JOURNAL OF APPLIED MEASUREMENT, *16*(1), 60-75

Measuring Teaching Assistants' Efficacy using the Rasch Model

Zi Yan

Chun Wai Lum

Rick Tze Leung Lui

Steven Sing Wa Chu
The Hong Kong Institute of Education, Hong Kong

Ming Lui
Hong Kong Baptist University, Hong Kong

Teaching assistants (TAs) play an influential role in primary and secondary schools. But there is an absence in literature about the TA's efficacy, and to date no instrument is available for measuring TA's efficacy. The present study aims to develop and validate a scale (Teaching Assistant Efficacy Scale, TAES) for measuring TA's efficacy on identified capabilities. A total of 531 teaching assistants from Hong Kong schools participated in the survey. The multidimensional Rasch model was used to analyse the data. The results revealed that a 5-dimension structure of TA's efficacy was supported. The final 30-item version of TAES assesses TA's efficacy on learning support, teaching support, behaviour management, cooperation, and administrative support. The Rasch reliabilities for all five dimensions were around 0.90. The 6-category response structure worked well for the scale. Further research was recommended to validate and test the robustness of the TAES both in Hong Kong and elsewhere.

Requests for reprints should be sent to Zi Yan, Department of Curriculum and Instruction, The Hong Kong Institute of Education, 10 Lo Ping Road, Tai Po, Hong Kong. e-mail: zyan@ied.edu.hk.

Background

Increased number of teaching assistants (TAs) has been deployed in primary and secondary schools to support teachers and students (Blatchford, Bassett, Brown, and Webster, 2009) towards the ends of raising students' performance and reducing teachers' unnecessary workload. In England, the number of TAs or TA equivalent staff deployed in schools increased from 60,600 in 1997 to 176,900 in 2008 (DCSF, 2008). In Hong Kong, although there are no official statistics, rough estimates based on the authors' experience suggest that around 10% of staff in schools are working in the role of TAs or TA equivalent. There is no doubt that TAs play an important role in schooling that is likely to have substantial impact on students' learning outcomes and teacher's practice (Blatchford, Bassett, Brown, Martin, Russell, and Webster, 2006). However, there is very limited research about TAs working in primary and secondary schools. Given that teacher efficacy is seemingly a popular topic in educational research and has been shown to be associated with many student outcomes such as achievement (Bozman, 2012; Deepa, 2007), motivation (Nelson, 2008), as well as teachers' teaching practice (Allinder, 1994; Gibson and Dembo, 1984), professional commitment (Coladarci, 1992), and attitude (Soodak, Podell, and Lehman, 1998), TA's efficacy might be a topic which deserves the foremost research attention in this field.

According to Bandura's socio-cognitive theory, efficacy refers to "beliefs in one's capacity to organize and execute the course of action required to produce given attainments" (1997, p.3). Various instruments (e.g., Gibson and Dembo, 1984; Tschannen-Moran and Woolfolk Hoy, 2001) have been developed and used widely to assess teachers' efficacy. There are also some studies examining the teaching efficacy of graduate TAs working in universities (e.g., Kim, 2009; Komarraju, 2008; McCrea, 2006; Mills, 2011). In those research studies, the instruments used were either originally designed for use with teachers, such as the Teacher Efficacy Scale (TES) (Gibson and Dembo, 1984), and the Teachers' Sense of Efficacy Scale (TSES) (Tschannen-Moran and Woolfolk Hoy, 2001), or specifically for use with graduate TAs working in universities but not in primary and secondary schools, e.g., the Self-Efficacy Toward Teaching Inventory (SETI) (Tollerud, 1990).

Since efficacy is a context-, subject-, and task-specific construct (Bandura, 1997; Chan, 2008; Tschannen-Moran and Woolfolk Hoy, 2001), efficacy scales must be tailor-made to activity domains and assess the many aspects of efficacy with regard to the specific domains (Bandura, 2006). Therefore, although it might be possible to use teacher or graduate TA efficacy scales to assess the efficacy of TAs working in primary and secondary schools due to the similarity of their work related to teaching, it is more defensible to use an efficacy scale which captures the unique contextual features of TA's work and the skills required of TAs to be successful in schools. This study attempts to fill in this gap in literature by developing an efficacy scale for use with TAs work in primary and secondary schools in Hong Kong.

Theoretical foundation of the Teaching Assistant Efficacy Scale (TAES)

The theoretical model of teacher efficacy proposed by Tschannen-Moran, Woolfolk Hoy, and Hoy (1998) was adopted in the present study to develop the instrument for assessing TA's efficacy. In this model, teacher's efficacy derives from the interaction of two components, namely, analysis of teaching task and assessment of teaching competence. Analysis of teaching task refers to examining the task in terms of the resources and constraints in particular teaching contexts. Assessment of personal teaching competence requires examination of the personal characteristics related to teaching, including skills, subject knowledge, and personality. Under this framework, two implications were posed for construction of TA efficacy scale items. First, the items in the instrument should invite respondents to assess personal competence in carrying out

particular TA tasks, taking into consideration the working contexts. Second, the instrument should assess a broad range of activities and tasks relevant to important aspects of TA's work. Earlier researchers (e.g., Pintrich and Schunk, 1996; Tschannen-Moran and Woolfolk Hoy, 2001) have argued that it is quite challenging to determine the optimal level of specificity in efficacy assessment. Neither extremely general nor extremely specific items are desirable. The former might fail to provide clear and meaningful efficacy assessments, and the later is likely to reduce the practicality of the scale.

Therefore, an important basis for the scale construction is to determine the task domains of TA's daily work in schools. Blatchford et al. (2009) reported that the work of TA (and equivalent) staff is usually related to providing learning support to students, assisting teachers, as well as supporting the school more generally. In order to find out the main task domains for TAs in Hong Kong primary and secondary schools—and in accordance with the guidelines proposed by Bandura (2006) to construct an efficacy scale—focus group interviews with in-service TAs and experts with experience in TA training were conducted to inform item construction. The results of content analysis of interviewees' responses showed that TAs working in Hong Kong schools have responsibilities in five domains: learning support, teaching support, behaviour management, cooperation, and administrative support. Learning support refers to direct support to students' learning provided by TAs without teacher involvement. Teaching support is the support of teachers' instruction. It could be in the form of replacement teaching or providing assistance during the teachers' teaching. Behaviour management refers to TA's responsibilities in managing students' behaviour in and out of classroom. Cooperation refers to the obligation TAs have to liaise with internal (e.g., professionals, non-teaching staff, etc.) and external parties (e.g., community, parents, etc.) in order to help students. Administrative support refers to the range of clerical duties TAs performs to facilitate the school's function.

