Invited Lecture

Online Cognitive Diagnostic Assessment with Ordered Multiple-Choice Items for Grade Four Topic of Time

Cheng Meng Chew[1] and Huan Chin[2]

ABSTRACT It is a great challenge for the teachers to practice differentiated instruction in the heterogeneous mathematics classroom because there is a great demand for a valid and reliable diagnostic assessment. To address this issue, this study sought to develop and validate an online Cognitive Diagnostic Assessment (CDA) with Ordered Multiple-Choice (OMC) items for Grade Four Topic of Time. However, this paper only focuses on the results of six cognitive models for conversion between time units. Each cognitive model was measured by an assessment comprising seven OMC items. The quality of the online CDA with OMC items was assured with robust psychometric properties, convincing reliability, and satisfactory model-data fit. Perhaps this instrument could support the teachers in diagnosing pupils' cognitive strengths and weaknesses, followed by practicing differentiated instruction in the mathematics classroom.

Keywords: Cognitive diagnostic assessment; Ordered multiple choice; Time.

1. Introduction

Students' diversity is a critical issue to be addressed for promoting equal opportunity in mathematics learning in a heterogeneous classroom (Csapó and Molnár, 2019). Thus, teachers are encouraged to practise differentiated instruction in the mathematics classroom to tailor to students' needs. However, it is a great challenge for the teachers to practise differentiated instruction because there is a great demand for a valid and reliable diagnostic assessment (Brendefur et al., 2018) that could provide detailed information on students' skill deficits. The teachers could only identify students' learning needs by conducting informal assessments through classroom interactions, evaluating the students' learning artefacts, or assessing students' understanding of a narrow scope using the teacher-made instrument, which might have quality concerns (Csapó and Molnár, 2019).

By integrating educational measurement with learning psychology, cognitive diagnostic assessment (CDA) emerged as an alternative assessment that could support teachers' classroom assessment practice. The CDA consisted of three components: (i)

[1] School of Educational Studies, Universiti Sains Malaysia, 11800, Penang, Malaysia.
E-mail: cmchew@usm.my; School of Education, Humanities and Social Sciences, Wawasan Open University, 10050, Penang, Malaysia. Email: cmchew@wou.edu.my
[2] School of Educational Studies, Universiti Sains Malaysia, 11800, Penang, Malaysia.
E-mail: chin_huan_@hotmail.com

cognition, (ii) observation, and (iii) interpretation. Different from the diagnostic assessment commonly used in the classroom, the CDA was developed based on a cognitive model (cognition component) which illustrates the skill acquisition hierarchy. While each item (observation component) in CDA was designed to elicit the students' response on the subskill (attribute) included in the cognitive model, the use of measurement model (interpretation component) to analyse students' responses could reflect their knowledge states, which composed of attribute mastery combinations and corresponding misconceptions or systematic errors (Kuo et al., 2016). Thus, the diagnostic inference such as students' weakness in skill acquisition could be made based on the results obtained from the CDA. Based on this imperative information, the teachers could adjust the instruction to support the needs of the struggling students in mathematics learning (Ketterlin-Geller et al., 2019). In short, CDA could be essential to support teachers in implementing differentiated teaching in the mathematics classroom.

In fact, CDA is rarely being used in the mathematics classroom (Wu, 2019) due to practical constraints. To locate students' skill deficits, the teachers would have to apply a measurement model on the binary pattern of students' responses, after scoring the answer script dichotomously. This eventually discourages the practical implementation of CDA in the mathematics classroom because the application of highly technical measurement models might barely be understood by teachers. To address this issue, this study sought to develop the CDA as a web application with an automated scoring feature. Following this, the complex scoring procedure of CDA could be mechanised and thereby increase the practicability of the CDA to be used in the mathematics classroom.

Several CDAs have been developed in the past using multiple-choice questions (e.g., Ketterlin-Geller et al., 2019; Roberts et al., 2014) or constructed response questions (e.g., Sia and Lim, 2018; Sia et al., 2019). Rather than developing a CDA with multiple-choice questions or constructed response questions, we developed the online CDA with a novel item format, named ordered multiple-choice (OMC) items, which was introduced by Briggs et al. (2006). Apparently, OMC items and multiple-choice questions look quite similar (Briggs and Alonzo, 2012). However, each option of OMC items is linked to the developmental level of students' skill acquisition as depicted in the construct map (Briggs et al., 2006). Thus, OMC items have a higher diagnostic value compared to typical multiple-choice questions, while retaining their scoring efficiency advantage (Briggs et al., 2006).

In this study, we only focused on the topic of "Time" which is important in daily life yet hardly being mastered by the students (Kamii and Russell, 2012, Tan et al., 2017). This topic is included in the domain of measurement (National Council of Teachers of Mathematics [NCTM], 2000; Malaysian Ministry of Education [MOE], 2016). In grade four, the Malaysian students (average age of 10 years old) learn about the conversion between the time units, performing an arithmetic operation on the measurement with time units and solving problems involving time units (MOE, 2016).

