., & Iavarone, A. (2020). Daily exposure to digital displays may affect the clock-drawing test: from psychometrics to serendipity. *Neurological Sciences*, *41*(12), 3683–3690. https://doi.org/10.1007/s10072-020-04498-z

Ismail, Z., Rajji, T. K., & Shulman, K. I. (2010). Brief cognitive screening instruments: An update. *International Journal of Geriatric Psychiatry*, *25*(2), 111–120. https://doi.org/10.1002/gps.2306

Kørner, E. A., Lauritzen, L., Nilsson, F. M., Lolk, A., & Christensen, P. (2012). Simple scoring of the Clock-Drawing test for dementia screening. *Danish Medical Journal*, *59*(1).

Lam, L. C. W., Chiu, H. F. K., Ng, K. O., Chan, C., Chan, W. F., Li, S. W., & Wong, M. (1998). Clock-face drawing, reading and setting tests in the screening of dementia in Chinese elderly adults. *Journals of Gerontology - Series B Psychological Sciences and Social Sciences*, *53*(6), 353–357. https://doi.org/10.1093/geronb/53B.6.P353

Linacre, J. M. (1994). *Many-Facet Rasch Measurement*. MESA Press.

Linacre, J. M. (2011). *Winsteps Help for Rasch Analysis*. http://homes.jcu.edu.au/~edtgb/%5Cnpapers3://publication/uuid/D56B724A-62FF-4D00-84E1-ECC888298B70

Linacre, J. M. (2022). *A User's Guide to FACETS*. https://eur-lex.europa.eu/legal-content/PT/TXT/PDF/?uri=CELEX:32016R0679&from=PT%0Ahttp://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CELEX:52012PC0011:pt:NOT

Mainland, B. J., Amodeo, S., & Shulman, K. I. (2014). Multiple clock drawing scoring systems: Simpler is better. *International Journal of Geriatric Psychiatry*, *29*(2), 127–136. https://doi.org/10.1002/gps.3992

Manos, P. J., & Wu, R. (1994). The ten point clock test: A quick screen and grading method for cognitive impairment in medical and surgical patients. *International Journal of Psychiatry in Medicine*, *24*(3), 229–244. https://doi.org/10.2190/5a0f-936p-vg8n-0f5r

Markon, K. E., Chmielewski, M., & Miller, C. J. (2011). The reliability and validity of discrete and continuous measures of psychopathology: A quantitative review. *Psychological Bulletin*, *137*(5), 856–879. https://doi.org/10.1037/a0023678

Mendez, M. F., Ala, T., & Underwood, K. L. (1992). Development of scoring criteria for the Clock Drawing Task in Alzheimer's Disease. *Journal of the American Geriatrics Society*, *40*(11), 1095–1099. https://doi.org/10.1111/j.1532-5415.1992.tb01796.x

Munson, B., Schellinger, S. K., & Edwards, J. (2017). Bias in the perception of phonetic detail in children's speech: A comparison of categorical and continuous rating scales. *Clinical Linguistics and Phonetics*, *31*(1), 56–79. https://doi.org/10.1080/02699206.2016.1233292

Nasreddine, Z. S., Phillips, N. A., Be´dirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J. L., & Chertkow, H. (2005). The Montreal Cognitive Assessment, MoCA: A Brief screening tool for mild cognitive impairment. *Brief Methodological Reports*, *53*, 659–699. https://doi.org/10.1177/0891988716666381

Natanael, Y. (2021). Analisis Rasch model Indonesia Problematic Internet Use Scale (IPIUS). *Persona:Jurnal Psikologi Indonesia*, *10*(1), 167–186. https://doi.org/10.30996/persona.v10i1.4827

Park, K. H., Hong, I., & Park, J. H. (2021). Development and validation of the Yonsei Lifestyle Profile-Satisfaction (YLP-S) using the Rasch measurement model. *Inquiry (United States)*, *58*. https://doi.org/10.1177/00469580211017639

Petrillo, J., Cano, S. J., McLeod, L. D., & Coon, C. D. (2015). Using classical test theory, item response theory, and rasch measurement theory to evaluate patient-reported outcome measures: A comparison of worked examples. *Value in Health*, *18*(1), 25–34. https://doi.org/10.1016/j.jval.2014.10.005

Pinto, E., & Peters, R. (2009). Literature review of the Clock Drawing Test as a tool for cognitive screening. *Dementia and Geriatric Cognitive Disorders*, *27*(3), 201–213. https://doi.org/10.1159/000203344

Reisberg, B., Ferris, S., De Leon, M., & Crook, T. (1982). The Global Detoriation Scale for assessment of primary degenerative dementia. *Gerontologist*, *139*(9), 1136–1139.