None of above-mentioned efficacy scales (i.e., TES, TSES, and SETI) covers all the identified domains of TA's work. TES consists of two subscales assessing personal teaching efficacy and general teaching efficacy respectively; TSES is a measure of teachers' efficacy in student engagement, instructional strategies, and classroom management; SETI investigates efficacy in course preparation, instructor behaviour, materials, evaluation and examination, and clinical skills training. These available instruments focus on teaching or teaching-related aspects which, to some extent, overlap three domains of TA's work—learning support, teaching support, and behaviour management. The other two domains (i.e., cooperation and administrative support) seem unique to TAs in primary and secondary schools. Even though the responsibilities in these two domains may be shared by teachers and graduate TAs in practice, they have not been included into research agenda related to efficacy.

Rasch model

The true-score model (TSM) has been routinely used to validate psychometric scales in social sciences. However, there are inherent weaknesses in this approach. First and foremost, analytical techniques in TSM, including factor analysis, require linear, interval scale data input (Wright, 1997). However, raw data collected through Likert-type scales are ordinal, but not interval data since the categories of Likert-type scales indicate only ordering without any proportional levels of meaning. Applying parametric statistics on ordinal data can lead to misleading results (Bond and Fox, 2007; Wright, 1997). Second, as TSM uses total score to indicate the levels of respondent's abilities, the estimates of person ability are item-dependent (e.g., person ability estimates appear high when the items are easy but low when items are difficult even though the person ability actually remains invariant) and the estimates of item difficulty are similarly sample-dependent (e.g., item difficulty estimates appear high when the respondents' competence is low but low when the respondents' competence is high; even though the item, itself, remains invariant).

The Rasch model (Rasch, 1960) overcomes the above-mentioned problems and provides a sound framework for objective measurement. Rasch model can converts ordinal data into interval measures by employing the logarithmic transformation to calculate log-odds (logit) if the data fits the model. Such interval measures, transformed from ordered category responses, facilitate objective and linear measurement (Linacre, 2006a). The estimates of person ability are independent from the item difficulty and the estimates of item difficulty are independent from the sample. Furthermore, the Rasch analysis prevails over TSM by calibrating the persons and items simultaneously onto a same metric (Bond and Fox, 2007; Wright, 1992). Person measures are placed from low to high and item difficulties are placed from easy to hard on an ordered trait continuum and direct comparisons between person measures and item difficulties can be easily conducted, based on their locations on that continuum.

Rasch model has been increasingly used to analyse assessment data (Baird, 2012; Barkaoui, 2011; MacMillan, 2000) as well as to examine psychometric properties of instruments (e.g., Kyriakides, Kaloyirou, and Lindsay, 2006; Muís, Winne, and Edwards, 2009; Yan and Mok, 2012; Yan and Coniam, 2013) and the advantages of Rasch model were echoed. For example, Muís et al. (2009) undertook traditional and Rasch analyses on the psychometric properties of two instruments that are widely used in educational research settings—the Achievement Goals Questionnaire (AGQ) and the Patterns of Adaptive Learning Scale (PALS)—and found that, although traditional analyses based on raw scores showed the AGQ and PALS to be reliable and valid instruments, subsequent Rasch analysis revealed several problems which had been neglected in the traditional psychometric analysis but had important empirical implications. These included mismatch between distribution of respondent abilities and item difficulties, low person reliability, and dysfunction of lower portion of the response scale. The Rasch model also proved to be a sound methodological framework for developing or validating efficacy scales, such as Safety Efficacy of Domestic Food-Handling Practices Scale (Fischer, Frewer, and Nauta, 2006), Self-efficacy/Social Support for Activity for persons with Intellectual Disability (Lee, Peterson, and Dixon, 2010), The Career Decision Self-efficacy Scale (Nam, Yang, and Lee, 2011), Korean version of Psychomotor Self-efficacy (Zhu and Kang, 1994), and the Memory Self-efficacy Scale (Zelinski and Gilewski, 2004). Consequently, the Rasch model appears to be a promising tool for developing a scale for measuring TA's efficacy from both methodological and content-related perspectives.

Measurement should focus on only one latent trait at a time (Bond and Fox, 2007) and, consequently, unidimensionality is a basic requirement of standard Rasch analysis. However, if a scale contains several unidimensional subscales, a multidimensional Rasch model (Adams, Wilson, and Wang, 1997) may be appropriate as this approach takes into account the correlations between different but related unidimensional latent traits. A multidimensional model simultaneously calibrates all the dimensions and, further, increases measurement precision by making use of the correlations between these dimensions (Wang, Chen, and Cheng, 2004) and this advantage is especially evident when the scale for each dimension are relatively short and the correlations among them are generally high (Wang, Yao, Tsai, Wang, and Hsieh, 2006).

Considering the unique nature of TA's work and the fact that TAs are playing an important role in schooling, the present study aims to fill the gap in existing research to examine the range of capacities that are essential to and important for TA's work in the Hong Kong school context and to develop and validate a scale for measuring TA's self-reported efficacy on those identified capacities. These findings should increase our understanding of the nature of TA's efficacy related to their work in schools, and the instruments developed in this study could provide a foundation for future research in this field.

Method

Scale Development

The research team responsible for item development consisted of two researchers specialized in measurement and three colleagues specializing in teacher education and running a TA training programme at the time of instrument development. The procedure recommended by DeVellis (2012) was adopted to develop the instrument. The scale development started from a conceptual map which depicted the structure of the target latent traits to be measured. An item pool was generated based on literature review and the results of consultative meetings with TAs and teachers. Given that TAs share some functions of teacher, 68 items were selected, with necessary modifications, from available teachers' efficacy scales, such as the TSES (Tschannen-Moran and Woolfolk Hoy, 2001), TES (Gibson and Dembo, 1984), Bandura's Teacher Efficacy Scale (undated), and Teacher Efficacy for Inclusive Practice Scale (Sharma, Loreman, and Forlin, 2011). These items are mainly about learning support, teaching support, and behaviour management. The research team generated other 20 items to reflect the unique aspects of TA's work in school which have not been covered by teachers' efficacy scales, such as cooperation and administrative support.