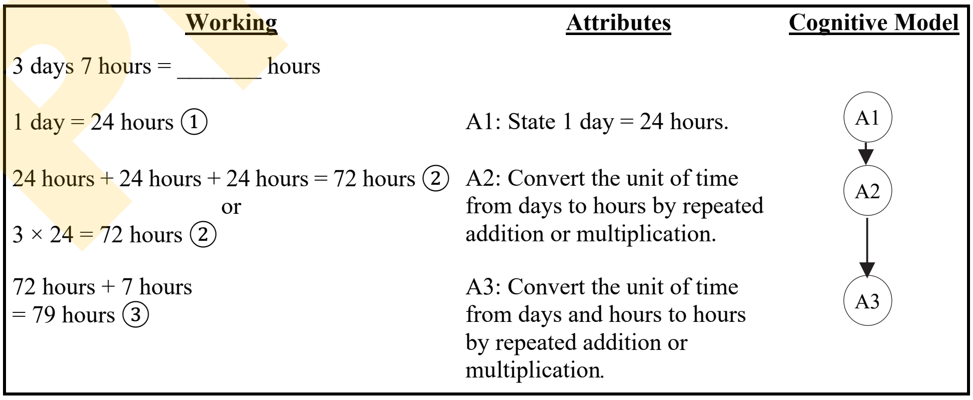
To support teachers' classroom assessment practice, we developed an online CDA with OMC items for the Grade Four topic of "Time". As part of the larger research project, this paper only discusses the development and validation of the online CDA for conversion between the time units.

2. Development of Online CDA with OMC Items

To promote evidentiary coherence for enhancing the validity of result interpretation and use (Nichols et al., 2017), the online CDA with OMC items was developed based on the principled assessment design adapted from the assessment system of Berkeley Evaluation and Assessment Research (BEAR) Centre (Wilson and Sloane, 2000). The usefulness of the BEAR assessment system to guide the development of OMC items has been illustrated in several studies (i.e., Briggs et al., 2006; Hadenfeldt et al., 2013; 2016). By adapting the BEAR assessment system, the online CDA with OMC items was developed sequentially following the five building blocks as described in the following sections.

2.1. *Building block 1: construct map*

The first building block involves the development of construct maps, which serve as the basis for OMC items construction. Due to the absence of a substantive theory on conversion between time units, two associate professors in the field of mathematics education were invited to specify the attributes by conducting the expert review and task analysis. With a sophisticated understanding of the curriculum, the two experts listed out the six skills related to conversion between time units, based on the review of the Grade Four mathematics curriculum document and textbook. Then, the experts selected a task for each skill and conducted the task analysis to specify the attributes required to master the skill. The example of task analysis for *conversion of the unit of time involving day and hour from a larger unit to a smaller unit* is as shown in Fig. 1.



Working	**Attributes**	**Cognitive Model**
3 days 7 hours = _______ hours		
1 day = 24 hours ①	A1: State 1 day = 24 hours.	A1
24 hours + 24 hours + 24 hours = 72 hours ② or 3 × 24 = 72 hours ②	A2: Convert the unit of time from days to hours by repeated addition or multiplication.	A2
72 hours + 7 hours = 79 hours ③	A3: Convert the unit of time from days and hours to hours by repeated addition or multiplication.	A3

Fig. 1. Example of task analysis and cognitive model derivation

The process began with solving the task and showing the working in a detailed manner. This was followed by specifying the attributes based on each step by considering the measurability of the attributes (Alves, 2012). Then, the attributes were arranged in a hierarchical order to form the cognitive model. The construct map was then derived based on the cognitive model, where each level indicates an accumulation of attributes mastered by the students. This was supported by the claim made by Szilagyi et al. (2013) in which mathematics learning is conceptualised as an accumulation of knowledge and skills. The six construct maps developed are as shown in the Appendix.

2.2. *Building block 2: item design*

The second building block involved the construction of the stem, key, and distractors of the OMC items. To ensure the alignment between the OMC items developed and the cognition component of CDA, the second building block started with generating the Q-matrices which outline the item-attribute relationship in the CDA developed. The Q-matrices are generated sequentially as shown in Fig. 2.

