Testing content validity of teacher-made test

Musah, Mohammed Borhandden; Tahir, Lokman Mohd; Al-Hudawi, Shafeeq Hussain; Issah, Mohammad; Hussein, Abdul Rawuf; Ibrahim, Mikail

Published in:
International Journal of Evaluation and Research in Education

DOI:
10.11591/ijere.v11i2.21992

Published: 01/06/2022

Document Version
Final published version

Document License
CC BY-SA

International Journal of Evaluation and Research in Education (IJERE)
Vol. 11, No. 2, June 2022, pp. 878~887
ISSN: 2252-8822, DOI: 10.11591/ijere.v11i2.21992

Testing content validity of teacher-made test: Profiling teacher perceptions and demographic variables

Mohammed Borhandden Musah[1], Lokman Mohd Tahir[2], Shafeeq Hussain Al-Hudawi[3], Mohammad Issah[1], Abdul Rawuf Hussein[4], Mikail Ibrahim[5]
[1]Department of Education Studies, Bahrain Teachers College, University of Bahrain, Sakhir, Bahrain
[2]School of Education, Faculty of Social Sciences and Humanities, Universiti Teknologi Malaysia, Johor Bahru, Malaysia
[3]Yanbu University College, Royal Commission of Yanbu Colleges & Institutes, Riyadh, Saudi Arabia
[4]Faculty of Education and Welfare Studies, Abo Akademi University, Turku, Finland
[5]Faculty of Major Languages Studies, Universiti Sains Islam Malaysia, Nilai, Malaysia

Article Info

Article history:

Received Nov 4, 2021
Revised Mar 9, 2022
Accepted Apr 5, 2022

Keywords:

Content validity
Demographic variable
Elementary school
Subject-matter expert
Table of specification
Teacher-made test

ABSTRACT

The content validity of teacher-made tests (TMT) in Malaysian Chinese primary schools was explored in this research. The study used teacher-made tests to analyze the selected Chinese primary schools, which have yet to be thoroughly explored in the context of these school types. It also investigated how well teachers in the sampled schools understand the table of specifications (TOS). The study further examined the extent to which included demographic variables affect the validity of teacher-made tests. A total of 660 questionnaires were distributed randomly to teachers from 21 Chinese primary schools. There were 381 completed questionnaires (58% of total rate) received and analyzed. The findings demonstrated that most teachers have a basic understanding of the TOS and that teacher-made test levels are valid. K-group MANOVA analysis also revealed that work experience and age had a significant influence on teacher understanding. Work experience and age, on the other hand, had no effect on the teacher-made test. Furthermore, the findings demonstrated that both teacher understanding and teacher-made tests were unaffected by teacher gender.

Corresponding Author:

Mohammed Borhandden Musah
Department of Education Studies, Bahrain Teachers College, University of Bahrain
Sakhir, Kingdom of Bahrain
Email: mmusah@edu.bh

1. INTRODUCTION

Teachers frequently use test scores as a yardstick for evaluating students' performance. The test results of students are combined from many exams taken over a period of time. Teachers should guarantee that the test instruments are legitimate and relevant [1]–[4]. since these tests affect the future job success of various learners. In the educational sector, there is still a lot of concern regarding test validity [5], [6]. With the emergence of alternate types of formative assessments, there is a global push to decentralize testing [7]. Malaysia has adopted a ground-breaking approach to the issue.

Until the mid-1990s, the Malaysian educational system mainly relied on test-based evaluations to determine student achievement. In order to address progressive expectations, the Ministry of Education (MoE) issued educational reform directives in 2011. The then Prime Minister of Malaysia, Najib bin Tun Abdul Razak stated that "in order to meet our high aspirations amidst an increasingly competitive global environment, we cannot stand still. Our country requires a transformation of its entire education system" (Preliminary Report, n.d.). As a result, the Ministry of Education (MoE) adopted school-based assessment

(SBA) in the educational system to replace test-based evaluations of students' performance with SBA [8]. SBA have been implemented in Malaysian elementary schools from 2011, followed by secondary schools in 2012. The former Minister of Education, Tan Sri Musa Mohamed is quoted to have said on May 7, 2003: "we need a fresh and new philosophy in our approach to exams … we want to make the education system less exam-oriented and (we) are looking at increasing SBA as it would be a better gauge of students' abilities" [9].

Students' performance in SBA is evaluated in terms of cognitive, emotional, and psychomotor components, using the National Philosophy of Education as a guide (NPE). The mentioned components must be developed in order for pupils' potential to grow in a balanced and comprehensive manner. However, teachers and students' impressions of the NPE's implementation were studied in a research, which found "major absence across the NPE element in developing the potentials of individuals according to students' perspectives and in education is an ongoing effort with reference to both students and teachers' perspectives" [10]. Teachers are enticed to broaden their attention and role in assisting students' overall development as a result of the approach. This transition revealed that teachers play a crucial role in the new assessment system, not only in measuring learner performance but also in successful instructional planning, where teacher-made tests (TMTs) play a significant role [11]. This is particularly essential given [12] that it has been revealed that more than half of the examinations used in classrooms are created by teachers. To put it another way, judgments based on student test results must guarantee that the exams are valid, relevant, and trustworthy [2], [13]. As a result, the test should be a precise measurement instrument.