This resulted in an item pool with 88 items. DeVellis (2012, p. 80) suggested to have an initial item pool that is three or four times as large as the final scale. Therefore, an item pool with 88 items is probably ideal to produce an intended final scale with 22 to 30 items. This item pool was subject to review by an expert panel consisting of the research team and another group of TAs and teachers. Content validity was confirmed and redundant items were removed. The relevance of items, ambiguities and possible bias in item wording were checked and necessary revisions were made. This resulted in a scale with 45 items which was piloted on a small group of TAs (N = 26). Pilot participants were asked to write down any suggestions for this scale, in addition to answering the items. Further revisions were made based on pilot results. The resultant TAES contained 38 items. A six-point Likert-type response scale, ranging from *Strongly Disagree* (1), *Disagree* (2), *Somewhat Disagree* (3), *Somewhat Agree* (4), *Agree* (5), to *Strongly Agree* (6) was used for all items. The TAES dimensions, operational definitions, and exemplar items are shown in Table 1.

Sample

Participants for this study were a convenience sample of 531 TAs from a TA training course. 271 (51.0%) TAs came from primary schools, 239 (45.0%) from secondary schools, while 21 (4.0%)

Table 1

TAES Dimensions, Operational Definitions, and Exemplar Items

Dimension	Abbreviation	Operational Definition	Number of Items	Exemplar item
Learning Support	LS	Direct support to students' learning	6	I can motivate students who show low interest in school work.
Teaching Support	TS	Support to teachers' instruction	8	I can provide an alternate explanation or example when students are confused.
Behaviour Management	BM	Managing students' behaviour in and out of classroom	9	I can control disruptive behaviour in the classroom.
Cooperation	CO	Liasion and work with internal and external parties	8	I can cooperate with non-teaching staff in my school.
Administrative Support	AS	Cleric duties to facilitate the school's function	7	I can assist to improve the leaning environment.

did not provide school information. 141 (26.6%) were identified as males, 366 (68.9%) as females, while 24 (4.5%) did not provide gender information. The majority of participants (416, 78.3%) were aged between 21 and 30, 92 (17.3%) aged 31 or above, 12 (2.3%) aged 20 or below, and 11 (2.1%) without age information.

Procedure

Research ethical guidelines of the authors' institution were followed and permission for the study was granted by the Human Research Ethics Committee of that institution. The questionnaire was printed in (Cantonese) Chinese and was administered before the training course in order to avoid the influence of the course on participants' conceptions of efficacy. The participants were guaranteed that the participation was on voluntary basis, all personal information and data collected would be kept strictly confidential and used only for research purposes.

Data Analysis

The Rasch rating scale model (Wright and Masters, 1982) was adopted for data analysis since the 6-point response Likert-style scale was invariant across all 38 items. Considering that TA's efficacy is likely to be a multidimensional construct given the multifaceted nature of TA's work, a multidimensional Rasch model is considered more appropriate than a single unidimensional model to investigate the measurement properties of TAES.

The Rasch model adopts a "the data fit the model" position. The empirical data must meet prior requirements of Rasch model in order to achieve objective measurement (Andrich, 2004). The prior standards ensure that the measurement results under different circumstances can be communicated within a stable framework and a scale constructed in one study can be applied to data collected in another context. A number of quality assurance indicators were applied in this study to examine the psychometric quality of the developed scale.

The first check examined whether the response category options function as intended. Thresholds between consecutive categories should increase monotonically to ensure that higher measures represent higher levels of latent trait. Threshold distances should be neither too short nor too far apart on the latent trait scale (Bond and Fox, 2007). Ideally, thresholds should increase by at least 1.4 logits but less than 5.0 logits (Linacre, 2002). For four or more category scales, however, a shorter threshold distance is acceptable.

A principal components analysis (PCA) of Rasch residuals (Linacre, 1998; Wright, 1996) using Winsteps 3.70 (Linacre, 2006b) was conducted on each dimension as well as on the whole scale to check whether or not each was measuring a unidimensional latent trait. ConQuest version 2.0 software (Wu, Adams, Wilson, and Haldane, 2007) was then used to undertake the multidimensional Rasch analysis.

Several criteria including Rasch reliability, item fit statistics, and differential item functioning (DIF) were used to investigate the quality of the items and the psychometric properties of the scale. ConQuest provides EAP/PV reliability which is calculated using an expected a posteriori estimation based on plausible values (Wu et al., 2007). EAP/PV reliability is explained variance according to the estimated model divided by total variance of the person ability.

Item fit statistics indicates the extent to which the data match specifications of a Rasch model. Outfit and Infit mean square (MNSQ) as well as their standardized forms (ZSTD) are commonly used item fit statistics. Since MNSQ is more sensitive to sample size, especially for Infit statistics when the sample size is larger than 500 (Smith, Schumacher, and Bush, 1998), Outfit ZSTD will be used to detect misfit items in this study. The Outfit ZSTD has approximately normalized t distribution with an expected value of 0 and a SD of 1.0. Underfit (values much higher than +2.0) suggest that variation in the observed data is greater than that predicted by the Rasch model and overfit (values much lower than −2.0) indicate redundancy between the information carried by the item in question and the other items in the scale (Linacre, 2006a). From a scale development

perspective, underfit is detrimental to a scale while overfit usually has little practical implications (Bond and Fox, 2007). Therefore, +2 will be adopted as the critical value of Outfit ZSTD. Items were considered to have misfit to Rasch model if the Outfit ZSTD is larger than +2.0.

DIF analysis should be used to check the construct equivalence across groups (Wang, 2000). The existence of DIF indicates that different groups may have different interpretation or perspectives on the items. In other words, different groups perform differently on the same items, after controlling their difference in the latent trait levels. As suggested by Adams and Wu (2010) that whether the effect of a DIF is of substantive importance is largely determined by the magnitude of that DIF, a difference equal to or larger than 0.5 logits (Wang et al., 2006) was regarded as evidence of substantial DIF in this study.

Result

Dimensionality

A PCA of Rasch residuals was conducted on each dimension as well as the whole scale. The eigenvalues of the Rasch dimension and the first contrast, i.e., first PCA component in the correlation matrix of the residuals, for each dimension and for the whole TAES scale were presented in Table 2.

Linacre (2006a) suggested an eigenvalue of 2.0 as the cut-off value. If the eigenvalue of the first contrast is greater than 2.0, then there is probably a "second dimension" besides the Rasch dimension; if the eigenvalue is smaller than 2.0, the residuals could be regarded as at random noise level. As shown in Table 2, the eigenvalues for the five dimensions range from 6.4 to 13.1, whereas the eigenvalues of first contrast are all less than 2.0 except that for dimension BM which is marginally higher than 2.0. This result supports the claim that the items in the five dimensions probably each measure a single latent trait. On the contrary, the eigenvalue of 3.5 for the first contrast for the whole scale suggests that items in TAES contain more than one dimension

Table 2

The Eigenvalues of Rasch Dimension and the First Contrast.