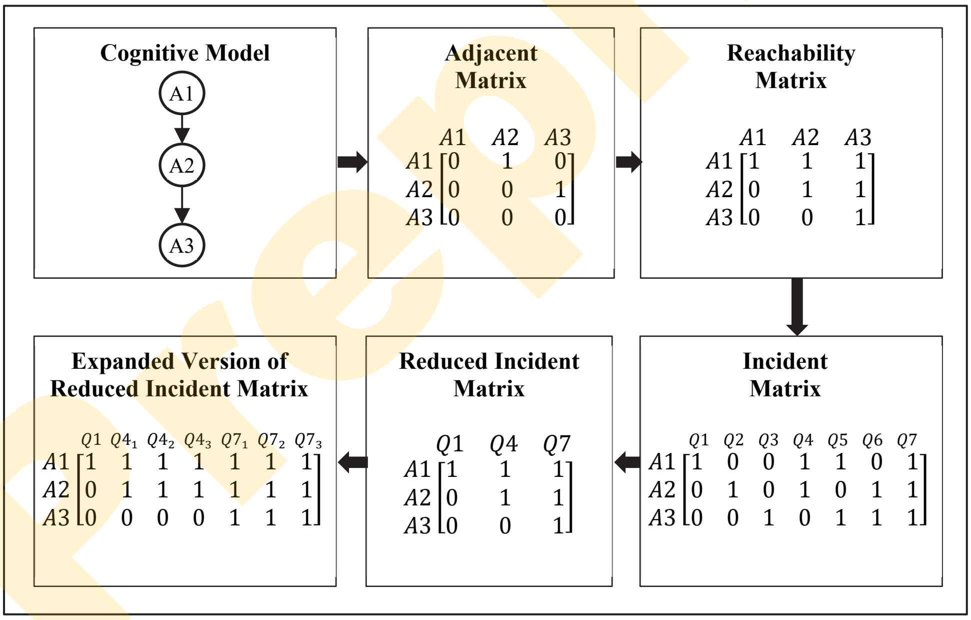

Cognitive Model
A1
A2
A3
Adjacent Matrix
$\begin{array}{cccc} & A1 & A2 & A3 \\ A1 & 0 & 1 & 0 \\ A2 & 0 & 0 & 1 \\ A3 & 0 & 0 & 0 \end{array}$
Reachability Matrix
$\begin{array}{cccc} & A1 & A2 & A3 \\ A1 & 1 & 1 & 1 \\ A2 & 0 & 1 & 1 \\ A3 & 0 & 0 & 1 \end{array}$
Incident Matrix
$\begin{array}{cccccccc} & Q1 & Q2 & Q3 & Q4 & Q5 & Q6 & Q7 \\ A1 & 1 & 0 & 0 & 1 & 1 & 0 & 1 \\ A2 & 0 & 1 & 0 & 1 & 0 & 1 & 1 \\ A3 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \end{array}$
Reduced Incident Matrix
$\begin{array}{cccc} & Q1 & Q4 & Q7 \\ A1 & 1 & 1 & 1 \\ A2 & 0 & 1 & 1 \\ A3 & 0 & 0 & 1 \end{array}$
Expanded Version of Reduced Incident Matrix
$\begin{array}{cccccccc} & Q1 & Q4_1 & Q4_2 & Q4_3 & Q7_1 & Q7_2 & Q7_3 \\ A1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ A2 & 0 & 1 & 1 & 1 & 1 & 1 & 1 \\ A3 & 0 & 0 & 0 & 0 & 1 & 1 & 1 \end{array}$

Fig. 2. Generation of Q matrices

The first step is converting the cognitive model which is illustrated as the directed network into the adjacent matrix (*A matrix*) by positioning the value '1'and '0' at the respective entry to indicate the presence or absence of a direct prerequisite relationship between the attribute pair. As shown in Fig. 2, the value '1' is positioned at Row 1, Column 2 and Row 2, Column 3 because A1 is the direct prerequisite attribute for A2, while A2 is the direct prerequisite attribute for A3. The ***A*** matrix is used to represent the direct prerequisite relationship of the attributes in CDA (Tatsuoka, 1990).

The second step is deriving the reachability matrix (***R*** *matrix*) by performing Boolean addition and multiplication using the formula $\boldsymbol{R} = (\boldsymbol{A} + \boldsymbol{I})^{n}$, where ***A*** refers to the ***A*** *matrix*, ***I*** refers to the Identity Matrix and n refers to the integer required to reach invariance, where $n = 1, 2, 3, \ldots, c$. In other words, ***R*** *matrix* is obtained if the matrix remains the same when the integer, *n* is substituted with two subsequent integers. The value of '1'or '0'in the ***R*** *matrix* indicates the presence or absence of the direct and indirect relationships between the attributes.

The third step is deriving the incident matrix (***Q*** *matrix*) that illustrates the involvement of the attributes in each item of the potential item pool. The number of potential items, *i* can be determined by using the formula, $i = 2^{k} - 1$, where *k* is the number of attributes. Since there are three attributes in the cognitive model as shown in Figure 2, the CDA might consist of seven items with different involvement of attributes. The attributes involve in each item was determined based on the subsets of attributes (i.e., A1, A2, A3), such as {A1}, {A2}, {A3}, {A1, A2}, {A2, A3}, {A1, A3}, {A1, A2, A3} and {}. Except for the empty set, each subset illustrates the attribute(s) involved in each potential item in the ***Q*** *matrix*. Then, the value of '1'or '0'is positioned at the row entry for each column in the ***Q*** *matrix* based on the involvement of attributes.

The fourth step is deriving the reduced incident matrix ($\boldsymbol{Q_r}$ *matrix*) from the ***Q*** *matrix* by imposing the prerequisite relationships among the attributes. Based on the cognitive model as shown in Fig. 2, attribute A2 has a prerequisite attribute (i.e., A1), while attribute A3 has two prerequisite attributes (i.e., A1 and A2). Notably, items Q2, Q3, Q5, and Q6 do not comply with the prerequisite relationship among the attributes. Thus, they were removed from the item pool. This brings about the reduction of columns Q2, Q3, Q5 and Q6 in the ***Q*** *matrix*. Following this, the $\boldsymbol{Q_r}$ *matrix* formed only consists of three columns, namely Q1, Q4, and Q7. Thus, the $\boldsymbol{Q_r}$ *matrix* serves as the basic item pool of CDA.

The last step is deriving the expanded version of the reduced incident matrix (expanded version of $\boldsymbol{Q_r}$ *matrix*) which serves as the test blueprint of CDA. To increase the reliability of the CDA, each attribute is measured by three parallel items (Gierl et al., 2009). Following this, a total of nine items (3 attributes × 3 items per attribute = 9 items) should be constructed for the three-attribute cognitive model as shown in Figure 2. However, only seven items were constructed based on the cognitive model because attribute A1 could only be elicited by one item. To accommodate all items and the respective attributes required to solve the CDA items, the $\boldsymbol{Q_r}$ matrix is expanded as shown in Fig. 2.

Upon the generation of the expanded version of $\boldsymbol{Q_r}$ *matrix*, the construction of OMC items began. The process started with the construction of the OMC item stem based on the expanded version of $\boldsymbol{Q_r}$ *matrix*. A total of 42 items (6 cognitive models × 7 items per cognitive models = 42 items) were constructed for the six cognitive models specified. The English written OMC items were then being translated into Malay, Mandarin and Tamil languages to match the medium of instruction in the three types of primary schools. This was followed by the content validation of the OMC item stems by two subject matter experts from each school type with at least eight years of Grade Four mathematics teaching experience. After the validation process, the OMC item stems were piloted to a total of 192 grade four students from each school type to collect the students' common mistakes for each item. For each OMC item, the correct answer would form the key, while the incorrect answers associated with the common mistakes would form the three distractors. It is worth noting that the pupils seldom made mistakes when answering the item which measured the basic attribute such as A1. As a result, the distractors could barely be extracted from the mistake made by the pupils. In this situation, the distractors could be derived from the common mistake reported in the literature (Sadler, 1998).