Despite the previous change, learners continue to doubt the accuracy of exam findings [7], [14]–[16]. Furthermore, students believe that evaluation is sometimes wholly or relatively unrelated to the material covered in class. Simply put, the subject covered in class and the material scored at the end of unit exams are incompatible [17]. This problem might lead to a lack of trust in assessment tools as a source of evidence for teachers to make accurate assessments of students' development. Creating a table of specifications (TOS) is thus a technique for educators to constructively resolve this issue. However, despite the availability of standards and objectives surrounding TOS, it is said that Malaysian teachers lack understanding and abilities in conducting SBA [11]. In terms of using SBA in education, Chinese primary schools in Malaysia have given teachers the opportunity to measure the performance of their students. However, it is vibrantly debated whether the TMT at the mentioned schools is credible and meets the content validity and dependability requirements.

An assessment is a decision made after a thorough examination [18]. In schools, determining whether students have learned previously taught content within a specific lesson is a standard procedure. As a result, assessment is the systematic collecting and analysis of information about what students know, understand, and can perform in connection to achieving specific learning objectives. The instruments used to obtain such data should fulfill exact requirements that are subject to public examination, based on this essential aim [19].

Validity is an important aspect of TMT. It refers to the degree to which an instrument measures what it intends to measure [20]–[24]. Moreover, content validity, face validity, criterion-related validity, and construct validity are all types of validity [21]. They went on to say that content validity refers to how well the instrument analyzes the concept of interest [21], [24]. The characteristic or attribute of interest is measured by face validity. When assessing the association between scores and a test with reference to a specific criteria, criterion-related validity is tested, whereas construct validity analyzes the degree to which an instrument assesses the trait or theoretical construct that it is designed to measure. The form of validity used in this study is content validity, which is applied to TMT.

Content validity has been commonly held to be the most important type of validity needed for criterion-referenced measures [6], [24], [25]. It involves content relevance and content coverage [24], [26], [27]. Content relevance evaluates the extent to which the aspects of the ability to be assessed are actually tested by the task, indicating the requirement to specify the ability domain and the test method facets [28]. Content coverage addresses the extent to which the test tasks adequately demonstrate the performance in the target context, which may be achieved by randomly selecting representative samples [28].

TOS is a test blueprint which helps teachers to align lesson/learning objectives, instruction, content, learning activities and test items in assessments systematically [17], [29], [30]. TOS can be used in conjunction with lesson and unit planning to help teachers make a clear connection between planning, instruction, activity and assessment [17]. The primary goal of a TOS is to improve the validity of a teacher's evaluation in relation to a given assessment [7]. Teachers should utilize the TOS since it specifies not only the topic areas taught in class but also the performance objectives at each level of Bloom's Taxonomy's cognitive domain [30], [31].

TOS is important to develop content-valid tests. While assessment instruments, teachers should care about evidence based on the validity of test content and response process evidence, because such instruments form the basis of significant decision-making on students and programs [17], [29]. The former encompasses the extent to which a test assesses the ability it is designed to assess [32]. The latter emphasizes the kind of

abilities a student must demonstrate throughout instruction and evaluation activities. It connects the degree of thinking necessary for a specific assessment with the knowledge conveyed through instructional procedures and activities [32]. In-valid tests over- or under-representing important content areas at the suitable cognitive level would not be helpful to accountably measure what it is intended to measure and reach valid judgements about students' attainment of learning outcomes [24], [29], [31], [33] except when it precisely and accurately determines learner ability.

As a result, it is critical for a teacher to create his or her own TOS since it serves as a supplement to lesson and unit preparation, allowing teachers to draw a clear line between planning, instruction, and assessment [17]. Variables such as experience, age, and gender, in addition to the teacher's preparation, are thought to have an influence on the TMT. In other words, the researchers hypothesized the foregoing mentioned demographic variables are likely affect TMT content validity.

The TMT's validity may be impacted by number of factors, one of which is the teacher's years of experience in the field of education [34]. This is because the teacher administers tests and reflects on the results, giving him or her a greater grasp of how to build a legitimate test. Many studies have revealed that experienced teachers (those who have been teaching for more than ten years) create more valid tests on TMT sub-units than less experienced teachers (those who have been teaching for one to five years) [34]–[36]. Conversely, teachers with less experience (1-5 years) utilize more TMT alternative approaches than those who have been teaching for 31-35 years [37]. Additional empirical investigations with a focus on the influence of job experience on the TMT are needed to resolve this discrepancy in the results.

TMT is thought to be age insensitive. The young and the old have different reactions to the concerns. Senior teachers, on the other hand, may be more concerned about the growth of others than junior teachers. However, this assumption is not supported by the examined research. Subject topics, grade level, teaching experience, gender, and in-service assessment training were all investigated in the literature as factors impacting TMT [35], processes and techniques [37], However, to the best of the researchers' knowledge and the reviewed literature, no empirical investigation has investigated at the impacts of age on the TMT. The gender of the teacher is generally cited as a contributing element to the TMT. According to the data, female teachers outperformed male counterpart on the TMT sub-units [35].