	Eigenvalue	
Dimension/Scale	Rasch Dimension	First Contrast
LS	9.7	1.8
TS	9.3	1.6
BM	13.1	2.1
CO	10.7	1.8
AS	6.4	1.7
Overall TAES	35.5	3.5

Furthermore, the correlations among person measures on the five dimensions of TAES range from 0.582 to 0.782 (see Table 3). These correlations indicate that multidimensional Rasch model is appropriate to be used since the five scales are not measuring the same dimension of the latent trait but the correlations among them could be utilized to increase measurement precision (Wang et al. 2006).

Table 3

Correlations among Dimensions of TAES.

	LS	TS	BH	CO
TS	0.782**		–	–
BM	0.692**	0.670**		
CO	0.582**	0.672**	0.611**	–
AS	0.651**	0.713**	0.699**	0.754**

Note. $^{**}p < .01$

Model-data fit

The multidimensional Rasch analysis on TAES was conducted using ConQuest. The data to model fit was examined through Outfit ZSTD for each item.

It can be seen from Table 4 that 7 items showed misfit to the Rasch model. Those items include TS Item 2 (I can provide suggestions on teaching activities to teachers), TS Item 4 (I can get students to learn together in small groups), TS Item 33 (I can assist teachers in preparing teaching materials), CO Item 35 (I can cooperate with parents who need assistance), AS Item 8 (I can manage after-school class), AS Item 11 (I can support students during extracurricular activities), and AS Item 38 (I can provide cleric/

Table 4

Item Outfit ZSTD and DIF.

Item	Outfit ZSTD	Gender DIF (M-F)	School Type DIF (P-S)
Learning Support			
Item 3	−2.7	−0.18	0.31
Item 7	−3.7	0.07	0.21
Item 17	−4.7	0.01	−0.11
Item 21	−7.8	0.09	−0.27
Item 28	−0.8	0.36	−0.19
Item 30	−3.9	−0.12	−0.22
Teaching Support			
Item 2*	*5.5*		
Item 4*	*2.8*		
Item 6	−5.1	−0.24	−0.24
Item 15	−5.0	−0.16	0.11
Item 20	−3.1	−0.21	0.09
Item 31	−4.5	−0.13	0.03
Item 33*	*11.6*		
Item 37	−1.5	−0.10	0.34
Behaviour Management			
Item 9	−2.5	0.07	0.26
Item 12	−3.2	−0.27	0.29
Item 13	−5.4	0.02	−0.14
Item 14	−5.5	0.03	0.08
Item 16	−3.7	−0.10	0.17
Item 19	−5.0	0.09	−0.30
Item 22	−3.8	0.31	−0.20
Item 27**	−1.1	*−0.52*	0.28
Item 29	−4.6	0.36	−0.44
Cooperation			
Item 5	2.0	0	−0.26
Item 23	0.2	0.14	0.02
Item 24	2.1	−0.1	0.13
Item 25	0.6	0.09	−0.27
Item 32	−0.8	−0.38	0.26
Item 34	0.5	0.22	−0.10
Item 35*	*4.4*		
Item 36	1.6	0.04	0.23
Administrative Support			
Item 1	0.4	0.15	−0.10
Item 8*	*9.1*		
Item 10	−0.1	−0.06	−0.17
Item 11*	*5.7*		
Item 18	−0.8	−0.22	−0.12
Item 26	2.2	0.13	0.38
Item 38*	*12.4*		

Note. *misfit item (Outfit ZSTD > +2.0); **item with substantial DIF (the difference of item difficulty > 0.5 logits)

administrative support to teachers). Two items including CO Item 24 and AS Item 26 has a marginal ZSTD value (+2.1 and +2.2 respectively). They were kept in the scale considering that this was the first validation study for this scale. Since deleting items probably affects the accuracy of the remaining scale and error rate of the model estimates (Hart and Wright 2002), the identified 7 misfit items were removed from the scale one by one, in descending misfit order, and Rasch analysis were re-applied until all remaining items showed fit to the Rasch model.

In order to examine the construct equivalence across gender (male vs female) and school type (primary vs secondary), DIF analysis was undertaken after removing all the misfit items. Only BM Item 27 (I can deal with students who are physically aggressive) appeared to have substantial DIF (a difference larger than 0.5 logits) across gender. The difficulty of this item seems substantially higher for female TAs than for males after controlling their overall efficacy levels. This DIF item was removed and the EAP/PV reliabilities reported by ConQuest for the five dimensions are quite good (LS: 0.90, TS: 0.90, BM: 0.90, CO: 0.88, and AS: 0.89).

The distribution of item difficulty and person ability

Table 5 presents the item difficulties with associated standard errors for the 30 remaining items. The item difficulty covers a range from −2.26 to 1.91 logits. Analysis for LS shows that LS Item 7 (I can get students to believe they can do well in school work) to be the easiest to endorse (−0.46 logits) and LS Item 3 (I can help students improve their academic achievements) to be the most difficult (0.38 logits). TS Item 6 (I can respond to difficult questions from my students) is the easiest item (−0.83 logits) and TS Item 37 (I can use a variety of assessment strategies) is the most difficult item (0.82 logits) in TS. BM ranges from the easiest BM Item 9 (I can make my expectations clear about student behaviour) (−0.69 logits) to the most difficult BM Item 22 (I can prevent disruptive behaviour in the classroom before it occurs) (0.46 logits). CO Item 34 (I can

cooperate with non-teaching staff in my school) has lowest difficulty (−1.29 logits) in the CO dimension and CO Item 24 (I can make use of community resources to help students' learning) has the highest CO difficulty (1.91 logits). For dimension AS, AS Item 1 (I can accomplish the work assigned by my school) and AS Item 18 (I can assist to improve the learning environment) are the easiest (−2.26 logits) and the most difficult (1.09 logits) item respectively.

The Rasch model estimates person and item measures on the same metric. As shown in the item-person map (Figure 1), the five frequency distributions on the left side indicate the TA's measures on each of the five dimensions of efficacy. The TAs with higher efficacy are placed at the top of the scale and those with lower efficacy are placed at the bottom. In a similar way, the item difficulty distributions for the five dimensions are displayed on the right side of the scale. The items that are more difficult to endorse are placed at the top, and those that are easier to endorse are placed at the bottom. The notation on the right side of the map are used to represent the item response category thresholds so that $x.y$ represents the y-th threshold of item x.