2.3. *Building block 3: outcome space*

The third building block involved the construction of the outcome space, which specifies the relationship of the OMC items with the construct map (Wilson, 2005). In this study, the outcome space refers to the relationship of each option of OMC items with the level of mastery of attributes as shown in the construct map. Two mathematics education experts were invited for assigning the level to each option of the OMC items based on the construct map. Specifically, the correct option of each item was assigned to the mastery level which includes the corresponding attribute being measured. On the other hand, each incorrect option was assigned to the lower mastery level based on the incorrect working and associated with the common mistakes presented to the experts. A sample of the OMC item with both correct and incorrect workings associated with each option is illustrated in Fig. 3.

Stem	88 hours = ______ days ______hours			
Option	A. 1 day	B. 3 days	C. 3 days 16 hours	D. 8 days 8 hours
Working	24 hours = 1 day	24 hours = 1 day 3 → 3 days 24)88 72 16	24 hours = 1 day 3 → 3 days 24)88 72 16 → 16 hour	10 hours = 1 day 8 → 8 days 10)88 80 8 → 8 hours
Mistake	State 24 hours = 1 day but do not proceed with the conversion of days into days and hours.	Perform correct long division but do not write down the remainder as the number of hours.	None	Divide the number of hours with the wrong divisor
Level	Level 1	Level 2	Level 3	Level 0

Fig. 3. Assigning the level to each option of the OMC items based on the construct map

To ensure the validity of the level assigned, each option and the corresponding level was then validated by an associate professor and a senior lecturer in the field of psychometric and educational measurement with mathematics teaching experience. After the validation process, the English version of the OMC items was translated into Malay, Mandarin and Tamil languages to match the medium of instruction in the three types of primary schools.

2.4. *Building block 4: online CDA web application*

The fourth building block involved the development of an online CDA web application by the web developer. The online CDA is a cohesive and comprehensive online assessment system that could be used to assess pupils' understanding of the topic of 'Time'and to profile their skill acquisition. To match the medium of instruction in the three types of primary schools, the online CDA developed possesses a language switching function. Specifically, the Malay, Mandarin, and Tamil language versions of the online CDA would be used by the pupils from National Primary School (NPS), National Type Chinese Primary School (NTCPS) and National Type Tamil Primary School (NTTPS), respectively.

The pupils, teachers and researchers were the three main users of the online CDA. The pupils would be able to take the assessment and view their scoring report if they logged in to their account in the online CDA. The teachers would be able to view the content of the assessment and the pupil's scoring report at both individual and class levels in the online CDA after their pupils had answered the assessment. The researchers would be able to manage the item bank of the online CDA, manage the users, access, and extract the scoring reports at the assessment level. Upon the completion of the development process, the researchers created the item bank by keying in all the attributes, followed by the OMC items and the respective answer key. After entering all the attributes, the OMC items, and the answer key in the item bank, the Online CDA was ready to be used in the classroom.

2.5. *Building block 5: measurement model*

The fifth building block involved applying the measurement model to determine the psychometric properties of each item, to evaluate the reliability of the assessment, to assess the model-data fit, and to map the pupil's responses onto the construct map. The responses collected during the pilot test were analysed by using the measurement model, named Classical Test Theory (CTT) to determine the psychometric properties of each item such as item difficulty (proportion of correct response [p-value]) and item discrimination (point-biserial correlation [r_{pb}]). This is because the result of item analysis could be communicated easily to the teachers during the item revision stage.

Likewise, the reliability of the dichotomously scored CDA was determined using Kuder Richardson 20 [KR-20] coefficient, which is rooted in CTT. In this study, the CDA was developed to measure students' attribute mastery. In other words, the students'results would be reported at attribute level, besides total score. Thus, the consistency of the attribute-level measurement was determined using the attribute reliabilities coefficient (α_{AHM}) in the measurement model, named Attribute Hierarchy Method (AHM). The α_{AHM} is derived from the Cronbach's Alpha reliability coefficient by imposing the prerequisite relationships among the attributes. The formula of α_{AHM} is as follows:

$$\alpha_{AHM_k} = \frac{n_k}{1-n_k}\left[1 - \frac{\sum_{i\in S_k} W_{ik}{}^2\sigma^2 x_i}{\sigma^2 \sum_{i\in S_k} W_{ik}x_i}\right], \qquad (1)$$

where n_k = number of items which involves attribute k, S_k = a set which consists of items which involve attribute k, i = an element in the set S_k, W_{ik}=$P(X_i = 1|\, A_k = 1) - P(X_i = 1|\, A_k = 0)$, where $P(X_i = 1|\, A_k = 1)$ is the conditional probability that an examinee to answer item i correctly, given that he has mastered the attribute k (i.e., percent score for attribute k more than .50), $P(X_i = 1|\, A_k = 0)$ is the conditional probability that an examinee who has not mastered attribute k (i.e., percent score for attribute k at most .50) can answer item i correctly, $\sum_{i\in S_k} W_{ik}{}^2\sigma^2{}_{X_i}$ = sum of the weighted variances of the observed score for item i, and $\sigma^2{}_{\sum_{i\in S_k} W_{ik}X_i}$= variance of the weighted observed total score (Gierl et al., 2009, p. 300). To explain the result of the attribute reliability, the effect of adding parallel items to the CDA was determined using the attribute-based Spearman-Brown formula as follows:

$$\alpha_{AHM-SB_k} = \frac{n_k\alpha_{AHM_k}}{1 + (n_k - 1)\alpha_{AHM_k}}, \qquad (2)$$

where, n_k is the number of additional parallel items involving attribute k added into the assessment, and α_{AHM_k} is the reliability of attribute k (Gierl et al., 2009, p. 300).

