Id: 1080121

Participatory Educational Research (PER)
Vol.9(6), pp. 436-452, November 2022
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.22.147.9.6

Validating the Cognitive Diagnostic Assessment and Assessing Students' Mastery of 'Parallel and Perpendicular Lines' Using the Rasch Model

Huan Chin
School of Educational Studies, Universiti Sains Malaysia, Penang, Malaysia
ORCID: 0000-0003-0991-7299

Cheng Meng Chew *
School of Educational Studies, Universiti Sains Malaysia, Penang, Malaysia
ORCID: 0000-0001-6533-8406

Wun Thiam Yew
School of Educational Studies, Universiti Sains Malaysia, Penang, Malaysia
ORCID: 0000-0002-2714-9636

Muzirah Musa
School of Educational Studies, Universiti Sains Malaysia, Penang, Malaysia
ORCID: 0000-0003-3803-0208

Article history

Received:
28.02.2022

Received in revised form:
24.07.2022

Accepted:
25.10.2022

Key words:
Cognitive diagnostic assessment; elementary school; geometry; parallel and perpendicular lines; validity

'Parallel and Perpendicular Lines' is an important topic that serves as a basis for the learning of a more advanced geometric concept in later years. Yet, this topic is hard to master by the students. To pinpoint students' weaknesses in this topic, this study sought to develop a cognitive diagnostic assessment (CDA) to assess students' mastery of 'Parallel and Perpendicular Lines'. The validation of the CDA and the use of CDA in measuring students' mastery of 'Parallel and Perpendicular Lines' was documented in this article. The content validation involved two subject matter experts, while the pilot test involved 154 Year Four students from Kedah state of Malaysia selected using cluster sampling. The experts' consensus on the relevancy of test items was captured by calculating the content validity index. The psychometric properties of items and reliability of assessment were analysed based on Rasch Measurement Model. The validity of the assessment content was supported with an acceptable content validity index of 1.00 (>.80). The findings of Rasch analysis span across all ranges of abilities level and hence fit students' competence well. With an acceptable person separation index of 1.58 (> 1.50), person separation reliability of .74 (>.70), and KR-20 coefficient of .78 (>.70), the CDA developed is reliable. The findings of assessing students' mastery level highlighted their weaknesses in defining the properties of perpendicular

* Correspondency: cmchew@usm.my

lines and drawing perpendicular lines. The findings of this study would encourage practitioners to utilise it in the mathematics classroom for diagnosing students' weaknesses and hence plan for remedial instruction.

Introduction

'Parallel and Perpendicular Lines' is an important topic included in elementary mathematics syllabi. This topic serves as a basis for learning more advanced geometric concepts such as triangles and quadrilaterals in later years (Mansfield & Happs, 1992; Ulusoy, 2016). Despite the importance of these topics, Clements (2003) found that this topic was difficult to be mastered by the students. Even the students have learnt this topic since elementary school, Ulusoy (2016) reported that a huge number of them still failed to generate correct examples of parallel and perpendicular lines in middle schools. The students also confused the concept of parallel lines with perpendicular lines (Retnawati et al., 2017). Consequently, elementary school and middle school students were reported to have a higher tendency for failing to solve the geometry items compared to other domains in Trends in International Mathematics and Science Study 2019 (Mullis et al., 2020). The persistent errors made by the students in 'Parallel and Perpendicular Lines' could be due to the misconceptions held (Ulusoy, 2019, 2021). For example, Ulusoy (2016) found that students regarded parallelity as verticality. With this misconception, they failed to recognise the slanted parallel line pairs because they are not in a vertical position. Likewise, the students would also fail to state the properties of parallel lines due to the misconceptions held.

To measure the Turkish Grade 7 students' learning of 'Parallel and Perpendicular Lines', Ulusoy (2016, 2022) developed a test with the example generation tasks and the example determination tasks. The example generation tasks consisted of two sections. For the first section, the students were required to draw a pair of parallel lines and a pair of perpendicular lines on the square grid paper. For the second section, the students were required to draw line parallel or perpendicular to the eight lines given. Meanwhile, the example determination task only consists of one section. The students were required to determine whether the 11-line segment pairs shown in the grid paper are perpendicular or parallel. For both tasks, the students are required to provide a written justification for their answer given as follows: "*These sections are … because …"*. Although the test developed by Ulusoy (2016, 2022) has shown to be useful in revealing students' thinking about the parallel and perpendicular lines, it might be less suitable to measure the elementary students' understanding due to their limited language ability in justifying their thoughts.

While classroom assessments play a predominant role in informing instructional practice, students' mastery of parallel and perpendicular lines should be tapped into a set of sub-skills which contribute to the formation of the concept. In other words, the items developed should adhere to the cognitive models illustrating the sequence of related sub-skills acquisition. Yet, past studies (i.e., Marnizam & Ali, 2021; Nortvedt & Buchholtz, 2018) indicate that the current classroom assessments are mainly developed by the teachers based on the curriculum documents. Rather than measuring students' mastery of concept using sub-skill mastery, these assessments used percent score or grade as a unidimensional measure which indicates students' competence on the concept being assessed. Thus, the available formative assessments have limited potential in providing detailed diagnostic information to teachers (Brendefur et al., 2018; Herrera et al., 2012).

To support students' learning of 'Parallel and Perpendicular Lines', it is crucial to pinpoint

students' weaknesses and provide targeted instruction to rectify their academic deficiencies. In fact, determining students' weaknesses might require the decomposition of learning goals into subskills required to solve the mathematical task (Philipp, 2017). With this regard, the cognitive diagnostic assessment (CDA) that emerges from the fusion of cognitive psychology and educational measurement (Alves, 2012) could be appropriate to pinpoint students' weaknesses. Different from other assessments, the development and inference-making of CDA were guided by cognitive models which are commonly illustrated as hierarchically ordered subskills or attributes required to solve the mathematical tasks correctly (Roberts et al., 2014). Thus, the diagnostic information could be linked to students' cognition.