Mean estimates of TA's efficacy for the five dimensions range from 0.79 to 2.21 logits. Table 6

Table 5

Item Difficulty and Standard Error (SE).

Item	Item Difficulty (*SE*)	Item	Item Difficulty (*SE*)
Learning Support		Teaching Support	
Item 3	0.38(0.04)	Item 37	0.82(0.09)
Item 28	0.32 (0.04)	Item 31	0.52(0.04)
Item 21	0.18(0.04)	Item 20	0.09(0.04)
Item 17	−0.15(0.04)	Item 15	−0.61(0.04)
Item 30	−0.27(0.09)	Item 6	−0.83(0.04)
Item 7	−0.46(0.04)		
Behaviour Management		Cooperation	
Item 22	0.46(0.04)	Item 24	1.91(0.04)
Item 12	0.40(0.04)	Item 36	1.73(0.11)
Item 16	0.29(0.04)	Item 23	0.08(0.05)
Item 13	0.19(0.04)	Item 32	−0.50(0.05)
Item 19	0.00(0.04)	Item 5	−0.76(0.05)
Item 14	−0.33(0.04)	Item 25	−1.17(0.05)
Item 29	−0.34(0.11)	Item 34	−1.29(0.05)
Item 9	−0.69(0.05)		
Administrative Support			
Item 18	1.09(0.04)		
Item 26	0.88(0.08)		
Item 10	0.30(0.04)		
Item 1	−2.26(0.05)		

Table 6

Person Measures on the Five Dimensions.

	Mean	SD	Maximum*	Minimum*
LS	0.79	1.49	6.69	−4.23
TS	0.98	1.46	6.99	−3.91
BM	0.81	1.41	5.92	−4.75
CO	2.21	1.38	7.78	−2.26
AS	1.42	1.33	7.00	−2.67

Note. *Some persons with extreme measures are not shown in the item-person map since the frequency is less than 4.

and the item-person map (Figure 1) show that TA's efficacy levels spread across a wide range of the latent trait scale. As well, TAES items (along with their item thresholds), still cover the major range of TA's efficacy levels and, therefore, provide a sufficient, targeted measurement of TA's efficacy.

TAES has a 6-category rating scale which generates 5 thresholds. Rasch analysis reports that the calibrations of the 5 thresholds (intersections between consecutive categories) increases monotonically from −3.05, −1.53, −0.83, 1.15, to 4.25 logits. No threshold disorder was found and the threshold distances (1.52, 0.7, 1.98, and 3.1 logits respectively) are quite good for a 6-category response structure according to Linacre's (2002) criteria. The results indicated that the 6-category response structure function well for the TAES.

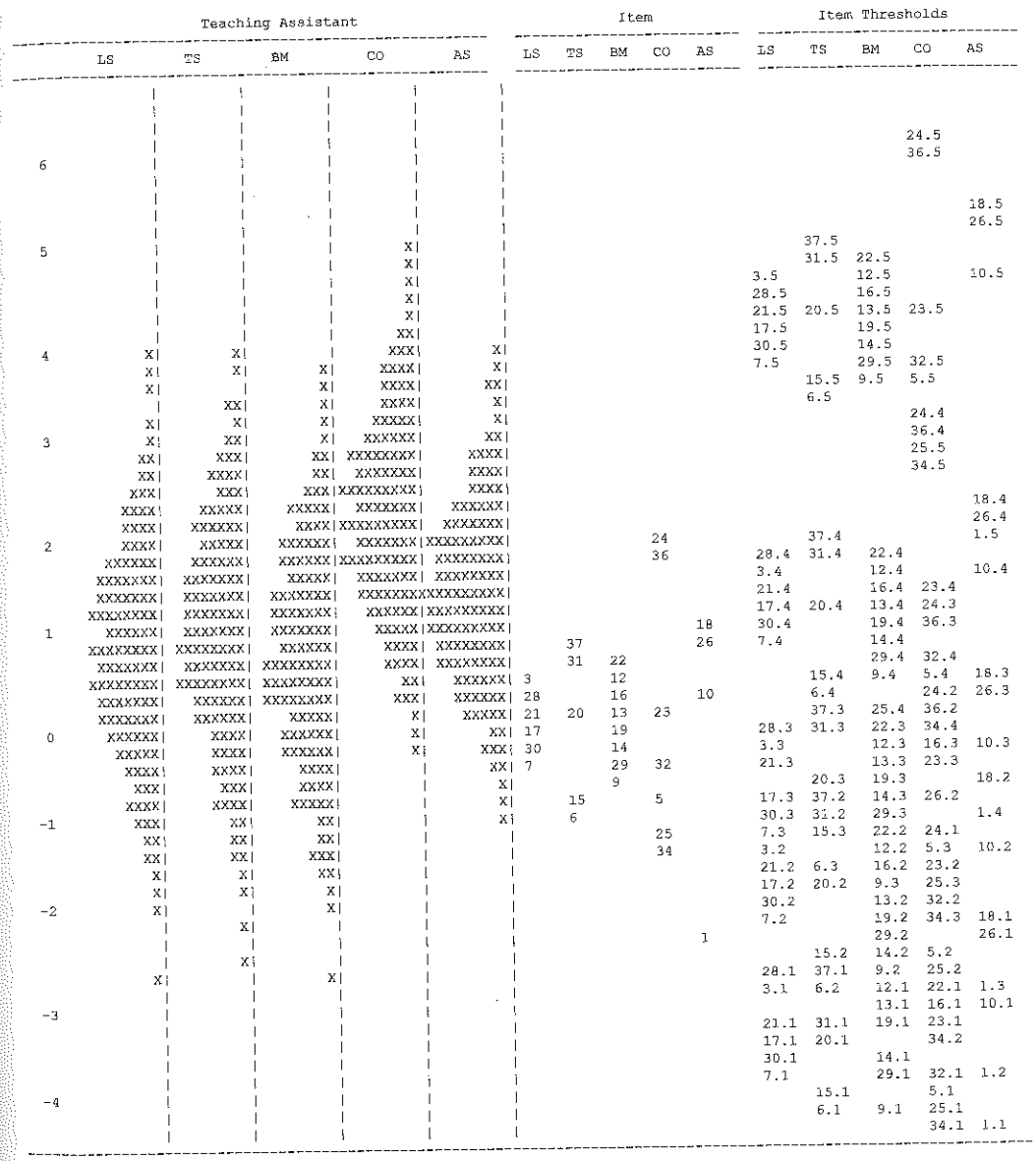

Figure 1. The Item-person Map for TAES.