Then, the model data fit was used to describe the extent to which the students' responses match the expected response derived based on the cognitive model. Following this, the evaluation of model data fit would provide validity evidence on the CDA developed. In this study, the model data fit was determined based on the Hierarchical Consistency Index (HCI) in AHM. The formula of HCI is as follows:

$$HCI_i = 1 - \frac{2\sum_{j\in S_{\text{correct}_i}}\sum_{g\epsilon S_j} X_{i_j}\left(1-X_{i_g}\right)}{N_{c_i}}, \qquad (3)$$

where S_{correct_i} is a set that consists of the items that are correctly answered by student I, X_{i_j} is the score (1 or 0) of student i for the item j, where item j is an element in the set $S_{\text{correct}_i}S_j$ is s set which consists of the items which required subset of attributes measured by item j, where item $j \notin S_j$, X_{i_g} is the score (1 or 0) of student i for the item

g, where item g is an element in the set S_j, N_{c_i} is the total number of comparisons for all the items that are correctly answered by the student i (Cui and Leighton, 2009, p. 436).

After analysing the HCI, the pupils' percentage subscore for each attribute was calculated. Then, the attribute mastery pattern of each cognitive model was determined by categorizing each attribute probability estimated into 'Mastery' and "Non-mastery"-based on the cut-off score proposed by Bradshaw (2017). Since the attribute mastery is presented as a row vector, the attribute probability exceeding the minimum threshold of .50 would be categorised as "Mastery" and coded as '1', whereas the attribute probability less than or equal to .50 would be categorised as "Non-mastery" and coded as '0'. For each assessment, the attribute mastery was then mapped onto the corresponding levels in the construct map based on the guidelines given in the Appendix.

3. Reliability and Validity Study

Reliability and validity are the important facets of assessment that should be evaluated during assessment development for ensuring the consistency of the measurement and meaningfulness of test score interpretation. Since Malaysia practices the vernacular school system, the reliability and validity study of online CDA developed were conducted in a Malay-medium NPS, a Mandarin-medium NTCPS, and a Tamil medium NTTPS. The sample of the study consisted of 90 Year Four pupils from NPS (30), NTCPS (48) and NTTPS (12) in Penang, Malaysia. The findings of the reliability and validity study are reported in the following sections.

3.1. *Psychometric properties of OMC items*

The psychometric properties of OMC items were evaluated based on CTT. The range and mean of item difficulty index (p-value) and item discrimination index (r_{pb}) of OMC items are tabulated in Tab. 1. Although the three versions of the assessments consisted of the same items, the findings indicated that the item difficulty of the assessments was not the same. Specifically, all assessments were considered as very easy for the pupils from NTCPS with the mean difficulty index ranging from .91 to .94 (>.80) (Tavakol and Dennick, 2011). However, only four out of the six assessments and three out of the six assessments were considered as very easy for the pupils from NPS and NTTPS respectively (Tavakol and Dennick, 2011). Nonetheless, all items in the six assessments were still considered as very good discriminating items with the minimum mean r_{pb} of .52 (>.40) (Ebel and Frisbie, 1991), regardless of the difference in terms of language. In other words, these items were good in differentiating pupils from high mastery level and the low mastery level across the school type.

Tab. 1. Psychometric properties of OMC items

	NPS				NTCPS				NTTPS			
	p-value		r_{pb}		*p*-value		r_{pb}		*p*-value		r_{pb}	
CDA	range	*M*	range	*M*	range	*M*	range	*M*	range	*M*	range	*M*
CDA 1	.83−.97	.88	.45−.89	.61	.85−.96	.92	.26−.82	.56	.83−.92	.86	.47−.93	.79
CDA 2	.63−.97	.82	.39−.74	.63	.88−.98	.94	.33−.75	.58	.42−.92	.70	.29−.89	.69
CDA 3	.57−.97	.80	.31−.90	.56	.85−.98	.93	.38−.82	.54	.58−.92	.79	.39−.93	.61
CDA 4	.50−.97	.79	.38−.81	.65	.77−.98	.91	.35−.83	.58	.50−.92	.79	.28−.93	.64
CDA 5	.63−.97	.84	.20−.87	.61	.81−.98	.93	.23−.76	.52	.75−.92	.83	.48−.94	.72
CDA 6	.47−.97	.83	.39−.85	.61	.81−.98	.92	.37−.75	.54	.58−.92	.81	.33−.90	.66

3.2. *Reliability of assessment*

The reliability of the assessment was evaluated based on AHM (i.e., α_{AHM}) and CTT (KR-20). The attribute reliabilities (α_{AHM}) and KR-20 of each assessment are as shown in Tab. 2. Attributes A1 and A2 in each CDA for NPS, NTCPS and NTTPS showed moderately high reliability, with the alpha coefficient ranging from .51 to .93 (Hinton, McMurray and Brownlow, 2014). Compared to attributes A1 and A2, the reliability of attribute A3 in each CDA for NPS, NTCPS and NTTPS were found to be lower with the range of .17 to .70.