Understanding the usage and building of the TOS by teachers is essential for creating meaningful and accurate assessments. Given the anticipated subject objectives, TOS is a blueprint that directs and gives instructions in interwoven areas of Bloom's Taxonomy. Students' disagreements over assessment outcomes necessitate a fresh attention on the TOS, which serves as a framework for teacher-assessment methods. Valid and reliable evaluation is the result of sufficient teacher comprehension of the TOS. Despite the paucity of TOS-related literature, recent surveys have indicated that most of primary school teachers have a poor comprehension of the TOS as a result of a lack of attending TOS-related workshops and training programs [38]–[40]. It is obvious from the reviewed literature that teacher knowledge of the TOS variable has received little attention. As a result, there was no empirical study that investigated the impacts of work experience, age, or gender on teacher knowledge in the literature. This emphasizes the study's substantial contribution to the assessment literature.

Given the preceding discussion and the scarcity of literature on the variables under consideration, this research examined the content validity of TMT. It also investigated teachers' understanding of the TOS. The effects of demographic variables on TMT and teacher understanding in Malaysian Chinese elementary schools are also examined in this study.

2. RESEARCH METHOD

2.1. Sample

The survey involved 381 teachers from 21 Chinese primary schools in Johor Bahru, Malaysia. The teachers were chosen randomly from each class group. The researchers used a two-stage cluster sampling approach. Based on school distributions, schools were chosen at random from each zone in the state of Johor. Teachers were chosen at random from each of the schools. Researchers used a hypothetical two-stage multi-cluster sampling method. The volunteers were chosen without regard to their gender, age, or job experience.

2.2. Instrument

The researchers employed a survey instrument to collect data on the factors under inquiry [38]. The instrument is divided into two pieces. Section one consisted of three questions in which respondents were asked to supply demographic information such as gender, age, and job experience. The second section of the TOS has 10 elements [17], [30], [38]. Items 1 through 4 assessed teachers' comprehension of the TOS, whereas Items 5 through 10 assessed the TMT's content validity. For teacher replies, a four-point Likert scale was used. The scale is interpreted as 1=strongly disagree, 2=disagree, 3=agree, and 4=strongly agree.

The data were interpreted using two steps of mean computations. The responses were initially defined using the mean score of each item [41]. The cumulative mean of the replies was then calculated using the midway point of the scales in the questionnaire [41]. Convention of mean score range were used [42]. Low scores were defined as those ranging from 1.50-2.50. Mean scores ranging from 2.51-3.50 were considered average, while mean scores ranging from 3.51-5.00 were considered high. Given the scale's structure, the greater the mean score, the better the teacher understanding and TMT. To analyze the data, descriptive and K-group MANOVA and t-test analyses were used.

3. RESULTS AND DISCUSSION

3.1. Respondents' characteristics

According to the frequency data, 159 (41.7%) of the participants were male teachers, while 222 (58.3%) were female teachers. The age distribution of the participants indicated that most of the participants were between the ages of 21 and 30 (n=132, 34.7%) and 31-40 (n=132, 34.7%). The 41–50-year age group was next (n=79, 20.7%). Furthermore, instructors over the age of 51 made up the smallest proportion of participants (n=38, 10.0%). The results also indicated that most of the participants (n=118, 31.0%) have being teaching for 1-5 years. This was followed by individuals with more than 16 years of teaching experience (n=96, or 25.2%). Participants with 11-15 years of teaching experience received a high score (n=86, 22.5%). Participants with 6-10 years of teaching experience (n=81, 21.3%) were the fewest. Table 1 presents the details of the participant.

Table 1. Frequency and percentages of participants' gender, age, and work experience

Variable		Frequency	%
Gender	Male	159	41.7
	Female	222	58.3
Total		381	100
Age	21-30	132	34.7
	31-40	132	34.7
	41-50	79	20.7
	51 and above	38	10.0
Total		381	100
Work experience	1-5	118	31.0
	6-10	81	21.3
	11-15	86	22.5
	16 and above	96	25.2
Total		381	100

The reliability of the collected data was then investigated. The results indicated that the teacher understanding component had an alpha coefficient of 0.71 and the TMT factor had an alpha coefficient of 0.80. Because the Cronbach's Alpha was greater than the cutoff value .70, this result demonstrated that the chosen instrument was reliable [24], [31], [43], [44]. As a result, the findings indicated that the items used in this research study were sufficient and interconnected.

After that, confirmatory factor analysis (CFA) was used to examine into the instrument's validity. The two-factor construct of the TMT and teacher understanding were assessed using a CFA with maximum likelihood (ML). The first measurement model had less-than-ideal fit statistics; χ^2 (26)=100.430, CFI=.91, TLI=.90, NFI=.89, RMSEA=0.08 and χ^2/df=3.963. Although the values of RMSEA, χ^2/df, TLI, and CFI were within acceptable bounds, the model's overall fit was less than desired. However, after modifying the model and eliminating Item 5, which had a low factor loading and large cross loading issue, the model showed a better fit based on the parameters of modification indices (MIs); χ^2 (19)=52.126, CFI=.96, TLI=.94, NFI=.94, RMSEA=0.06 and χ^2/df=2.743. The average variance extracted (AVE) and the results of the composite reliability index (CRI) were likewise positive, with AVEs of. 82, .51 and CRIs of .83, .81 for teacher understanding and TMT respectively [45]–[47].

3.2. Teacher understanding of the TOS

The average mean scores were found in the findings of the individual item means. Item 6 had a mean score of M=3.33, SD=.552 and Item 9 had a mean score of M=3.17, SD=.513. This indicates that most of the participants have taken TOS training courses. As a result, they have the capacity and expertise to create a TOS for the courses that they teach. This finding contradicted [38], [39] who concluded that many teachers are unclear of how to construct valid assessments due to their inability to use the TOS.