Discussion

The present study aimed to investigate capacities that are essential and important to TA's work in Hong Kong primary and secondary schools and to develop a sound, psychometrically appropriate instrument for measuring TA efficacy. The results of the Rasch analyses reveal that the conceptual design of the instrument matched the requirements of the multidimensional Rasch model. TA's efficacy could be understood from a multidimensional perspective and a 5-dimension structure was supported in the present study. The final version of TAES (see Appendix A) reported here consists of 30 items that access TA's efficacy on five dimensions: learning support (LS, 6 items), teaching support (TS, 5 items), behaviour management (BM, 8 items), cooperation (CO, 7 items), and administrative support (AS, 4 items). The EAP/PV reliabilities for the five dimensions are all around 0.90. The 6-category response category structure also works well for the TAES since the calibrations of the 5 thresholds increases monotonically and the threshold distances are reasonable for a 6-category response structure.

Seven items that were relevant and conceptually valid were excluded from the scale since they showed misfit to the Rasch model's measurement requirements. However, these items and the efficacy components they represent should not be prematurely regarded as unrelated to TA's efficacy. Closer investigation and contextual interpretation are necessary to reveal the potential underlying causes for the misfit of those items. For example, the reason for misfit of AS Item 8 (I can manage after-school class) and AS Item 11 (I can support students during extracurricular activities) could lie in the fact that these tasks—managing after-school class and supporting students during extracurricular activities—are purely administrative work for some TAs, while they might involve teaching duties for other TAs in accordance with the particular school's requirements. Such dual functions embraced by these items could, therefore, result in the misfit. Future research could explore this point further by rewording those items to check whether they can perform in a way that is consistent with TA's responses to other items in the AS dimension. BM Item 27 (I can deal with students who are physically aggressive.) showed substantial DIF across gender. Female TAs found this item much more difficult to endorse than did male TAs, after controlling for overall level of TA efficacy. TA's responses to this item were obviously influenced by gender factors: physical aggression by students is more confidently responded to by male TAs rather than female TAs. Discarding this item would be inappropriate if TAs are really expected to deal with students' physical violence. In that case, splitting the item response by gender would allow BM Item 27 to be retained by scoring it separately for males and females.

Although the range of item difficulty (with thresholds) is well targeted with TA's efficacy levels along the latent trait scale, two dimensions cover only a relatively small range of item difficulties (i.e., LS and BM), while others (i.e., TS, CO, and AS) spread over a wider range but gaps exist in their coverage. For dimension CO, some items could be added to bridge the 1.65 logits gap between items CO Items 23 and 36. In dimension AS (only four items), AS Item 1 is so extraordinarily easy that it cannot help to differentiate TA's efficacy levels. Further items could be developed to shrink the (2.56 logits) gap between AS Items 1 and 10. The items in dimensions LS and BM tend cluster along the scale. It indicates that TA's duties in these two dimensions have similar difficulties and TA's efficacy does not vary much. In contrast, the dimension CO items spread over a wide range along the scale. The likely explanation is that duties in the CO dimension involve many different stakeholders and TA's level of efficacy is somewhat dependent on the objectives with which they are obliged to cooperate. In Hong Kong schools, it seems that TAs feel more confident in working with non-teaching staff in their schools, but least confident in cooperating with the wider community.

TA's mean estimates of efficacy levels are substantially higher than the mean item difficulty on all the five dimensions. There are two potential interpretations for this result. One is that the current items are relatively easy for

TAs, and more difficult items need to be added. In other words, there is a ceiling effect that limits TAs from expressing their efficacy levels on these dimensions. Indeed, some TAs report perfect (average raw score equals 6) or nearly perfect scores. Another interpretation, which might deserve more attention, is that many TAs are overqualified for their positions. In Hong Kong, many TAs are qualified teachers or newly graduated university students already holding teaching certificates. Their knowledge and skills make them very competent as TAs but most of them regard their TA job as merely a stepping-stone to employment as a teacher. This might also be one of the reasons for the turnover problems associated with TA posts reported in previous studies (Blatchford et al. 2006). The differences between TA's mean estimates of efficacy levels and the mean item difficulties are especially salient for the dimensions CO (TAs are 2.21 logits higher than the item mean) and AS (1.42 logits higher than the mean item difficulty). This is likely to be because, in Hong Kong schools, TAs are frequently involved in administrative duties and working in cooperation with other school staff or parents. The plethora of those experiences lend TAs to have more confidence in conducting such duties compared to independently undertaking duties such as supporting learning/teaching or managing student behaviour.

The advantages of Rasch analysis over traditional analysis were echoed in this study. First of all, the Rasch analysis transforms the ordinal Likert type raw scores into interval measures on a logit scale, which have consistent unit and stable meaning on the underlying trait scale (i.e., TA efficacy) when the data fit the model. Second, Rasch analysis estimates persons and items on the same metric according to their efficacy levels and item difficulties. Comparisons could be made between person efficacy levels and item difficulties based on their locations on the trait continuum for each dimension (Yan and Bond, 2011).

The development of the TAES provides a crucial step toward better understanding of TA's work and associated efficacy beliefs. This instrument provides an appropriate tool for measuring TA's efficacy as a context-specific construct and opens new possibilities for future research. For example, how do the efficacy beliefs of TAs impact their work performance? What are the factors influencing TA's efficacy? How to enhance and maintain a high level of efficacy for TAs?

One point which needs attention is that the results support use of the dimension scores of the TAES, but do not suggest use of a single composite score of TA self-efficacy. Users of the instrument should use the five dimension scores to arrive at inferences about efficacy. As for other newly-developed instruments, TAES has its own limitations and further validation and empirical research will be needed to consolidate this instrument and contribute to literature in efficacy research. First, items of the TAES are open for further development and revision. More items with higher difficulty level could be added, especially for dimensions CO and AS. Second, this study exclusively relies on data from Hong Kong TAs. This might be a limitation for immediate generalization of the TAES to other education systems. As indicated by the TAES as well as our focus group interviews, TAs working in Hong Kong primary and secondary schools play a multi-faceted role. The whole TAES could be applied in a similar research/education context to gather more validation evidence for this scale and to check the invariance of item estimates, while in an education system in which TAs are not required to play such multi-faceted roles, the TAES can be used piecemeal (i.e., one or two dimensions may be used in isolation) and this lends generalizability to the instrument precisely.

Acknowledgement

The authors would like to thank Professor Trevor G. Bond for his invaluable advice on conducting the Rasch analyses.