Tab. 2. Attribute reliabilities and KR-20 of each assessment

	CDA 1				CDA 2				CDA 3				CDA 4				CDA 5				CDA 6			
School-Type	α_{AHM} *CM1A1*	α_{AHM} *CM1A2*	α_{AHM} *CM1A3*	*KR-20*	α_{AHM} *CM2A1*	α_{AHM} *CM2A2*	α_{AHM} *CM2A3*	*KR-20*	α_{AHM} *CM3A1*	α_{AHM} *CM3A2*	α_{AHM} *CM3A3*	*KR-20*	α_{AHM} *CM4A1*	α_{AHM} *CM4A2*	α_{AHM} *CM4A3*	*KR-20*	α_{AHM} *CM5A1*	α_{AHM} *CM5A2*	α_{AHM} *CM5A3*	*KR-20*	α_{AHM} *CM6A1*	α_{AHM} *CM6A2*	α_{AHM} *CM6A3*	*KR-20*
NPS	.72	.66	.48	.69	.77	.72	.49	.73	.73	.71	.70	.73	.82	.73	.61	.78	.74	.71	.55	.71	.77	.74	.63	.74
NTCPS	.69	.70	.61	.68	.68	.59	.47	.65	.62	.59	.47	.59	.70	.66	.36	.66	.60	.60	.67	.58	.58	.51	.17	.53
NTTPS	.93	.89	.66	.90	.84	.78	.77	.84	.76	.73	.68	.74	.79	.74	.56	.76	.87	.84	.55	.84	.83	.78	.59	.80

This is because the number of items that measured attribute A3, directly and indirectly, is relatively less compared to that of A1 and A2 (Alves, 2012). Due to the prerequisite relationship among the attributes, the simple attribute (i.e., A1) will be measured indirectly using the items which elicit the response for the more complex attributes (i.e., A2 and A3) in the attribute hierarchy. Thus, attribute A1 is measured by 7 items (1 item$_{A1}$ + 3 items$_{A2}$ + 3 items$_{A3}$), attribute A2 is measured by 6 items (3 items$_{A2}$ + 3 items$_{A3}$), and attribute A3 is only measured by 3 items (3 items$_{A3}$).

Although the attribute A3 of some CDA for NPS and NTCPS was found to have low reliability with α_{AHM} ranging from .17 to .43 (Hinton et al., 2014), the result was acceptable as Gierl et al. (2009) argued that the short diagnostic tests with less than 12 items per cognitive model could hardly yield satisfactory attribute reliability. Based on the attribute-based Spearman-Brown formula, these unsatisfactory attribute reliabilities could be increased to at least .50 by increasing the number of parallel items by two-fold. Nonetheless, it comes with the price of an increase of time allocation for each CDA (Gierl et al., 2009). This might cause the participants with low performance to be opted out from the main research project as nearly 30 assessments had been developed in total.

In general, all CDAs were reliable with the values of KR-20 ranging from .54 to .86 surpassing the minimum threshold (KR-20 = .50) for an assessment with less than 15 items (Kehoe, 1994). This indicates that the CDAs could provide a consistent measurement of students'understanding using the total score.

3.3. *Model data fit*

The model data fit of each assessment was evaluated based on AHM using HCI. The HCI of the cognitive model corresponded to each assessment is as shown in Tab. 3. The mean HCI for NPS pupils and NTTPS ranged from .70 to .80 for the six cognitive models. This indicates that the pupils from NPS and NTTPS exhibited a moderate fit (.60 ≤ mean HCI ≤ .80) for the six cognitive models (Roberts et al., 2014). Compared to NPS and NTTPS, the six cognitive models for NTCPS were found to have a better model-data fit with the mean HCI ranging from .86 to .88. This might be due to the higher mathematical proficiency of the pupils from NTCPS compared to their counterparts from NPS and NTTPS as reported by Ghazali and Sinnakaudan (2014).

Tab. 3. HCI of the cognitive model corresponded to each assessment

	NPS		NTCPS		NTTPS		Overall	
CDA [CM]	M_{HCI}	Fit category	M_{HCI}	Fit category	M_{HCI}	Fit category	M_{HCI}	Fit category
CDA 1 [CM 1]	.79	Moderate	.88	Excellent	.72	Moderate	.83	Excellent
CDA 2 [CM 2]	.77	Moderate	.86	Excellent	.72	Moderate	.81	Excellent
CDA 3 [CM 3]	.80	Moderate	.88	Excellent	.73	Moderate	.83	Excellent
CDA 4 [CM 4]	.72	Moderate	.86	Excellent	.70	Moderate	.81	Excellent
CDA 5 [CM 5]	.76	Moderate	.88	Excellent	.71	Moderate	.82	Excellent
CDA 6 [CM 6]	.75	Moderate	.88	Excellent	.70	Moderate	.81	Excellent

Notes. CM indicates a cognitive model. M_{HCI} less than .60 indicates poor fit, M_{HCI} between .60 and .80 indicates moderate fit, and M_{HCI} more than .80 indicates excellent fit (Roberts et al., 2014)

In general, the six cognitive models that corresponded to each assessment were found to be excellently consistent with the student's responses with the mean HCI ranging from .81 to .83 (Cui and Leighton, 2009). This implies that the mathematics experts correctly identified the relevant attributes and their ordering through the task analysis (Robert et al., 2014). Following this, the attributes used by the pupils in solving the tasks were consistent with the prediction of the mathematics education experts. Thus, the diagnostic inferences made based on the six cognitive models would be valid.

4. Concluding Remarks

CDA is an alternative assessment that can provide a clear picture of the pupils' learning process to education stakeholders so that instructional strategies can be designed to tailor to pupils' needs. In this paper, we describe the process of development and validation of the online CDA with OMC items for conversion between the time units. The findings of the validity and reliability study indicated that the OMC items developed were of good quality with high discrimination power even though most of the OMC items were considered very easy. Besides that, the online CDA with OMC items developed in this study was found to be reliable at both attribute level and assessment level. With the satisfactory model-data fit, the inferences made about pupils' attribute mastery based on their performance in the online CDA with OMC items were valid. Perhaps this instrument could support the teachers in diagnosing pupils' cognitive strengths and weaknesses, followed by practising differentiated instruction in the mathematics classroom.