Interestingly, the overall mean score demonstrated that the teachers' grasp of the TOS was within an average mean score zone, as the cumulative mean was M=2.99, SD=0.49, which was greater than the 2.50

cutoff mark. This study also contradicted with the previous findings that discovered the sampled teachers in the selected Chinese primary schools had a poor understanding of the usage of TOS [38]. Table 2 demonstrates the details.

Table 2. Mean, standard deviation and cumulative mean score of the items examining the teacher understanding of the TOS

No.	Item	N	Range	Min	Max	Mean	SD
1	I know about the TOS	381	3	1	4	3.08	.578
2	I can build a TOS for the subject that I involved	381	3	1	4	2.96	.629
3	I discuss with my colleagues how to build the TOS instruments for assessment	381	3	1	4	3.19	.652
4	I attended courses related to the building of the TOS instruments for an assessment	381	3	1	4	2.73	.820
			Cumulative mean: 2.99				

3.3. The validity of teacher-made tests

The findings revealed that practically all elements pertaining to the validity of TMT had exceeded the middle point and fell inside the average mean score zone. Item 6 (I am clear with the contents of the subject which I involved in) M=3.33, SD=.552 had the highest mean score, which was in the higher quartile of the average mean score. This indicates that most of the participants had a clear notion of how to relate learning content with the evaluation instrument. This conclusion matched with previous studies [30], [31] that proposed instructors adopt the TOS because it defines not only the topic areas taught in the classroom but also the performance objectives at each level of Bloom's Taxonomy's cognitive domain. Teachers who do not follow conventional construction principles when developing a test are more likely to fall short of gauging real student achievement and have poor content validity [30], [48].

The results also revealed that the TMT's overall mean score was greater than the lower average point score, surpassing the middle point of 2.51 with a cumulative mean of 3.28. This score suggested that TMT in the selected Chinese primary schools demonstrated both response process and test content evidence. Thus, despite the fact that TMT addresses classroom topic given a specific amount of relevance (i.e., students' reactions), it connects the content with instructional or guideline driven by the TOS rules. This conclusion is consistent with the previous findings [38] stated that Chinese primary school teachers' TMT methods had both process evidence and evidence-based validity in the studied schools. Table 3 presents the details.

Table 3. Mean, standard deviation and cumulative mean of the items examining the content validity of the TMT

No.	Item	N	Range	Min	Max	Mean	SD
6	I am clear with the contents of the subject which I involved	381	3	1	4	3.33	.552
7	I am clear with the format of assessment for the subject which I involved	381	3	1	4	3.30	.608
8	I am clear with the time allocated for answering the assessment of the subject which I involved	381	3	1	4	3.25	.558
9	I am able to categorize the items of assessment based on their level of difficulties	381	3	1	4	3.17	.513
10	I am clear with the level of understanding of my students for the subject which I involved	381	3	1	4	3.31	.541
			Cumulative mean: 3.28				

3.4. Effects of teacher demographic profiles

3.4.1. Teacher-made assessment with work experience

On two dependent variables, teacher understanding and TMT, a K-group MANOVA analysis was performed. Work experience was used as an independent variable. Work experience was presented in the following way as 1-5 years n=118, 6-10 years n=81, 11-15 years n=86 and above 16 years n=96. A statistical significant Box's M test (Box's M=21.281, p<0.012) Pillai's trace was used to analyze multivariate effects since it revealed different variance-covariance matrices of the dependent measures across degrees of work experience, Pillai's=.062, F(6,754)=4.05, p<.001, partial η^2=.031. To find the locus of the statistically significant multivariate impact, we used univariate ANOVAs on each dependent measure independently. Table 4 shows that teacher understanding was strongly influenced by work experience, F(3,377)=7.03, p<.001, partial η^2=.053.

When there is heterogeneity of variance, Tamhane's post hoc tests are appropriate. Table 5 shows teachers who have been teaching for 11-15 years and those who have been teaching for more than 16 years (M=3.137, SD=0.431 and M=3.138, SD=0.413) had a considerably higher understanding score than their counterparts with 6-10 years of experience in the classroom (M=2.935, SD=0.467) and teachers having only 1 to 5 years of experience in the classroom (M=2.853. SD=0.557). For TMT, there were no statistically significant respondent work experience impacts, $F(3, 377)=2.18$, $p>.091$, partial $\eta^2=.017$. Because of the heterogeneity of variance-covariance matrices reported in this investigation, these results should be interpreted with caution.