In view of its strength in locating students' cognitive strengths and weaknesses, several CDAs have been developed in the past. Broaddus (2011) developed a CDA with multiple-choice items for the middle school mathematics topic, named 'Slope'. In the elementary school context, Alves (2012) developed a CDA with a mixture of multiple-choice items and open-ended items for diagnosing students' cognitive strengths and weaknesses for 'two-digit numeral subtraction'. Recently, the CDAs with open-ended items (Sia & Lim, 2018) and ordered multiple-choice items (Chin et al, 2021a; Chin et al., 2021b; Chin & Chew, 2022) for the topic of 'Time' have also been developed.

Even though 'Parallel and Perpendicular Lines' has been reported as a difficult topic by Clements (2003), Retnawati et al. (2017) and Ulusoy (2016), there is a paucity of studies on the development of a diagnostic instrument for elementary students on this topic. To fill the research gap, this study was conducted to develop and validate the CDA for the topic of "Parallel and Perpendicular Lines". While the learning of this concept is sequential (Szinger, 2008), the development of CDA based on cognitive models which specify the skill mastery sequence would ensure the inference made could be mapped onto the students' cognition. Hence, the development of CDA on "Parallel and Perpendicular Lines" would support teachers in obtaining rich feedback about students' mastery of 'Parallel and Perpendicular lines' and align the instruction to support students' needs. Rather than using the classical test theory (Alves, 2012; Chin et al., 2021a, 2021b; Chin & Chew, 2022) or the three-parameter logistic model item response theory (Brouddus, 2011), the CDA developed in this study was validated using Rasch Model which could produce the person-free and test-free ability estimates (Stemler & Naples, 2021). In other words, this study was conducted to fill the methodology gaps in the relevant literature. Besides that, this study also demonstrated the use of CDA in assessing students' mastery of 'Parallel and Perpendicular Lines'.

Purpose of the Study

This study aimed to develop a cognitive diagnostic assessment (CDA) for assessing students' mastery of 'Parallel and Perpendicular Lines'. Specifically, the validation of the CDA and the use of CDA in measuring students' mastery of 'Parallel and Perpendicular Lines' was documented in this article. Following this, the research questions addressed in this study are:

(1) To what extent are the items in the CDA relevant to the attribute intended to measure?
(2) To what extent do the items fit the students' competence for 'Parallel and Perpendicular Lines'?
(3) To what extent is the discrimination power of the items appropriate?
(4) To what extent is the CDA reliable?
(5) What is the student's mastery of 'Parallel and Perpendicular Lines'?

Method

Research Design

The study was conducted by adopting a four-step instrument development approach that involved (i) concept identification; (ii) item construction; (iii) validity testing and (iv) reliability testing (Davis, 1996). In this study, concept identification involved specifying the attributes intended to measure. This was followed by item construction. Upon the completion of item construction, validity testing, and reliability testing were conducted through expert review and field testing based on a cross-sectional research design. This is because the validity and reliability of the assessment could be determined based on the data that only needed to be collected at a specific time point.

Participants

The validity and reliability testing involved both subject matter experts and students. The expert review of instrument content involved an experienced mathematics teacher teaching in the national primary school and a mathematics instructional coach from the education district office. The two subject matter experts have a minimum of 10 years of mathematics teaching experience. Thus, they have a sufficient understanding of the mathematics curriculum in primary school.

The field test of the instrument involved 1069 Year Four students with an average age of 10 years old from 51 national primary schools in Kedah, Malaysia. These students had been introduced to various types of angles (i.e., acute angles, obtuse angles, right angles, and straight angles) as the pre-requisite skills before being exposed to the concepts of parallel and perpendicular lines. The sample was selected using two-stage cluster sampling. The sample selection process begins with selecting 51 schools in Kedah state, followed by selecting one class of students as the participant from the selected school. Since the 51 selected schools do not practice class streaming, the sample of this study consisted of participants with various ability levels. The demographic information of the participants is presented in Table 1.

Table 1. Demographic Information of Participants

	Gender		
Ethnicity	**Male**	**Female**	**Total**
Malay	484 (45.28%)	542 (50.70%)	1026 (95.98%)
Indian	11 (1.03%)	28 (2.62%)	39 (3.65%)
Others	1 (0.09%)	3 (0.28%)	4 (0.37%)
Total	496 (46.40%)	573 (53.60%)	1069 (100.00%)

Research Instrument

The data of this study was collected using the CDA developed in this study. The development of CDA begins with analysing the mathematical tasks (Akbay et al., 2018; Tang et al., 2020) in the mathematics textbook to specify the cognitive attributes involved in (i) recognizing parallel and perpendicular lines; and (ii) drawing parallel and perpendicular lines. Then, the researchers constructed the test items to measure the 10 attributes identified. The relevant tasks in the mathematics textbook were adapted into multiple-choice items. To improve the measurement precision, three parallel items were constructed for measuring each attribute. Thus, the CDA consisted of 30 multiple-choice items. The table of specifications of the CDA is shown in Table 2. After the item construction, the test items were translated into the Malay Language by a Malay native speaker with good English proficiency and a

sophisticated understanding of the elementary mathematics curriculum, to match the instruction medium of mathematics lessons in National Primary Schools. Then, the translated test items and corresponding answer keys were added to the Google Form.

Table 2. Table of Specifications

Attributes Intended to Measure	Items
A1: State the properties of parallel line	Q1, Q2, Q3
A2: State the lines with the distance which are always equal each other	Q4, Q5, Q6
A3: State the parallel lines.	Q7, Q8, Q9
A4: State the properties of the perpendicular lines	Q10, Q11, Q12
A5: State the lines which intersect with each other at a right angle.	Q13, Q14, Q15
A6: State the perpendicular lines.	Q16, Q17, Q18
A7: State the procedure of drawing parallel lines using a set square and ruler correctly	Q19, Q20, Q21
A8: Draw the parallel lines using a set square and ruler correctly	Q22, Q23, Q24
A9: State the procedure of drawing perpendicular lines using a set square and ruler correctly	Q25, Q26, Q27
A10: Draw the perpendicular lines using a set square and ruler correctly	Q28, Q29, Q30

Data Collection

The data collection was conducted in two stages. The first stage of the study involved content validation. The two subject matter experts were invited to validate the CDA developed in an online workshop. During the workshop, the table of specifications, the CDA, and the validation form were given to each subject matter through email. Then, clear instructions were given to the subject matter experts to rate the relevancy of the items to the attributes intended to measure on the validation form with a five-point Likert Scale: Rating 1 - Not Relevant; Rating 2 - Less Relevant; Rating 3 - Relevant; Rating 4 - Quite Relevant; and Rating 5 - Very Relevant.