Ricci, M., Pigliautile, M., D’Ambrosio, V., Ercolani, S., Bianchini, C., Ruggiero, C., Vanacore, N., & Mecocci, P. (2016). The clock drawing test as a screening tool in mild cognitive impairment and very mild dementia: a new brief method of scoring and normative data in the elderly. *Neurological Sciences*, *37*(6), 867–873. https://doi.org/10.1007/s10072-016-2480-6

Rouleau, I., Salmon, D. P., Butters, N., Kennedy, C., & McGuire, K. (1992). Quantitative and qualitative analyses of clock drawings in Alzheimer’s and Huntington’s disease. *Brain and Cognition*, *18*(1), 70–87. https://doi.org/10.1016/0278-2626(92)90112-Y

Shulman, K. I. (2000). Clock-drawing: Is it the ideal cognitive screening test? *International Journal of Geriatric Psychiatry*, *15*(6), 548–561. https://doi.org/10.1002/1099-1166(200006)15:6<548::AID-GPS242>3.0.CO;2-U

Shulman, K. I., Pushkar Gold, D., Cohen, C. A., & Zucchero, C. A. (1993). Clock-drawing and dementia in the community: A longitudinal study. *International Journal of Geriatric Psychiatry*, *8*(6), 487–496. https://doi.org/10.1002/gps.930080606

Shulman, K. I., Shedletsky, R., & Silver, I. L. (1986). The challenge of time: Clock-drawing and cognitive function in the elderly. *International Journal of Geriatric Psychiatry*, *1*(2), 135–140. https://doi.org/10.1002/gps.930010209

Storey, J. E., Rowland, J. T. J., Basic, D., & Conforti, D. A. (2002). Accuracy of the clock drawing test for detecting dementia in a multicultural sample of elderly Australian patients. *International Psychogeriatrics*, *14*(3), 259–271. https://doi.org/10.1017/S1041610202008463

Sumintono, B. (2018). Rasch Model Measurements as tools in assesment for learning. *173*(Icei 2017), 38–42. https://doi.org/10.2991/icei-17.2018.11

Sunderland, T., Hill, J. L., Mellow, A. M., Lawlor, B. A., Gundersheimer, J., Newhouse, P. A., & Grafman, J. H. (1989). Clock Drawing and Alzheimer’s Disease. *Journal of the American Geriatrics Society*, *37*, 725–729. https://doi.org/10.1111/j.1532-5415.1990.tb03530.x

Tabari, P., & Amini, M. (2021). Educational and psychological support for medical students during the COVID-19 outbreak. *Medical Education*, *55*(1), 125–127. https://doi.org/10.1111/medu.14376

Talwar, N. A., Churchill, N. W., Hird, M. A., Pshonyak, I., Tam, F., Fischer, C. E., Graham, S. J., & Schweizer, T. A. (2019). The neural correlates of the clock-drawing test in healthy aging. *Frontiers in Human Neuroscience*, *13*(February), 1–12. https://doi.org/10.3389/fnhum.2019.00025

Tavakol, M., & Pinner, G. (2019). Using the Many-Facet Rasch Model to analyse and evaluate the quality of objective structured clinical examination: A non-experimental cross-sectional design. *BMJ Open*, *9*(9), 1–9. https://doi.org/10.1136/bmjopen-2019-029208

Tranel, D., Rudrauf, D., Vianna, E. P. M., & Damasio, H. (2008). Does the Clock Drawing Test have focal neuroanatomical correlates? *Neuropsychology*, *22*(5), 553–562. https://doi.org/10.1037/0894-4105.22.5.553

Uto, M. (2021). A multidimensional generalized many-facet Rasch model for rubric-based performance assessment. In *Behaviormetrika* (Vol. 48, Issue 2). Springer Japan. https://doi.org/10.1007/s41237-021-00144-w

Wolf-Klein, G. P., Silverstone, F. A., Levy, A. P., Brod, M. S., & Breuer, J. (1989). Screening for Alzheimer's Disease by Clock Drawing. *Journal of the American Geriatrics Society*, *37*(8), 730–734. https://doi.org/10.1111/j.1532-5415.1989.tb02234.x

Zahirah, A., & Susanto, H. (2021). Aplikasi model Rasch pada adaptasi skala personal fable remaja di Jawa Barat. *Persona:Jurnal Psikologi Indonesia*, *10*(1), 63–80. https://doi.org/10.30996/persona.v10i1.5097