Table 4. Multivariate and univariate analysis of variance for teachers understanding and TMT measures

Source	Multivariate	Univariate	
	Fa	Teacher understanding	TMT
Teacher work experience	.062*	7.062*	2.172
SME		3.647	4.240

Note: multivariate F ratios were generated from Pillai's trace; * $p < .05$; ** $p < .01$
a. Multivariate df=6.754; b. Univariate df=3.377

Table 5. Mean and standard deviation for measures of teacher understanding and TMT as a function of teacher work experience

Group	Teacher understanding		TMT	
	M	SD	M	SD
1-5 years	2.853h	0.557	2.865	0.546
6-10 years	2.935f	0.467	2.974	0.466
11-15 years	3.137b	0.431	3.987	0.387
Above 16 years	3.138a	0.413	3.237	0.415

Note: Tamhane post hoc test revealed that means with distinct superscripted letters differed significantly at the .05 level

On the dependent measures of teacher understanding and TMT, another K-group MANOVA analysis was conducted. Age was used as an independent variable. The age group ranged from 21-30 years old n=132, 31-40 years old n=132, 41-50 years old n=79 and above 51 years n=38. A statistical significant Box's M test (Box's M=22.619, $p<0.008$) Pillai's trace was used to analyze multivariate effects since it revealed different variance-covariance matrices of the dependent variables across age levels, Pillai's trace=.062, $F(6,754)=3.991$, $p<.001$, partial $\eta^2=.031$. To find the locus of the statistically significant multivariate impact, we used univariate ANOVAs on each dependent measure independently. The findings imply that teacher understanding was strongly influenced by age, $F(3,377)=7.576$, $p<.001$, partial $\eta^2=.057$.

Tamhane's post hoc test was explored due to the heterogeneity of variance. Senior teachers aged 51 and up are the most effective (M=3.237, SD=0.415) had significantly greater understanding of validity practices than did teachers aged 41-50 years (M=3.087, SD=0.387), teachers aged 31-40 years (M=2.974, SD=0.466), and teachers aged 21-30 years (M=2.865, SD=0.546). For TMT, there were no statistically significant respondent age effects, $F(3,377)=1.981$, $p>.116$, partial $\eta^2=.016$. The insignificant TMT finding in this study contradicts previous research [34]–[36] that suggested that more experienced teachers produce more valid TMT sub-unit assessments than less experienced teachers. Because of the heterogeneity of variance-covariance matrices reported in this investigation, these results should be regarded with caution and the details are presented in Tables 6 and 7.

Table 6. Multivariate and univariate analysis of variance for teachers understanding and TMT measures

	Multivariate	Univariate	
Source	Fa	Teacher understanding	TMT
age	.062*	7.576*	1.981
SME		3.632	4.246

Table 7. Means and standard deviations for measures of teacher understanding and TMT as a function of teacher age

	Teacher understanding		TMT	
Group	M	SD	M	SD
21-30 years	2.865h	0.546	3.246	0.427
31-40 years	2.974d	0.466	3.243	0.373
41-50 years	3.087b	0.387	3.293	0.451
Above 51 years	3.237a	0.415	3.416	0.414

Finally, on teacher understanding and TMT, a two-group between subjects MANOVA was performed. The gender of Chinese elementary school teachers was the independent variable. The study included 158 males (41.5%) and 223 females (58.5%) participants. Gender had no effect on the composite dependent variable, according to Wilks' criterion (Wilks' λ, $F(2,378)=1.761^{a}$, $p>.173$, partial η =.009). The locus of the statistically significant multivariate impact was subsequently investigated using univariate ANOVAs on each dependent variable independently. According to Table 8, gender had no significant effect on the two dependent measures of teacher understanding and TMT, $F(1, 379)=1.538$, $p>.216$, partial η=.004, and TMT $F(1, 379)=3.154$, $p>.077$, partial η =.008. The locus of the statistically significant multivariate impact was subsequently investigated using univariate ANOVAs on each dependent variable independently. According to Table 8, gender had no significant effect on the two dependent measures of teacher understanding and TMT.

According to Table 9, teachers in the selected Chinese primary schools had an average understanding of the use of TOS and were using it to assess student achievement. This is because most instructors participated in TOS-related teacher professional development courses and so gained the skills to create a TOS for the subjects they teach throughout the selected schools. As a result, they have the skills and expertise to create a TOS for all of the courses they teach.

Table 8. Multivariate and univariate analysis of variance for teachers understanding and TMT measures

	Multivariate	Univariate	
Source	*df* F^{a}	Teacher understanding	TMT
F Ratios for gender	11.761	1.538	3.154
SME		5.866	13.420

Note: multivariate *F* ratios were generated from Wilks' criterion.
a. Multivariate *df*=2.378
b. Univariate *df*=1.379
* *p*<.01

Table 9. Means and standard deviations for measures of teacher understanding and TMT as a function of gender

	Teacher understanding	TMT	
Group	M	SD	M
Male	2.949	0.471	3.227
Female	3.012	0.500	3.304

The study also found that instructors in the selected schools had a clear understanding of how to use process-based and evidence-based methods to link learning material with the evaluation instrument. As a result, the TMT's content validity was upheld in the majority of their evaluation processes. It is worth noting that TMT's validity was determined by instructors' alignments of their assessments with TOS, which took into account both procedural and evidence-based approaches. This indicates that the TOS is used by the majority of instructors when creating instruments for formative learning performance evaluation. The results also show that while evaluating their students' performance, teachers stick to the standards, methods, and tactics acquired in teacher professional training rather of taking into account their students' talents.

The findings revealed that none of the demographic variables studied (job experience, age, and gender) had a significant impact on TMT practices across the investigated schools. As a result, the TMT generated by the most experienced and less experienced teachers has the same content validity. Furthermore, instructors' gender has little impact on how they build and apply TMT in evaluating student academic achievement. Teacher gender was likewise shown to have a negligible impact on teacher comprehension. Similarly, the age of the teachers has little bearing on the outcome. As a result, both elderly and new instructors approach TMT activities in the same way. This may be due to the fact that all instructors went through the same professional teacher training program.