During the second stage of the study, the CDA was administered to the participants by their mathematics teacher through the Google Form prepared by the researchers. The students were given 60 minutes to complete the CDA during their online mathematics lesson. The students were requested to answer the assessment individually without consulting their family members. After the test administration, the researchers extracted students' responses from the database and coded the responses into dichotomous scores using Microsoft Excel 2019 based on the answer key.

Data Analysis

To address Research Question 1, the ratings given by the subject matter experts were collapsed into two categories using Microsoft Excel 2019. Ratings 1 and 2 were categorized as 'irrelevant' and were coded as '0', while Ratings 3, 4, and 5 were categorized as 'relevant' and were coded as '1'. Then, the item-level content validity index (I-CVI) was calculated as the average of dichotomous codes derived from the rating given by each validator on the relevancy of each item. To capture the overall consensus of validators on the relevancy of items constructed on the attributes intended to measure, the scale-level content validity index (S-CVI) was calculated as the average of I-CVI (Polit & Beck, 2006). Then, the S-CVI was interpreted based on the cut score proposed by Polit and Beck (2006).

To address Research Question 2, Rasch analysis was conducted using Winsteps (Linacre, 2012). The data analysis begins with checking the two main assumptions of the Rasch model: (i) unidimensionality of the measured trait and (ii) local independence, by performing the

principal component analysis (PCA) of the Rasch residuals and examining the correlations of the residuals respectively. The items are local independent if the correlations are at most .70 (Linacre, 2012). The measured trait is unidimensional if the following criteria was satisfied:

(1) variance explained by measure is more than 20 percent (Reckace, 1979).
(2) the eigenvalue of the largest secondary dimension is less than 3.00 (Linacre, 2012)
(3) the unexplained variance of the largest dimension is less than 15.00 percent (Fisher, 2007)

Then, the item fit was determined based on the fit statistics such as 'inlier-sensitive or information-weighted fit' (infit) and 'outlier-sensitive fit' (outfit) (Linacre, 2012) computed using Winsteps. The infit statistics are sensitive to discrepancies of responses on the items with difficulty close to students' ability (Tavakol & Dennick, 2013). The items would be flagged if the items were wrongly answered by the students with an ability equivalent to the difficulty measure of the items. Thus, infit reflects the construct validity of the test (Alkhadim et al., 2021). The outfit statistic is sensitive to the discrepancies of responses on the items with difficulty far away from students' ability due to carelessness or guessing. For example, the easy items would be flagged if they are answered wrongly by high-performing students. Likewise, the difficult items would be flagged if they are answered correctly by low-performing students due to guessing. Thus, outfit provides information on items which is potentially affected by factors such as guessing or carelessness.

The fit statistics are commonly presented in the mean of squared residuals (MNSQ) and z-standardised of mean square values (ZSTD) (Bond & Fox, 2007). MNSQ is the sample size-independent indicator that describes the size of discrepancies whereas ZSTD is the sample-size-dependent indicator that describes the significance of the fit. With a large sample size, ZSTD could be ignored because substantive misfits might be small (Linacre, 2002). Thus, the item fit was only evaluated based on the MNSQ of infit and outfit based on the interpretation guideline [0.70 ≤MNSQ≤1.30] proposed by Linacre (2012) for non-high stakes assessment. To provide additional information on the match or mismatch of the item difficulty and students' ability, the Wright Map was generated using Winsteps. Besides that, the item separation and item reliability were analysed based on the cut score (i.e., item separation > 3; item reliability > .90) proposed by Linacre (2012) to provide validity evidence on the item difficulty hierarchy of CDA.

To address Research Question 3, the item discrimination index was calculated using Winsteps. While point-biserial correlation is commonly used as the item discrimination index for CTT, the partial correlation of each item with the total measure (PTCOR) was computed as the item discrimination index for Rasch Model. Then, the test items were categorised into three categories based on the guidelines suggested by Hassan and Hod (2017)]: (i) poor (PTCOR <.15); (ii) fair (.15 ≤ PTCOR < .25); and (iii) good (PTCOR ≥ .25).

To address Research Question 4, Kuder-Richardson 20 (KR-20) coefficient was calculated using Microsoft Excel 2019 to measure the reliability of the CDA which is dichotomously scored. Then, the reliability of the CDA was analysed based on the common rule of thumb (i.e., KR-20 ≥ .70) which is suggested by Thompson (2010). Besides that, person separation and person reliability indices computed using Winsteps provide additional reliability evidence for the CDA developed. According to Fisher (1992), the assessments are considered reliable if the person separation is higher than 1.50 and the person reliability is higher than .70.

To address Research Question 5, descriptive analysis was used to profile students' attribute

mastery after determining the person's ability and item difficulty using WINSTEPS based on the following formula:

$$P(X_m = 1|\theta_n, \delta_i) = \frac{e^{\theta n - \delta_i}}{1+e^{\theta n - \delta_i}} \quad (1)$$

where P(X_m=1 | θ_n,δ_i) is the probability that an examinee n (n= 1, …, N) with ability θ_n to answer item i (i = 1, …, I) with difficulty δ_i correctly (Rasch, 1960). Based on the Rasch model, the probability of the students answering the item correctly is more than .50, if his or her ability estimate is higher than the item difficulty. Since each attribute was measured using three items in this study, the complexity of the attribute is operationalised as median item difficulty in the Rasch model. Following this, the probability of students for mastery of the attribute is more than .50 if the ability estimate is higher than the attribute complexity in the Rasch model. According to Bradshaw (2017), the students are considered to master the attribute if the probability of attribute mastery is more than .50. Thus, the proportion of students who have mastered the attribute is operationalised as the proportion of students with the ability estimated higher than median item difficulty. The proportion of students who have mastered the attribute is computed using Excel based on the item difficulty and person ability measure estimated in Winsteps.