References

Adams, R. J., Wilson, M., and Wang, W. C. (1997). The multidimensional random coefficient multinomial logit model. *Applied Psychological Measurement, 21*, 1-23.

Adams, R. J., and Wu, M. L. (2010). *Differential item functioning*. Retrieved on March 11, 2014, at http://www.acer.edu.au/documents/Conquest-Tutorial-6-DifferentialItemFunctioning.pdf

Allinder, R. M. (1994). The relationship between efficacy and the instructional practices of special education teachers and consultants. *Teacher Education and Special Education, 17*, 86-95.

Andrich, D. (2004). Controversy and the Rasch model: A characteristic of incompatible paradigms? *Medical Care, 42*, 1-16.

Baird, J. (2012). Do we need marking at all? *Assessment in Education: Principles, Policy & Practice, 9*, 277-279.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York, NY: W. H. Freeman.

Bandura, A. (2006). Guide for constructing self-efficacy scales, In F. Pajares and T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307-337). Greenwich, CT: Information Age.

Bandura, A. (undated). *Bandura's teacher efficacy scale*. Available on-line at http://people.ehe.osu.edu/ahoy/files/2009/02/bandura-instr.pdf\

Barkaoui, K. (2011). Effects of marking method and rater experience on ESL essay scores and rater performance. *Assessment in Education: Principles, Policy & Practice, 18*, 279-293.

Blatchford, P., Bassett, P., Brown, P., Martin, C., Russell, A., and Webster, R. (2006). *The deployment and impact of support staff project: Strand 1, Wave 2 Report*. London, UK: Institute of Education, University of London.

Blatchford, P., Bassett, P., Brown, P., Martin, C., Russell, A., and Webster, R. (2009). *The deployment and impact of support staff project: Research summary*. Retrieved on December 15, 2011, at http://www.schoolsupportstaff.net/publications/DISS_reports/DISS_Res_Sum.pdf

Blatchford, P., Bassett, P., Brown, P., and Webster, R. (2009). The effect of support staff on pupil engagement and individual attention. *British Educational Research Journal, 35*, 661-686.

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

Bozman, C. E. (2012). *The effects of principals' leadership styles, teacher efficacy, and teachers' trust in their principals on student achievement*. Unpublished doctoral dissertation. Tennessee State University, Nashville, TN.

Chan, D. W. (2008). Teacher self-efficacy and successful intelligence among Chinese secondary school teachers in Hong Kong. *Educational Psychology, 28*, 735-746.

Coladarci, T. (1992). Teachers' sense of efficacy and commitment to teaching. *Journal of Experimental Education, 60*, 323-337.

Deepa, M. (2007). Students' and teachers' efficacy in use of learning strategies and achievement in Mathematics. *Issues in Educational Research, 17*, 207-231.

Department for Children, Schools and Families. (DCSF, 2008). *School workforce in England (including pupil:teacher ratios and pupil:adult ratios) January 2008 (provisional)*. SFR 10/2008. London, UK: Author.

DeVellis, R. F. (2012). *Scale development: Theory and applications* (3rd ed.). Thousand Oaks, CA: Sage.

Fischer, A. R. H., Frewer, L. J., and Nauta, M. J. (2006). Toward improving food safety in the domestic environment: A multi-item Rasch scale for the measurement of the safety efficacy of domestic food-handling practices. *Risk Analysis, 26*, 1323-1338.

Gibson, S., and Dembo, M. (1984). Teacher efficacy: A construct validation. *Journal of Educational Psychology, 76*, 569-582.

Hart, D. L., and Wright, B. D. (2002). Development of an index of physical functional health status in rehabilitation. *Archives of Physical Medicine and Rehabilitation, 83*, 655-665.

Kim, E. (2009). Beyond language barriers: Teaching self-efficacy among East Asian international teaching assistants. *International*

Journal of Teaching and Learning in Higher Education, 21, 171-180.

Komarraju, M. (2008). A social-cognitive approach to training teaching assistants. *Teaching of Psychology, 35*, 327-334.

Kyriakides, L., Kaloyirou, C., and Lindsay, G. (2006). An analysis of the Revised Olweus Bully/Victim Questionnaire using the Rasch measurement model. *British Journal of Educational Psychology, 76*, 781-801.

Lee, M., Peterson, J. J., and Dixon, A. (2010). Rasch calibration of Physical Activity Self-Efficacy and Social Support Scale for persons with intellectual disabilities. *Research in Developmental Disabilities, 31*, 903-913.

Linacre, J. M. (1998). Detecting multidimensionality: Which residual data-type works best? *Journal of Outcome Measurement, 2*, 266-283.

Linacre, J. M. (2002). Optimizing rating scale category effectiveness. *Journal of Applied Measurement, 3*, 85-106.

Linacre, J. M. (2006a). *A user's guide to WINSTEPS/MINISTEP: Rasch-model computer programs*. Chicago, IL: Winsteps.com.

Linacre, J. M. (2006b). WINSTEPS: Rasch measurement computer program [computer software]. Chicago, IL: Winsteps.com.

MacMillan, P. D. (2000). Classical, generalizability, and multifaceted Rasch detection of interrater variability in large, sparse data sets. *The Journal of Experimental Education, 68*, 167-190.

McCrea, L. B. G. (2006). *An investigation of the relationship between graduate teaching assistants' teaching self-efficacy and attributions for students' learning*. Unpublished doctoral dissertation. University of Akron, Akron, Ohio. Retrieved on 8 August 2012, from http://etd.ohiolink.edu/send-pdf.cgi/McCrea%20Laura%20Grove.pdf?akron1144943095

Mills, N. (2011). Teaching assistants' self-efficacy in teaching literature: Sources, personal assessments, and consequences. *Modern Language Journal, 95*, 61-80.

Muis, K. R., Winne, P. H., and Edwards, O. V. (2009). Modern psychometrics for assessing achievement goal orientation: A Rasch analysis. *British Journal of Educational Psychology, 79*, 547-576.

Nam, S. K., Yang, E., and Lee, S. M. (2011). A psychometric evaluation of the Career Decision Self-efficacy Scale with Korean students: A Rasch model approach. *Journal of Career Development, 38*, 147-166.

Nelson, S. L. (2008). *Teacher efficacy and student motivation: A link to achievement in elementary mathematics*. Unpublished doctoral dissertation. University of South Dakota, Vermillion, SD.

Pintrich, P. R., and Schunk, D. H. (1996). *Motivation in education: Theory, research, and applications*. Englewood Cliffs, NJ: Merrill/Prentice-Hall.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Copenhagen, Denmark: Danish Institute for Educational Research. (Expanded edition, 1980. Chicago, IL: University of Chicago Press.)