Given that an average TMT and teachers' understanding of the TOS were seen in the activities of teachers from selected primary schools, this study gives recommendation for future improvements. First, while the data suggested that teachers used the TOS on average when creating assessment instruments, additional increases in teacher understanding of the TOS and the benefits of applying it carefully are urged. As a result, it is critical for school administrators or principals to raise teacher understanding of the value of TOS and to provide skill follow-up training on how to use TOS to enhance teacher-assessment techniques. This may be accomplished through organizing short-term training programs with TOS specialists in order to build constructive course-based TOS, as well as in-school training and peer mentoring, particularly amongst experienced and beginner instructors.

Second, despite the fact that the majority of teachers took TOS-related courses and so gained skill in creating a TOS for legitimate TMT, there is still space for improvement. Training or seminars will be one option for keeping teachers up to date on the newest advancements in TMT practices. Information would give high-quality data as a prerequisite for making decisions and judgments on TMT practices related to student success of certain learning objectives.

Finally, teachers and administrators should be aware that every classroom practice and standardized assessment contains a mistake. To guarantee that TMT processes are fair and lead to a valid conclusion with the least amount of error, several assessment techniques should be used. As a result, teachers should not rely just on TMT techniques, but rather use a combination of ways to collect accurate and high-quality data on students' performance.

4. CONCLUSION

The findings concluded that in the selected schools, teacher job experience aided teacher grasp of the TOS. Teachers who have had a lot of experience with TOS have become masters at knowing how it works and how to use its processes in ongoing formative assessment. Furthermore, the data revealed that older instructors had a better comprehension of how to use TOS than novice teachers. This might be due to the fact that senior teachers were given more opportunity to evaluate TOS assessment-related components than their peers.

A cross-sectional survey approach was used to sample elementary school instructors who took part in the study. A hybrid method approach, using the strengths of both qualitative and quantitative methods, would better address the depth and breadth of TMT and teacher understanding throughout the studied schools. Another drawback was the sample size. Because the study only included 381 instructors from 21 Chinese elementary schools in Johor, the results should not be applied to other Chinese primary schools in the country. In addition, future research should include additional schools from other states. Furthermore, the study's participants were all Chinese primary school teachers. As a result, no Indian or Malay elementary school teachers were included, which would be considered if the findings were to be generalized throughout primary schools and comprehensive information on the TMT and teacher understanding of the TOS was to be captured across Malaysia.

ACKNOWLEDGEMENTS

This research was an extension of a previously published paper in 2015 entitled Validity of Teacher-Made Assessment: A Table of Specification Approach.

REFERENCES

[1] R. N. Amelia, A. R. P. Sari, and S. R. D. Astuti, "Assessment of Chemistry Learning: How is The Quality of The Tests Made by The Teacher?" *Journal of Educational Chemistry (JEC)*, vol. 3, no. 1, pp. 11–22, 2021, doi: 10.21580/jec.2021.3.1.6582.

[2] P. A. Griswold, "Assessing Relevance and Reliability To Improve the Quality of Teacher-Made Tests," *NASSP Bulletin*, vol. 74, no. 523, pp. 18–24, 1990, doi: 10.1177/019263659007452305.

[3] S. A. Karim, S. Sudiro, and S. Sakinah, "Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test," *Journal of English Education, Literature, and Culture*, vol. 6, no. 2, pp. 256–269, 2021.

[4] L. Hakim and Irhamsyah, "The analysis of the teacher-made test for senior high school at state senior high school 1 Kutacane, Aceh Tongogara," *Jurnal Ilmiah DIDAKTIKA*, vol. 21, no. 1, pp. 10–20, 2020.

[5] H. Aliah, "The Analysis of Junior High School Teacher-Made Tests for the Students in Enrekang," *FOSTER: Journal of English Language Teaching and Learning*, vol. 1, no. 2, pp. 122–138, 2020, doi: 10.24256/foster-jelt.v1i2.14.

[6] J. T. Akinboboye and M. A. Ayanwale, "Bloom Taxonomy Usage and Psychometric Analysis of Classroom Teacher Made Test," *African Multidisciplinary Journal of Development (AMJD)*, vol. 10, no. 1, p. 2021, 2021.

[7] P. J. Black, *Testing: Friend or foe: Theory and practice of assessment and testing*. Taylor and Francis e-Library, 2002.

[8] Ministry of Education Malaysia, *Malaysia Education Blueprint 2013-2025 (Preschool to Post-Secondary Education)*. Kementerian Pendidikan Malaysia, Putrajaya Malaysia, 2013.

[9] C. Y. Fook and G. K. Sidhu, *School-Based Assessment among ESL Teachers in Malaysian Secondary Schools*. Shah Alam, Selangor: Faculty of Education, Universiti Teknologi MARA, 2006.

[10] S. H. V. Al-Hudawi, R. L. S. Fong, M. B. Musah, and L. M. Tahir, "The actualization of the Malaysian national education philosophy in secondary schools: Student and teacher perspectives," *International Education Studies*, vol. 7, no. 4, pp. 57–68, 2014, doi: 10.5539/ies.v7n4p57.

[11] F. Majid, "School-Based Assessment in Malaysian Schools: The Concerns of the English Teachers.," *Online Submission*, vol. 3, no. 10, pp. 393–402, 2011.