Results

To what extent are the items in the CDA relevant to the attribute intended to measure?

The result of content validation is tabulated in Table 3. As shown in Table 3, most of the items (19 out of 30 items; 63.33 %) were rated as very relevant by the two experts. While all items were rated as relevant (rating 3), quite relevant (rating 4), or very relevant (rating 4), the two experts reached a full consensus on the item's relevancy with an I-CVI of 1.00. In other words, all items in the CDA are relevant to the attribute intended to measure. With S-CVI exceeding the minimum threshold of .80 (Polit & Beck, 2006), the content of the CDA is valid.

Table 3. Item Relevance, Item Fit Statistics, and Item Discrimination Analysis

	Item Relevance					Item Difficulty		Item Fit Statistics		I
	Ratings		Dichotomous Code			Measure	S.E.	Infit	Outfit	F
Item	Expert 1	Expert 2	Expert 1	Expert 2	I-CVI	(Logit)	(Logit)	MNSQ	MNSQ	(
Q1	5	5	1	1	1.00	-1.24	0.11	1.08	1.51	0
Q2	5	5	1	1	1.00	2.79	0.08	1.23	1.66	0
Q3	5	5	1	1	1.00	-0.41	0.09	1.02	1.12	0
Q4	5	5	1	1	1.00	-2.17	0.16	0.99	0.73	0
Q5	5	5	1	1	1.00	-0.96	0.10	1.08	1.08	0
Q6	5	5	1	1	1.00	2.63	0.08	1.10	1.25	0
Q7	5	5	1	1	1.00	-0.87	0.10	1.01	0.86	0
Q8	5	4	1	1	1.00	-0.43	0.09	0.96	0.81	0
Q9	5	5	1	1	1.00	0.17	0.08	0.92	0.83	0
Q10	4	5	1	1	1.00	-0.22	0.08	1.03	0.99	0
Q11	5	5	1	1	1.00	1.49	0.07	1.11	1.13	0
Q12	5	5	1	1	1.00	1.18	0.07	1.12	1.12	0
Q13	4	5	1	1	1.00	-0.17	0.08	0.99	0.87	0
Q14	4	5	1	1	1.00	-0.76	0.10	1.06	1.14	0
Q15	4	5	1	1	1.00	-0.02	0.08	1.07	1.13	0
Q16	5	5	1	1	1.00	-0.23	0.08	0.90	0.72	0
Q17	5	5	1	1	1.00	-0.75	0.10	0.95	0.75	0
Q18	5	5	1	1	1.00	1.11	0.07	1.02	1.00	0
Q19	3	5	1	1	1.00	-0.44	0.09	0.90	0.78	0
Q20	3	5	1	1	1.00	-0.47	0.09	0.88	0.71	0
Q21	3	5	1	1	1.00	-0.31	0.08	0.87	0.69	0
Q22	5	5	1	1	1.00	-0.93	0.10	0.95	0.96	0
Q23	5	5	1	1	1.00	0.10	0.08	0.97	1.02	0
Q24	5	5	1	1	1.00	0.50	0.07	0.93	1.02	0
Q25	3	5	1	1	1.00	-1.43	0.12	0.96	0.84	0
Q26	3	5	1	1	1.00	-0.78	0.10	0.89	0.69	0
Q27	3	5	1	1	1.00	-0.82	0.10	0.91	0.92	0
Q28	5	5	1	1	1.00	0.21	0.08	0.95	0.94	0
Q29	5	5	1	1	1.00	1.63	0.07	0.98	1.01	0
Q30	5	5	1	1	1.00	1.61	0.07	1.05	1.11	0
				Mean	1.00 (S-CVI)	0.00	0.09	1.00	.98	

To what extent do the items fit the students' competence for 'Parallel and Perpendicular Lines'?

Before performing the item analysis, a principal component analysis of Rasch residuals was performed to determine the dimensionality of the data. The result of the analysis is reported in Table 4. In this study, the Rasch model explained 29.30 percent of the empirical data. It almost matches the expected variance explained (26.70%). The largest secondary dimension (i.e., the first contrast of the residual) contributed to 6.50 percent of explained variance. With the size of 2.80 (≈3.00) eigenvalue units, the CDA consisted of three items measuring an alternative contrast. Since the variance explained by measure is more than 20 percent, the eigenvalue of the largest secondary dimension is less than 3.00 and the unexplained variance of the largest dimension is less than 15.00 percent, the CDA developed satisfied the unidimensionality assumption for Rasch analysis (Fisher, 2007; Linacre, 2012; Reckace, 1979).

Table 4. Result of Principal Component Analysis of Rasch Residuals

Standardized Residual variance	Eigenvalue units	Empirical		Modelled
Total raw variance in observations	42.40	100.00%		100.00%
Raw variance explained by measures	12.40	29.30%		26.70%
Raw variance explained by persons	4.70	11.10%		10.10%
Raw variance explained by items	7.70	18.20%		16.60%
Raw unexplained variance (total)	30.00	70.70%	100.00%	73.30%
Unexplained variance in 1st contrast	2.80	6.50%	9.20%	

To check the item dependency assumption, the standardized Rasch residuals of items were correlated with each other. The largest standardized residual correlations were as listed in Table 5. Item pair Q19 and Q20 has the largest residual correlation of .70. This indicates all item pairs are at most .70. Hence, the items in CDA are not locally dependent and the CDA developed satisfied the item dependency assumption of Rasch analysis (Linacre, 2012).