However, we identified some limitations in the process of developing the online CDA with OMC items. Due to the practical constraints, the sample size of the study was quite small. This could affect the generalisability of findings. To address this limitation, future studies are recommended to be conducted with a larger sample size. Since the online CDA with OMC items was translated into three different languages, future studies are suggested to analyse the Differential Item Functioning (DIF) to identify the potential item bias which might be present in the multi-lingual online CDA. Future studies should also explore the practical use of the online CDA with OMC items in the classroom setting.

Acknowledgements

This study was made possible with funding from the Research University Grant (RUI) Scheme 1001/PGURU/8011027 of Universiti Sains Malaysia (USM). The authors would like to thank all the subject matter experts, mathematics educators and psychometric experts who were involved in the development and validation process of the online CDA with OMC items. The authors also would like to thank the computer science experts, the web developer, and the technician from Centre for Knowledge, Communication and Technology, USM who were involved in the Online CDA web application development process. Lastly, the authors would like to thank all the teachers and pupils who voluntarily participated in this study.

Appendix

CDA/ Construct Map [CM]	Attributes [Code]	Descriptors [Attribute Mastery Pattern]
CDA 1 [CM 1]	1. State 1 day = 24 hours [CM1A1] 2. Convert the unit of time from days 3. to hours by repeated addition or multiplication [CM1A2] 4. Convert the unit of time from days and hours to hours by repeated addition or multiplication [CM1A3]	Level 0: Do not master any attribute [0 0 0] Level 1: Master attribute CM1A1 [1 0 0] Level 2: Master attributes CM1A1 and CM1A2 [1 1 0] Level 3: Master attributes CM1A1, CM1A2 andCM1A3 [1 1 1]
CDA 2 [CM 2]	1. State 24 hours = 1 day [CM2A1] 2. Convert the unit of time from hours 3. to days by repeated subtraction or 4. division [CM2A2] 5. Convert the unit of time from hours to days and hours by repeated subtraction or division [CM2A3]	Level 0: Do not master any attribute [0 0 0] Level 1: Master attribute CM2A1 [1 0 0] Level 2: Master attributes CM2A1 and CM2A2 [1 1 0] Level 3: Master attributes CM2A1, CM2A2 and CM2A3 [1 1 1]
CDA 3 [CM 3]	1. State 1 week = 7 days [CM3A1] 2. Convert the unit of time from weeks 3. to days by repeated addition or multiplication [CM3A2] 4. Convert the unit of time from weeks 5. and days to days by repeated addition 6. or multiplication [CM3A3]	Level 0: Do not master any attribute [0 0 0] Level 1: Master attribute CM3A1 [1 0 0] Level 2: Master attributes CM3A1 and CM3A2 [1 1 0] Level 3: Master attributes CM3A1, CM3A2 andCM3A3 [1 1 1]
CDA 4 [CM 4]	1. State 7 days = 1 week [CM4A1] 2. Convert the unit of time from days 3. to weeks by repeated subtraction or division [CM4A2] 4. Convert the unit of time from days to weeks and days by repeated subtraction 5. or division [CM4A3]	Level 0: Do not master any attribute [0 0 0] Level 1: Master attribute CM4A1 [1 0 0] Level 2: Master attributes CM4A1 and CM4A2 [1 1 0] Level 3: Master attributes CM4A1, CM4A2 andCM4A3 [1 1 1]
CDA 5 [CM 5]	1. State 1 year = 12 months [CM5A1] 2. Convert the unit of time from years 3. to months by repeated addition or multiplication [CM5A2] 4. Convert the unit of time from years 5. and months to months by repeated addition or multiplication [CM5A3]	Level 0: Do not master any attribute [0 0 0] Level 1: Master attribute CM5A1 [1 0 0] Level 2: Master attributes CM5A1 and CM5A2 [1 1 0] Level 3: Master attributes CM5A1, CM5A2 andCM5A3 [1 1 1]
CDA 6 [CM 6]	1. State 12 months = 1 year [CM6A1] 2. Convert the unit of time from months 3. to years by repeated subtraction or division [CM6A2] 4. Convert the unit of time from months 5. to years and months by repeated subtraction or division [CM6A3]	Level 0: Do not master any attribute [0 0 0] Level 1: Master attribute CM6A1 [1 0 0] Level 2: Master attributes CM6A1 and CM6A2 [1 1 0] Level 3: Master attributes CM6A1, CM6A2 andCM6A3 [1 1 1]

Notes.
CM 1: Conversion of the unit of time involving day and hour from a larger unit to a smaller unit
CM 2: Conversion of the unit of time involving day and hour from a smaller unit to a larger unit
CM 3: Conversion of the unit of time involving week and day from a larger unit to a smaller unit
CM 4: Conversion of the unit of time involving week and day from a smaller unit to a larger unit
CM 5: Conversion of the unit of time involving year and month from a larger unit to a smaller unit
CM 6: Conversion of the unit of time involving year and month from a smaller unit to a larger unit

References

C. B. Alves (2012). Making diagnostic inferences about student performance on the Alberta education diagnostic mathematics project: An application of the Attribute Hierarchy Method. (Doctoral Thesis), University of Alberta, Ann Arbor, Canada.

L. Bradshaw (2017). Diagnostic classification models. In A. A. Rupp and J. P. Leighton (Eds.), *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (1st ed., pp. 297–327). New York: Wiley Blackwell.

J. L. Brendefur, E. S. Johnson, K. W. Thiede, S. Strother, and H. H. Severson (2018). Developing a multi-dimensional early elementary mathematics screener and diagnostic tool: The primary mathematics assessment. *Early Childhood Education Journal*, 46(2), 153–157.