[12] I. J. Lehmann and W. A. Mehrens, "Using teacher-made measurement devices.," *NASSP Bulletin*, vol. 71, no. 496, pp. 36–44, 1987, doi: 10.1177/019263658707149605.

[13] B. E. Omoruan, "Valid Teacher-Made Tests: It's Implication on Students' Achievement in Business Education," *ATBU, Journal of Science, Technology & Education (JOSTE)*, vol. 6, no. 2, pp. 65–77, 2018.

[14] E. Feron, T. Schils, and B. ter Weel, "Does the Teacher Beat the Test? The Value of the Teacher's Assessment in Predicting Student Ability," *Economist (Netherlands)*, vol. 164, no. 4, pp. 391–418, 2016, doi: 10.1007/s10645-016-9278-z.

[15] K. Kinyua and L. O. Okunya, "Validity and Reliability of Teacher-Made Tests: Case Study of Year 11 Physics in Nyahururu District of Kenya," *African Educational Research Journal*, vol. 2, no. 2, pp. 61–71, 2014.
[16] N.-G. Ugwu and S. O. Mkpuma, "Ensuring Quality in Education: Validity of Teacher-made Language Tests in Secondary Schools in Ebonyi State," *American Journal of Educational Research*, vol. 7, no. 7, pp. 518–523, 2019, doi: 10.12691/education-7-7-12.
[17] H. Fives and N. Didonato-Barnes, "Classroom test construction: the power of table of contents," *Practical Assessment, Research, and Evaluation*, vol. 18, no. 3, pp. 1–7, 2013, doi: 10.7275/cztt-7109.
[18] A. Stevenson, *Oxford Dictionary of English*, 3rd ed. Oxford Uni. UK, 2010.
[19] D. Pratt, *Curriculum: Design and development*. New York: Harcourt Brace Jovanovich, 1980.
[20] D. Furwana, "Validity and Reliability of Teacher-Made English Summative Test at Second Grade of Vocational High School 2 Palopo," *Language Circle: Journal of Language and Literature*, vol. 13, no. 2, 2019, doi: 10.15294/lc.v13i2.18967.
[21] M. J. Lambert *et al*., "The reliability and validity of the outcome questionnaire," *Clinical Psychology and Psychotherapy*, vol. 3, no. 4, pp. 249–258, 1996, doi: 10.1002/(sici)1099-0879(199612)3:4<249::aid-cpp106>3.0.co;2-s.
[22] E. J. Pedhazur and L. P. Schemelkin, "Measurement, Design, and Analysis: An Integrated Approach," in *Measurement, Design, and Analysis: An Integrated Approach*, New York: Psychology Press; Taylor & Francis Group, 1991, pp. 1689–1699.
[23] L. G. Grimm and P. R. Yarnold, *Reading and Understanding Multivariate Statistic*. Washington, DC: American Psychological Association, 2006.
[24] A. Setiabudi, Mulyadi, and H. Puspita, "An Analysis of Validity and Reliability of a Teacher-Made Test (Case Study at XI Grade of SMA N 6 Bengkulu)," *Journal of English Education and Teaching (JEET)*, vol. 3, no. 4, pp. 522–532, 2019.
[25] R. L. Linn, "Issues of Validity for Criterion-Referenced Measures," *Applied Psychological Measurement*, vol. 4, no. 4, pp. 547–561, 1980, doi: 10.1177/014662168000400407.
[26] L. F. Bachman, *Fundamental Considerations in Language Testing*. New York: Oxford University Press, 1990.
[27] G. Özdemir, A. Özdemir, and P. D. S. Gelbal, "Determination of cyber accessibility of teacher made tests/exams," *International Journal of Assessment Tools in Education*, vol. 8, no. 3, pp. 553–569, 2021, doi: 10.21449/ijate.780556.
[28] Y. Liao, "Issues of Validity and Reliability in Second Language Performance Assessment," *TESOL & Applied Linguistics*, vol. 4, no. 2, pp. 2–5, 2002.
[29] G. Eweda, Z. A. Bukhary, and O. Hamed, "Quality assurance of test blueprinting Enhanced Reader," *Journal of Professional Nursing*, vol. 36, pp. 166–170, 2020, doi: 10.1016/j.profnurs.2019.09.001.
[30] C. E. Notar, D. C. Zuelke, J. D. Wilson, and B. D. Yunker, "The Table of Specifications: Insuring Accountability in Teacher Made Tests," *Journal of Instructional Psychology*, vol. 31, no. 2, pp. 115–129, 2004.
[31] H. Mulyani, H. Tanuatmodjo, and R. Iskandar, "Quality analysis of teacher-made tests in financial accounting subject at vocational high schools," *Jurnal Pendidikan Vokasi*, vol. 10, no. 1, pp. 1–9, 2020, doi: 10.21831/jpv.v10i1.29382.
[32] S. Wolming and C. Wikström, "The concept of validity in theory and practice," *Assessment in Education: Principles, Policy and Practice*, vol. 17, no. 2, pp. 117–132, 2010, doi: 10.1080/09695941003693856.
[33] D. M. Billings and J. A. Halstead, *Teaching in nursing: A guide for faculty*. St. Louis, Mo: Elsevier/Saunders, 2016.
[34] L. O. Kiragu Kinyua, "Validity and reliability of teacher-made tests: Case study of year 11 physics in Nyahururu District of Kenya," *African Educational Research Journal*, vol. 2, no. 2, pp. 61–71, 2014.
[35] H. Alkharusi, "Teachers' classroom assessment skills: Influence of gender, subject area, grade level, teaching experience and in-service assessment training," *Journal of Turkish Science Education*, vol. 8, no. 2, pp. 39–48, 2011.
[36] C. Magno, "The Profile of Teacher-Made Test Construction of the Professors of University of Perpetual Help Laguna," *UPHL Institutional Journal*, vol. 1, no. 1, pp. 48–55, 2003.
[37] C. A. Mertler, "Classroom Assessment Practices of Ohio Teachers," *The Annual Meeting of the Mid-Western Educational Research Association*, 1998.
[38] L. M. Ing, M. B. Musah, S. H. V. Al-Hudawi, L. M. Tahir, and N. M. Kamil, "Validity of teacher-made assessment: A table of specification approach," *Asian Social Science*, vol. 11, no. 5, pp. 193–200, 2015, doi: 10.5539/ass.v11n5p193.
[39] T. Cooper, A. Pittman, and S. Womack, "Using reliability, validity, and item analysis to evaluate a teacher-developed test in international business," *Evaluation and Testing Research Article*, pp. 1–11, 2014.
[40] E. Hartell and G. J. Strimel, "What is it called and how does it work: examining content validity and item design of teacher-made tests," *International Journal of Technology and Design Education*, vol. 29, no. 4, pp. 781–802, 2019, doi: 10.1007/s10798-018-9463-2.
[41] M. N. A. Ghafar, "Reka Bentuk Tnijauan Soal Selidik Pendidikan, Johor, Malaysia," in *Statistical Methods in Education and Psychology*, Johor Bahru: UTM Press, 1996.
[42] G. V. Glass and K. D. Hopkins, *Statistical Methods in Education and Psychology*. New Jersey: Prentice -Hall, Inc. Inglewood Cliffs, 1996.
[43] R. A. Santos, "Cronbach's Alpha: A Tool for Assessing the Reliability of Scales," *Journal of Extension*, vol. 37, no. 2, 1999.
[44] M. Tavakol and R. Dennick, "Making sense of Cronbach's alpha," *International Journal of Medical Education*, vol. 2, pp. 53–55, 2011, doi: 10.5116/ijme.4dfb.8dfd.
[45] C. Fornell and D. F. Larcker, "Evaluating Structural Equation Models with Unobservable Variables and Measurement Error," *Journal of Marketing Research*, vol. 18, no. 1, p. 39, 1981, doi: 10.2307/3151312.
[46] J. F. Hair Jr., W. C. Black, B. J. Babin, and R. E. Anderson, *Multivariate data analysis*, 7th ed. New Jersey: Pearson Prentice Hall, 2015.
[47] R. P. Bagozzi, "Evaluating structural equation models with unobservable variables and measurement error: a comment," *Journal of Marketing Research*, vol. 18, no. 3, pp. 375–381, 1981.
[48] A. Sugianto, "Item Analysis of English Summative Tests: EFL Teacher-Made Test," *Indonesian EFL Research & Practice*, vol. 1, no. 1, pp. 35–54, 2020.