Table 5. The Largest Standardized Residual Correlations between Item Pairs

Item pair	Residual Correlation
Q19 - Q20	.70
Q26 - Q27	.64
Q25 - Q27	.47
Q19 - Q21	.43
Q20 - Q21	.38
Q25 - Q26	.38
Q16 - Q17	.33
Q5 - Q7	.31
Q5 - Q14	.25
Q7 - Q14	.25

To evaluate the fit of each item and students' competence in accordance with the Rasch model, the fit statistics are calculated. As shown in Table 3, the infit MNSQ of the CDA items ranged from 0.87 to 1.23. With the infit MNSQ within the acceptable range of 0.70. to 1.30 (Linacre, 2012), the CDA could provide a valid measurement of students' understanding of parallel and perpendicular lines. The outfit MNSQ of the CDA items \ranged from 0.71 to 1.66. The outfit MNSQ of most of the CDA items falls within the acceptable range of 0.70 to 1.30 (Linacre, 2012), except Item Q1, Item Q2, Item Q21, and Item Q26. With outfit MNSQ exceeding 1.30, the student's responses on items Q1 and Item Q2 underfit the Rasch model (Linacre, 2012). With outfit MNSQ slightly lower than .70, the students' responses on items

Q21 and Item Q26 overfit the Rasch model (Linacre, 2012).

In Rasch analysis, the item difficulty was measured in logits. The smaller the value, the easier the test item. To compare the hierarchy of item difficulty and ability estimates, the two parameters were calibrated on the same scale using Wright Map. As shown in Figure 1, the item difficulty measures spread about 5 logits (range: -2.17 to 2.79), while the person ability measures span about 7 logits (range: -1.71 to 5.36). Since the minimum item difficulty measure (-2.17 logits) is lower than the minimum person measure (-1.71 logits), the CDA developed consists of an item [i.e., Item Q2] which is too easy for the students. In other words, the majority of the CDA items match the abilities estimates. Thus, the item difficulty level of CDA is appropriate, and the item fits well with the student's competence. With an item separation index which is more than 3.00 (i.e., 12.44) and item reliability which is more than .90 (i.e., .99), the CDA developed can be categorised into at least three levels of difficulty based on the student's competence.

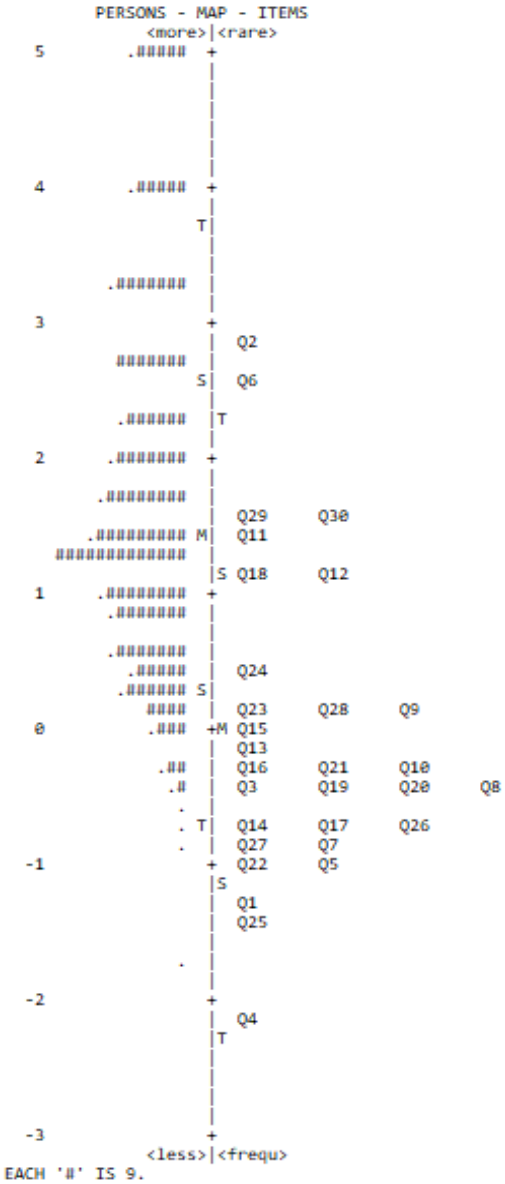

Figure 1. Wright Map

To what extent are the discrimination power of the items appropriate?

The item discrimination index (PTCOR) of each item is tabulated in Table 3. The PTCOR of the items in the CDA developed ranged from .14 to .50. This indicates that the CDA consisted of the items in all three discrimination levels, namely poor (PTCOR <.15), fair (.15 ≤ PTCOR < .25), and good (PTCOR ≥ .25) (Hassan & Hod, 2017). As shown in Table 3, most of the items have good (83.33%) or fair discrimination power (13.33%). There is only one poor discrimination item [i.e., Item Q1] in the CDA with a PTCOR of .14.

To what extent is the CDA reliable?

The reliability of the assessment is examined based on the person separation index, person separation reliability and KR-20. With a person separation index that is more than 1.50 (i.e., 1.68) and a person reliability index that is more than .70 (i.e., .74), the CDA developed is sufficient in separating the participants into two performance levels (Fisher, 1992). The KR-20 coefficient of the CDA is .78. Since the CDA was not developed as a high-stake examination, the reliability coefficient was accepted if the reliability coefficient surpassed the common rule of thumb (i.e, KR-20 ≥ .70) which is suggested by Thompson (2010). In fact, the KR-20 of the CDA developed is higher than the proposed reliability coefficient range for teacher-made assessment (i.e., .50 ≤ KR-20 ≤.60) as reported in the study conducted by Quaigrain and Arhin (2017).

What is the students' mastery of 'Parallel and Perpendicular Lines'?

The CDA was developed to examine students' mastery of five attributes on 'Parallel Lines' and 5 attributes on 'Perpendicular Lines'. The student's mastery of each attribute is as tabulated in Table 6. Most of the attributes have been mastered by at least 90 percent of students. Attribute A2 is the simplest attribute. With an attribute complexity of -0.96, attribute A2 has been mastered by nearly all students (99.91%). The two attributes with a proportion of students' mastery less than 90 percent are Attribute A4 and A10. With an attribute complexity of 1.18, attribute A4 has been mastered by 59.87 percent. With an attribute complexity of 1.61, attribute A10 has been mastered by 40.69 percent.

Notably, these two attributes are related to the content domain on 'Perpendicular Lines'. In other words, all attributes on 'Parallel Lines' has been mastered by at least 90 percent of students, but only three out of five attributes on 'Perpendicular Lines' has been mastered by at least 90 percent of students. The findings indicate that concept of 'Perpendicular Lines' could be more difficult to grasp compared to 'Parallel Lines'.