Sharma,U., Loreman, T., and Forlin, C. (2011). Measuring teacher efficacy to implement inclusive practices: An international validation. *Journal of Research in Special Educational Needs, 11*, 1-10.

Smith, R. M., Schumacher, R. E., and Bush, M. J. (1998). Using item mean squares to evaluate fit to the Rasch model. *Journal of Outcome Measurement, 2*, 66-78.

Soodak, L. C., Podell, D. M., and Lehman, L. R. (1998). Teacher, student, and school attributes as predictors of teachers' responses to inclusion. *Journal of Special Education, 31*, 480-497.

Tollerud, T. (1990). *The perceived self-efficacy of teaching skills of advanced doctoral students and graduates from counselor education programs*. Unpublished doctoral dissertation. University of Iowa, Iowa City, IA.

Tschannen-Moran, M., and Woolfolk-Hoy, A. (2001). Teacher efficacy: Capturing an elusive

construct. *Teaching and Teacher Education, 17*, 783-805.

Tschannen-Moran, M.,Woolfolk Hoy, A., and Hoy,W. K. (1998). Teacher efficacy: Its meaning and measure. *Review of Educational Research, 68*, 202-248.

Wang, W. C. (2000). Modeling effects of differential item functioning in polytomous items. *Journal of Applied Measurement, 1*, 63-82.

Wang, W. C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *The Journal of Experimental Education, 72*, 221-261.

Wang, W. C., Chen, P. H., and Cheng, Y. Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*, 116-136.

Wang, W. C., Yao, G., Tsai, Y, J., Wang, J. D., and Hsieh, C. L. (2006). Validating, improving reliability, and estimating correlation of the four subscales in the WHOQOL-BREF using multidimensional Rasch analysis. *Quality of Life Research, 15*, 607-620.

Wright, B. D. (1992). IRT in the 1990s: Which models work best? *Rasch Measurement Transactions, 6*, 196-200.

Wright, B. D. (1996) Local dependency, correlations and principal components. *Rasch Measurement Transactions, 10*, 509-511.

Wright, B. D. (1997). A history of social science measurement. *Educational Measurement: Issues and Practice, 16*(4), 33-45.

Wright, B. D., and Masters, G. N. (1982). *Rating scale analysis.* Chicago, IL: MESA Press.

Wu, M. L., Adams, R. J., Wilson, M. R., and Haldane, S. A. (2007). ACER ConQuest version 2.0: Generalised item response modelling software [Computer software]. Melbourne, Australia: ACER Press.

Yan, Z., and Bond, T. (2011). Developing a Rasch measurement physical fitness scale for Hong Kong primary school-aged students. *Measurement in Physical Education and Exercise Science, 15*, 182-203.

Yan, Z., and Mok, M. M. C. (2012). Validating the Coping Scale for Chinese Athletes using multidimensional Rasch analysis. *Psychology of Sport and Exercise, 13*, 271-279.

Yan, Z., and Coniam, D. (2013). Assessing the ease of use in the environment and markers' acceptance of onscreen marking: A Rasch measurement perspective. *Educational Research and Evaluation, 19*, 461-483.

Zelinski, E. M., and Gilewski, M. J. (2004). A 10-item Rasch modeled memory self-efficacy scale. *Aging & Mental Health, 8*, 293-306.

Zhu, W., and Kang, S-J. (1994). Cross-cultural stability of the optimal categorization of a self-efficacy scale: A Rasch analysis. *Measurement in Physical Education and Exercise Science, 2*, 225-241.

Appendix A: The final version of 30-item TAES

Strongly Disagree (1), Disagree (2), Somewhat Disagree (3), Somewhat Agree (4), Agree (5), to Strongly Agree (6)
Please indicate your degree of agreement for the following statements.

Learning Support							
Item 3	I can help students improve their academic achievements	①	②	③	④	⑤	⑥
Item 7	I can get students to believe they can do well in school work	①	②	③	④	⑤	⑥
Item 17	I can motivate students who show low interest in school work	①	②	③	④	⑤	⑥
Item 21	I can help students concentrate on classroom learning	①	②	③	④	⑤	⑥
Item 28	I can get through to the most difficult students	①	②	③	④	⑤	⑥
Item 30	I can make students enjoy classroom learning	①	②	③	④	⑤	⑥
Teaching Support							
Item 6	I can respond to difficult questions from my students	①	②	③	④	⑤	⑥
Item 15	I can provide an alternate explanation or example when students are confused	①	②	③	④	⑤	⑥
Item 20	I can adjust my lessons to the proper level for individual students	①	②	③	④	⑤	⑥
Item 31	I can gauge student comprehension of what you have taught	①	②	③	④	⑤	⑥
Item 37	I can use a variety of assessment strategies	①	②	③	④	⑤	⑥
Behaviour Management							
Item 9	I can make my expectations clear about student behaviour	①	②	③	④	⑤	⑥
Item 12	I can deal with students who are verbally aggressive	①	②	③	④	⑤	⑥
Item 13	I can keep a few problem students from ruining an entire lesson	①	②	③	④	⑤	⑥
Item 14	I can establish routines to keep activities running smoothly	①	②	③	④	⑤	⑥
Item 16	I can respond to defiant students	①	②	③	④	⑤	⑥
Item 19	I can control disruptive behaviour in the classroom	①	②	③	④	⑤	⑥
Item 22	I can prevent disruptive behaviour in the classroom before it occurs	①	②	③	④	⑤	⑥
Item 29	I can get students to follow classroom rules	①	②	③	④	⑤	⑥
Cooperation							
Item 5	I can cooperate with teachers	①	②	③	④	⑤	⑥
Item 23	I can cooperate with other professionals	①	②	③	④	⑤	⑥
Item 24	I can make use of community resources to help students' learning	①	②	③	④	⑤	⑥
Item 25	I can get well with students	①	②	③	④	⑤	⑥
Item 32	I can obtain trust from students	①	②	③	④	⑤	⑥
Item 34	I can cooperate with non-teaching staff in my school	①	②	③	④	⑤	⑥
Item 36	I can assist families in helping their children do well in school	①	②	③	④	⑤	⑥
Administrative Support							
Item 1	I can accomplish the work assigned by my school	①	②	③	④	⑤	⑥
Item 10	I can assist to make my school a safe place	①	②	③	④	⑤	⑥
Item 18	I can assist to improve the learning environment	①	②	③	④	⑤	⑥
Item 26	I can provide guidance to students who have emotional problems	①	②	③	④	⑤	⑥