BIOGRAPHIES OF AUTHORS

Mohammed Borhandden Musah is an Assistant Professor at the Education Studies Department, Bahrain Teachers College, University of Bahrain, Kingdom of Bahrain. His work focuses on the management of higher learning institutions, quality management, followership, professional learning communities, teacher leadership and workforce performance. He can be contacted at email: mmusah@uob.edu.bh.

Lokman Mohd Tahir is an Associate Professor at the School of Education, Faculty of Social Sciences and Humanities. His research work focuses on the principalship, School Leadership, educational policy, teacher leadership, ethics in education, and professional learning communities in education. He can be contacted at email: p-lokman@utm.my.

Shafeeq Hussain Vazhathodi Al-Hudawi is an Assistant Professor at Yanbu Industrial College, Yanbu, Saudi Arabia. His work focuses on curriculum and instruction, alignment of curriculum, alignment of assessment, significant learning, learning communities, academic and research writing and preparing measurement tools using Quantitative Skills. He can be contacted at email: hussains@rcyci.edu.sa.

Mohammad Issah is an Assistant Professor at the Education Studies Department, Bahrain Teachers College, University of Bahrain, Kingdom of Bahrain. His work focuses on Educational Leadership, policy and organizational change, and leading change with emotional intelligence. He can be contacted at email: missah@uob.edu.bh.

Abdul Rawuf Hussein Ph.D in counselling psychology and counsellor education. Previously, he worked with UNHCR on promoting psychosocial adjustment of Somali youth refugees in Malaysia. His work focuses on refugees, mental health, emotional intelligence, students' counselling, and group guidance. His current research project focuses on the refugee post-migration emotional health in Finland, where he works with a minority research group in Åbo Akademi, Finland. He can be contacted at email: ahussein@abo.fi.

Mikail Ibrahim is an Associate Professor at Faculty of Major Languages Studies, Universiti Sains Islam Malaysia, Nilai, Malaysia. His focuses on educational psychology, Motivation, human resource management, Arabic as a second language, quality management, teaching strategies. He can be contacted at email: mikail@usim.edu.my.