Table 6. Students' Attribute Mastery

Attribute	**Attribute Complexity [*Median* δ_i for each attribute]**	**Proportion of Students Mastered the Attribute [$P(\theta_n >$ Median δ_i for each attribute)]**
Parallel Lines		
A1	-0.41	99.25%
A2	-0.96	99.91%
A3	-0.43	99.25%
A7	-0.44	99.25%
A8	0.1	90.08%
Perpendicular Lines		
A4	1.18	59.87%
A5	-0.17	96.07%
A6	-0.23	96.07%
A9	-0.82	99.81%
A10	1.61	40.69%

Discussion and Conclusion

In this study, a CDA has been developed for assessing the student's mastery of 'Parallel and Perpendicular Lines'. The findings indicate that CDA developed was supported with convincing content validity evidence. This is supported by Haladyna and Rodriguez (2013) because the use of test specifications to guide the assessment development would ensure the content validity of the assessment. In general, the items developed fit the students' competence well, except for the Item Q1, Item Q2, Item Q21, and Item Q26.

The findings indicated that the student's responses on items Q1 and Item Q2 were under fitted the Rasch model (Linacre, 2012). In other words, the student's responses were unpredicted for these items. This could be due to random guessing or careless mistake made in answering the items (Bond & Fox, 2007). Items Q1 and Q2 are the true/false items related to the properties of parallel lines. With mean students' ability (1.45 logit) higher than item difficulty (-1.24 logit), item Q1 is considered as easy for most of the students. However, the item is flagged as a misfit because there are unpredicted responses which might be due to students' carelessness in answering the item. With item difficulty (2.79 logits) higher than mean students' ability (1.45 logits), item Q2 is considered difficult for most of the students. Most of the students failed to recognise the false statement (i.e., the length of parallel lines is equal). Due to the misconception of parallel lines. However, some of the students with an ability estimated less than 2.79 logits answered the items correctly due to guessing.

The findings indicated that the student's responses on items Q21 and Item Q26 overfit the Rasch model (Linacre, 2012). In other words, the student's responses for items Q21 and Q26 are slightly too predictable. This could be due to a small extent of the local dependence on the items (Baghaei, 2008). In this study, Item Q19, Item Q20, and Item Q21 measured students' understanding of the three steps in drawing a pair of parallel lines. Meanwhile, Item Q25, Item Q26, and Item Q27 measured students' understanding of the three steps in drawing a pair of perpendicular lines. Whilst each item corresponds to each step, standardized residuals of response for item Q21 are correlated with item Q19 and item Q20. Likewise, the standardized residuals of response for item Q26 are correlated with items Q25 and 19(iii). Nonetheless, the local dependence is not significant because the correlations are less than .70 (Linacre, 2012). Thus, the outfit MNSQ (.69) was still very near the low limit of the acceptable range (i.e., .70).

Since the items in CDA, these items spanned across all difficulty ranges they can discriminate the high-performing students and low-performing students well. Nonetheless, there is an item (i.e., item Q1) with a poor discrimination index. With item difficulty (-1.24 logit) lower than the mean person ability (1.45 logit), Item 1(i) could be an easy item for most of the students. Yet, it is flagged as misfit items due to the presence of unpredicted responses rooted in the guessing of students with ability estimates lower than item difficulty. Although the CDA has a very easy item (i.e., Item Q1) with poor discrimination power, Rush et al. (2016) argued that the item could be retained if it is intentionally developed to assess a simple skill. In this study, Item Q1 was developed to assess students' understanding of the basic property of parallel lines. Regardless of ability level, the students could state that the distance between a pair of parallel lines is equal because it is the most basic properties of parallel lines which can be directly observed in any pair of parallel lines.

To ensure the valid use of the CDA, a reliability study was conducted. The findings indicated that the CDA developed was reliable to be used for assessing students' mastery of 'Parallel and Perpendicular Lines'. Even though the CDA developed was not high-stakes assessments,

the CDA developed was more reliable than the teacher-made assessment reported in the study conducted by Quaigrain and Arhin (2017). This could be due to the multiple items used to measure each attribute (Gierl et al., 2009). In this study, each attribute was measured using three items. This would increase the precision of the measurement, and hence increase the reliability of CDA.

To illustrate the use of CDA developed, the student's mastery of 'Parallel and Perpendicular Lines' was also reported in this study. In general, most of the attributes of "Parallel and Perpendicular Lines" have been mastered by the Grade Four students. The finding was supported by the study conducted by Ulusoy (2021). This is because the implementation of the instruction followed the developmental phase of parallel and perpendicular concepts suggested by the Van Hieles' Theory (1986). Notably, this study reveals that the concept of 'Perpendicular Lines' could be more difficult to grasp compared to 'Parallel Lines'. This is in line with the study conducted by Paksu and Bayram (2019), as well as Ulusoy (2016). This might be due to the misconceptions held by the students in the learning of perpendicular lines compared to parallel lines (Ulusoy, 2016). They might treat the non-examples as examples of perpendicular lines. While defining geometric properties involves the generalisation of examples (Park & Kim, 2017; Yao & Manouchehri, 2019), students with misconceptions might make false claims about the properties of perpendicular lines. Without adequate knowledge of perpendicular lines, they might fail to draw perpendicular lines correctly.

In sum, the validity claims of the CDA developed in this study were supported with convincing content validity evidence, as well as satisfactory item psychometric properties and assessment reliability. This would encourage the teachers to utilise it in the mathematics classroom to diagnose students' mastery of attributes on parallel and perpendicular lines. Based on the detailed diagnostic information, the teachers could plan for remedial instruction to support the students' learning of parallel and perpendicular lines.

Notably, it is reported that the students had poor mastery in defining the properties of perpendicular lines and drawing perpendicular lines. According to Park and Kim (2017), defining properties of geometric shape begin with recognizing the visual appearance of the geometric figure, followed by describing the perceptually sensed properties of the figure, determining the local commonalities of the examples, and expressing generalities of the examples. To avoid over-generalisation and under-generalization which might lead to misconceptions, the teachers are encouraged to engage the students in comparing more sophisticated examples of perpendicular lines. As such, the students would have a better understanding of the concept of perpendicular lines.