D. C. Briggs and A. Alonzo (2012). The psychometric modeling of ordered multiple-choice item responses for diagnostic assessment with a learning progression. In A. C. Alonzo and A. W. Gotwals (Eds.), *Learning Progressions in Science: Current Challenges and Future Directions* (pp. 293–316). Rotterdam, The Netherlands: Sense Publishers.

D. C. Briggs, A. Alonzo, C. Schwab, and M. Wilson (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment,* 11(1), 33–63.

B. Csapó and G. Molnár (2019). Online diagnostic assessment in support of personalized teaching and learning: The eDia System. *Frontiers in Psychology*, 10, Article 1522. doi: 10.3389/fpsyg.2019.01522

Y. Cui and J. P. Leighton (2009). The Hierarchy Consistency Index: evaluating person fit for cognitive diagnostic assessment. *Journal of Educational Measurement,* 46(4), 429–449.

R. L. Ebel and D. A. Frisbie (1991). *Essentials of Educational Measurement* (5th ed.). Englewood Cliffs: Prentice-Hall

M. Ghazali and S. Sinnakaudan (2014). Reasearch on teachers'beliefs about mathematics teaching and learning between Sekolah Kebangsaan (SK), Sekolah Jenis Kebangsaan Cina (SJKC) ans Sekolah Jenis Kebangsaan Tamil (SJKT). *Journal of Education and Practice*, 5(31), 10–19.

M. J. Gierl, Y. Cui, and J. Zhou (2009). Reliability and attribute-based scoring in cognitive diagnostic assessment. *Journal of Educational Measurement,* 46(3), 293–313.

J. C. Hadenfeldt, S. Bernholt, X. Liu, K. Neumann, and I. Parchmann (2013). Using ordered multiple-choice items to assess students'understanding of the structure and composition of matter. *Journal of Chemical Education,* 90(12), 1602–1608.

J. C. Hadenfeldt, K. Neumann, S. Bernholt, X. Liu, and I. Parchmann (2016). Students'progression in understanding the matter concept. *Journal of Research in Science Teaching,* 53(5), 683–708.

P. R. Hinton, I. McMurray, and C. Brownlow (2014). *SPSS Explained*. New York: Routledge.

C. Kamii and K. A. Russell (2012). Elapsed time: Why is it so difficult to teach*? Journal for Research in Mathematics Education,* 43(3), 296–345.

Kehoe, J. (1994). Basic item analysis for multiple-choice tests. *Practical Assessment, Research, and Evaluation*, 4(10), 1–4.

L. R. Ketterlin-Geller, P. Shivraj, D. Basaraba, and P. Yovanoff (2019). Considerations for using mathematical learning progressions to design diagnostic assessments. *Measurement: Interdisciplinary Research and Perspectives*, 17(1), 1–22.

B. C. Kuo, C. H. Chen, C. W. Yang, and M. M. C. Mok (2016). Cognitive diagnostic models for tests with multiple-choice and constructed-response items. *Educational Psychology*, 36(6), 1115–1133.

Malaysian Ministry of Education [MOE] (2016). *Kurikulum standard sekolah rendah: Standard kurikulum dan pentaksiran Matematik Tahun 4* [Primary School Standard Curriculum: Year 4 Mathematics Curriculum and Assessment Standard Document]. Putrajaya: Ministry of Education Malaysia.

National Council of Teachers of Mathematics. (2000). *Principles and Standards for School Mathematics.* NCTM.

P. D. Nichols, J. L. Kobrin, E. Lai, and J. D. Koepfler (2017). The role of theories of learning and cognition in assessment design and development. In A. A. Rupp and J. P. Leighton (Eds.), *The Handbook of Cognition and Assessment: Frameworks, Methodologies, and Applications* (1st ed., pp. 41–74). New York: Wiley Blackwell.

M. R. Roberts, C. B. Alves, M. W. Chu, M. Thompson, L. M. Bahry, and A. Gotzman (2014). Testing expert based versus student based cognitive models for a Grade 3 diagnostic mathematics assessment. *Applied Measurement in Education,* 27(3), 173–195.

P. M. Sadler (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in Science Teaching*, 35(3), 265–296.

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

C. J. L. Sia, C. S. Lim, C. M. Chew, and L. K. Kor (2019). Expert-based cognitive model and student-based cognitive model in the learning of "Time": Match or mismatch? *International Journal of Science and Mathematics Education*, 17(6), 1–19.

J. Szilágyi, D. H. Clements, and J. Sarama (2013). Young children's understandings of length measurement: Evaluating a learning trajectory. *Journal for Research in Mathematics Education*, 44(3), 581–620.

P. L. Tan, C. S. Lim, and L. K. Kor (2017). Diagnosing primary pupils'learning of the concept of "after" in the topic "Time" through knowledge states by using cognitive diagnostic assessment. *Malaysian Journal of Learning and Instruction,* 14(2), 145–175.

K. K. Tatsuoka (1990). Toward an integration of item-response theory and cognitive error diagnosis. In N. Frederiksen, R. Glaser, A. Lesgold, and M. G. Shafto (Eds.), *Diagnostic Monitoring of Skill and Knowledge Acquisition* (pp. 453–488). Hillsdale, NJ: Lawrence Erlbaum Associates

M. Tavakol and R. Dennick, R. (2011). Post-examination analysis of objective tests. *Medical Teacher*, 33(6), 447–458.

M. Wilson and K. Sloane (2000). From principles to practice: An embedded assessment system. *Applied Measurement in Education,* 13(2), 181–208.

M. Wilson (2005). *Construct Measures: An Item Response Modelling Approach*. New Jersey, London: Lawrence Erlbaum Associates.

H. M. Wu (2019). Online individualised tutor for improving mathematics learning: A cognitive diagnostic model approach. *Educational Psychology*, 39(10), 1218–1232.