Limitations and Suggestions

This study was subjected to several limitations. Firstly, the findings of the study could be less representative due to a limited sampling frame. Secondly, the items developed could be subjected to bias in terms of gender and ethnicity. This is because the group differences such as gender and ethnicity might lead to different item functioning (DIF) (Liu et al, 2019; Oliveri et al., 2018). Thus, future studies are suggested to be conducted with a larger sampling frame. Besides that, differential item functioning should be evaluated to avoid ethnicity bias.

References

Akbay, L., Terzi, R., Kaplan, M., & Karaaslan, K.G. (2018). Expert-based attribute identification and validation: A cognitively diagnostic assessment application. *Journal on Mathematics Education*, *9*(1), 103-120. http://dx.doi.org/10.22342/jme.9.1.4341.103-120

Alkhadim, G. S., Cimetta, A. D., Marx, R. W., Cutshaw, C. A., & Yaden, D. B. (2021). Validating the Research-Based Early Math Assessment (REMA) among rural children in Southwest United States. *Studies in Educational Evaluation*, 68, Article 100944. https://doi.org/10.1016/j.stueduc.2020.100944

Alves, C. B. (2012). *Making diagnostic inferences about student performance on the Alberta Education Diagnostic Mathematics Project: An application of the Attribute Hierarchy Method* (Doctoral thesis). Available from ProQuest Dissertations and Theses database. (Publication No. 919011661)

Baghaei, P. (2008). Local dependency and Rasch measures. *Rasch Measurement Transactions*, *21*(3) 1105-1106.

Bond, T. G. & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in the human sciences* (2nd ed.). Lawrence Erlbaum Associates.

Bradshaw, L. (2017). Diagnostic classification models. In A. A. Rupp & J. P. Leighton (Eds.), *The handbook of cognition and assessment: Frameworks, methodologies, and applications* (1st ed., pp. 297–327). Wiley Blackwell.

Brendefur, J. L., Johnson, E. S., Thiede, K. W., Strother, S., & Severson, H. H. (2018). Developing a multi-dimensional early elementary mathematics screener and diagnostic tool: the primary mathematics assessment. *Early Childhood Education Journal*, *46*(2), 153-157. https://doi.org/10.1007/s10643-017-0854-x

Broaddus, A. E. (2011). *An investigation into foundational concepts related to slope: An application of the Attribute Hierarchy Method* (Doctoral thesis). Available from ProQuest Dissertations and Theses Global database. (Publication No. 3487353)

Chin, H., Chew, C. M., & Lim, H. L. (2021a). Development and validation of online cognitive diagnostic assessment with ordered multiple-choice items for 'Multiplication of Time'. *Journal of Computers in Education*, *8*(2), 289-316. https://doi.org/10.1007/s40692-020-00180-7

Chin, H., Chew, C. M., Lim, H. L., & Thien, L. M. (2021b). Development and validation of a cognitive diagnostic assessment with ordered multiple-choice items for Addition of Time. *International Journal of Science and Mathematics Education*. [Advance Online Publication] http://doi.org/10.1007/s10763-021-10170-5

Chin, H. & Chew, C.M. (2022). Cognitive diagnostic assessment with ordered multiple-choice items for word problems involving 'time'. *Current Psychology*. [Advance Online Publication]. https://doi.org/10.1007/s12144-022-02965-8

Clements, D. H. (2003). Teaching and learning geometry. In J. Kilpatrick, W. G. Martin, & D. Schifter (Eds.), *A research companion to principles and standards for school mathematics* (pp. 151-178). National Council of Teachers of Mathematics.

Davis, A. E. (1996). Instrument development: getting started. *Journal of Neuroscience Nursing*, *28*(3), 204-208.

Fisher, W. P. (1992). Reliability, separation, strata statistics. *Rasch Measurement Transactions*, *6*(3), 238.

Fisher, W. P. (2007). Rating scale instrument quality criteria. *Rasch Measurement Transactions*, *21*(1), 1095.

Gay, L. R., Mills, G. E., & Airasian, P. W. (2012). *Educational research: Competencies for analysis and applications* (10th ed.). Merrill.

Gierl, M. J., Cui, Y., & Zhou, J. (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement, 46*(3), 293–313. https://doi.org/10.1111/j.1745-3984.2009.00082.x

Haladyna, T. M., & Rodriguez, M. C. (2013). *Developing and validating test items*. Routledge.

Hassan, S., & Hod, R. (2017). Use of item analysis to improve the quality of single best answer multiple choice question in summative assessment of undergraduate medical students in Malaysia. *Education in Medicine Journal, 9*(3), 33-43. https://doi.org/10.21315/eimj2017.9.3.4

Herrera, S. G., Murry, K. G., & Cabral, R. M. (2012). *Assessment accommodations for classroom teachers of culturally and linguistically diverse students*. Pearson Higher Education.

Linacre, J. M. (2002) What do infit and outfit, mean-square and standardized mean? *Rasch Measurement Transactions*, *16*(2), 878.

Linacre, J. M. (2012). *A user guide to Winsteps Ministep Rasch model computer programs: Program manual 3.75.0.* Retrieved from http://www.winsteps.com/a/winstepsmanual.pdf

Liu, Y., Yin, H., Xin, T., Shao, L., & Yuan, L. (2019). A comparison of differential item functioning detection methods in cognitive diagnostic models. *Frontiers in Psychology*, *10*, Article 1137. https://doi.org/10.3389/fpsyg.2019.01137

Mansfield, H. M. & Happs, J. C. (1992). Using grade eight students' existing knowledge to teach about parallel lines. *School Science and Mathematics, 92*(8), 450-454. https://doi.org/10.1111/j.1949-8594.1992.tb15628.x

Marnizam, F. I., & Ali, S. R. (2021). Penilaian Pelaksanaan Pentaksiran Bilik Darjah (PBD) dalam Kalangan Guru Matematik Sekolah Rendah [Evaluation of The Implementation of Classroom Assessment (PBD) Among Primary School Mathematics Teachers]. *Jurnal Pendidikan Sains Dan Matematik Malaysia*, *11*(2), 81-94. https://doi.org/10.37134/jpsmm.vol11.2.7.2021

Mullis, I. V. S., Martin, M. O., Foy, P., Kelly, D. L., & Fishbein, B. (2020). *TIMSS 2019 International Results in Mathematics and Science*. Retrieved from Boston College, TIMSS & PIRLS International Study Center website: https://timssandpirls.bc.edu/timss2019/international-results/

Nortvedt, G. A., & Buchholtz, N. (2018). Assessment in mathematics education: Responding to issues regarding methodology, policy, and equity. *ZDM*, *50*(4), 555-570. https://doi.org/10.1007/s11858-018-0963-z

Oliveri, M. E., Lawless, R., Robin, F., & Bridgeman, B. (2018). An exploratory analysis of differential item functioning and its possible sources in a higher education admissions context. *Applied Measurement in Education*, *31*(1), 1-16. https://doi.org/10.1080/08957347.2017.1391258

Paksu, A. D., & Bayram, G. (2019). Altıncı sınıf öğrencilerinin paralel ve dik doğru/doğru parçalarını belirleme ve çizme durumları [The sixth-grade students' identification and drawings of parallel and perpendicular line/line segments]. *Gazi University Journal of Gazi Educational Faculty*, *39*(1), 115-145. https://doi.org/10.17152/gefad.346360

Park, J., & Kim, D. W. (2017). How can students generalize examples? Focusing on the generalizing geometric properties. *Eurasia Journal of Mathematics, Science and Technology Education*, *13*(7), 3771-3800. https://doi.org/10.12973/eurasia.2017.00758a

Philipp, K. (2018). Diagnostic competences of mathematics teachers with a view to processes and knowledge resources. In T. Leuders T., K. Philipp, & J. Leuders (Eds.*)*,

Diagnostic competence of mathematics teachers. Mathematics teacher education (vol 11, pp. 109-127). Springer.

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health, 29*(5), 489–497. https://doi.org/10.1002/nur.20147

Quaigrain, K., & Arhin, A. K. (2017). Using reliability and item analysis to evaluate a teacher-developed test in educational measurement and evaluation. *Cogent Education, 4*(1), 1–11. https://doi.org/10.1080/2331186X.2017.1301013

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Danmarks Paedogogiske Institute.

Retnawati, H., Kartowagiran, B., Arlinwibowo, J., & Sulistyaningsih, E. (2017). Why are the mathematics national examination items difficult and what is teachers' strategy to overcome it?. *International Journal of Instruction*, *10*(3), 257-276. https://doi.org/10.12973/iji.2017.10317a

Roberts, M. R., Alves, C. B., Chu, M. W., Thompson, M., Bahry, L. M., & Gotzmann, A. (2014). Testing expert based versus student based cognitive models for a Grade 3 diagnostic mathematics assessment. *Applied Measurement in Education, 27*(3), 173–195. https://doi.org/10.1080/08957347.2014.905787

Rush, B. R., Rankin, D. C., & White, B. J. (2016). The impact of item-writing flaws and item complexity on examination item difficulty and discrimination value. *BMC Medical Education*, *16*(1), 1-10. https://doi.org/10.1186/s12909-016-0773-3

Sia, C. J. L., & Lim, C. S. (2018). Cognitive diagnostic assessment: An alternative mode of assessment for learning. In D. R. Thompson, M. Burton, A. Cusi, & D. Wright (Eds.), *Classroom assessment in mathematics* (pp. 123–137). Springer

Stemler, S. E., & Naples, A. (2021). Rasch Measurement v. Item Response Theory: Knowing when to cross the line. *Practical Assessment, Research, and Evaluation*, *26*(1), Article 11. https://doi.org/10.7275/v2gd-4441

Szinger, I. S. (2008). The evolvement of geometrical concepts in lower primary mathematics. *Annales Mathematicae et Informaticae*, *35*, 173-188. http://publikacio.uni-eszterhazy.hu/id/eprint/3111

Tang, W. L., Tsai, J. T., & Huang, C. Y. (2020). Inheritance coding with Gagné-based learning hierarchy approach to developing mathematics skills assessment systems. *Applied Sciences*, *10*(4), 1465–1483. https://doi.org/10.3390/app10041465

Tavakol, M., & Dennick, R. (2013). Psychometric evaluation of a knowledge based examination using Rasch analysis: An illustrative guide: AMEE guide no. 72. *Medical teacher*, *35*(1), 838-848. https://doi.org/10.3109/0142159X.2012.737488

Thompson, N. A. (2010). KR-20. In N. Salkind (Ed.), *Encyclopedia of research design* (pp. 667–668). Sage.

Ulusoy, F. (2016). The role of learners' example spaces in example generation and determination of two parallel and perpendicular line segments. In C. Csíkos, A. Rausch, & J. Szitányi (Eds.), *Proceedings of the 40th Conference of the International Group for the Psychology of Mathematics Education* (Vol. 4, pp. 299-306). PME.

Ulusoy, F. (2019). Early-years prospective teachers' definitions, examples and non-examples of cylinder and prism. *International Journal for Mathematics Teaching and Learning*, *20*(2), 149–169. https://cimt.org.uk/ijmtl/index.php/IJMTL/article/view/213/72

Ulusoy, F. (2021). Prospective early childhood and elementary school mathematics teachers' concept images and concept definitions of triangles. *International Journal of Science and Mathematics Education*, *19*(5), 1057–1078. https://doi.org/10.1007/s10763-020-10105-6

Ulusoy, F. (2022). Middle school students' reasoning with regards to parallelism and perpendicularity of line segments. *International Journal of Mathematical Education in Science and Technology* [Advance Online Publication]. https://doi.org/10.1080/0020739X.2022.2049384

Van Hiele, P. M. (1986). *Structure and insight*. Academic Press.

Yao, X., & Manouchehri, A. (2019). Middle school students' generalizations about properties of geometric transformations in a dynamic geometry environment. *The Journal of Mathematical Behavior*, *55*, Article 100703. https://doi.org/10.1016/j.jmathb.2019.04.002