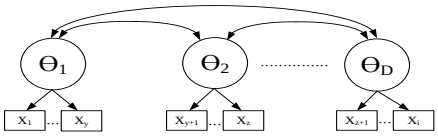

X_i = Each item i belongs to a specific component latent trait.
θ_d = Estimate of the component latent trait *d*, *d* = 1, 2, …, D.
Correlations exist between component latent traits.

Figure 1(c). Multidimensional approach of Rasch models

Figure 1. Diagrams of the unidimensional, consecutive, and multidimensional Rasch approaches

The Rasch model (Rasch, 1960) is the simplest model in item response theory (IRT), and requires all the items to measure the same latent trait. This is the unidimensionality assumption of the Rasch model. In some cases, respondents' abilities are composed of several component latent traits. The typical approach, called the consecutive Rasch model, is to calibrate the Rasch model for each component latent trait with its own items separately. The consecutive Rasch models, which ignore the possibility that performance of items might be interrelated across different component latent traits (Briggs and Wilson, 2003), might yield imprecise results when models are estimated independently. Figures 1(a) and 1(b) (Chang and Shih, 2012) depict the unidimensional approach and consecutive approach, respectively. To take the correlations between component latent traits into account, the multidimensional Rasch model that simultaneously calibrates the component latent traits has been developed to increase measurement precision (as illustrated in Figure 1(c)).

2.1 The unidimensional Rasch model and consecutive Rasch model

To simplify our introduction to the Rasch model and its extended models, we shall first consider the dichotomous responses The items in the questionnaire are first assumed to be the type of "Can you easily achieve the following necessary tasks while conducting red-light running enforcement?" The response is either "yes" or "no". A score of 1 is assigned to an item to which the police officer responds "yes, I can"; otherwise, a score of 0 is assigned. The probability that an officer n will respond with "yes, I can" for item i is then expressed as

$$P_{ni}(1|\theta_n,\delta_i)=\frac{\exp(\theta_n-\delta_i)}{1+\exp(\theta_n-\delta_i)} \tag{1}$$

The probability that an officer *n* will respond with "no, I can't" for item *i* is then expressed as

$$P_{ni}(0|\theta_n,\delta_i)=1-p_{in}(1|\theta_n,\delta_i)=\frac{1}{1+\exp(\theta_n-\delta_i)} \tag{2}$$

Thus the odds ratio that an officer *n* can achieve item *i* is

$$\frac{P_{ni}(1|\theta_n,\delta_i)}{P_{ni}(0|\theta_n,\delta_i)} = \exp(\theta_n - \delta_i) \quad (3)$$

and the logarithm of the odds ratio, known as the "logit", is

$$\ln\frac{P_{ni}(1|\theta_n,\delta_i)}{P_{ni}(0|\theta_n,\delta_i)} = \theta_n - \delta_i, \quad (4)$$

which isolates the parameters of interest.

The person and item parameters can then be estimated through the response odds ratios in the data set using the formulation of Eq. (4).

In addition to dichotomous responses, the Rasch model can also be modified to be applicable to polytomous rating scale instruments, such as the five-point Likert scale (Andrich, 1978; Masters, 1982). The modified Rasch model decomposes a polytomous response into several dichotomous responses and formulates one rating scale problem into several binary-choice problems. That is, it assigns δ_{ik} as the value of the item parameter for rating category *k* to item i and assumes that Equation (5) refers to the probability of officer n responding with rating category *k* rather than *k*-1 to item *i*. Thus, we can model the log odds of the probability that an officer responds in category *k* for item i, compared with category *k*-1, as a linear function of the person parameter (i.e., the police officer's perceived ability in this study) θ_n and the relative parameter of category *k*, namely, δ_{ik}, for item i:

$$\ln\left[\frac{P_{nik}}{P_{ni(k-1)}}\right] = \theta_n - \delta_{ik} \quad (5)$$

Per Andrich's (1978) modification of the Rasch model for polytomous responses, two types of formulation are widely applied in assessing the values of item and person parameters: the rating scale Rasch model and the partial-credit Rasch model. The rating scale Rasch model is used for instruments in which the definition of the rating scale is identical for all items, whereas the partial-credit Rasch model is used when the definition of the rating scale differs from one item to another. The partial-credit Rasch model differs from the rating scale Rasch model in the possession of its own threshold parameters, F_{ik}, for each category *k* (Wright, 1977). This is achieved by a reparameterization of Equation (6):

$$\delta_{ik} = \delta_i + F_{ik} \quad (6)$$

and the partial-credit model can be demonstrated as

$$\ln\left[\frac{P_{nik}}{P_{ni(k-1)}}\right] = \theta_n - \delta_i - F_{ik} \tag{7}$$

The partial-credit model (Masters, 1982) is used for items where (1) credits are given for partially correct answers, (2) there is a hierarchy of cognitive demand on the respondents for each item, (3) each item requires a sequence of tasks to be completed, or (4) there is a batch of ordered response items with individual thresholds for each item. In assessing the police officer's ability to conduct red-light running enforcement, it is not necessary to assume that the rating scales of the items are the same, and thus we adopted the partial-credit model for our model formulation.

The consecutive Rasch approach assigns each item to its corresponding component latent trait exclusively, and separately measures each component latent trait with its own items like a unidimensional Rasch model.

2.2 Multidimensional Rasch model

The statistical problems of multidimensional item response models have been addressed with the development of the multidimensional random coefficients multinomial logit (MRCML) model (Adams, et. al, 1997; Wang, et. al, 2004). The MRCML advances sufficient flexibility to present a wide range of Rasch family models, including those that apply to scales having either yes/no or Likert-type responses. The MRCML model is a direct extension of the unidimensional random coefficients multinomial logit (RCML), which has been described in earlier papers (Adams, et al., 1997; Wang, et al., 1997) and is built up from a basic conceptual building-block. It assumes a set of d (d = 1,…,D) component latent traits that determine test performances (i.e., underlie the individuals' responses). The persons responding to a given item are indexed by n (n = 1,…, N). Then, the log odds of the probability a person's response in category k of item i (P_{ik}) compared to category $k-1$ ($P_{i(k-1)}$) as a linear function of latent ability on that dimension (θ_d), and the relative difficulty of category k (δ_{ik}) can be modeled as (Briggs and Wilson, 2003):

$$\ln\left[\frac{P_{nik}}{P_{ni(k-1)}}\right] = \theta_{nd} - \delta_{ik} \tag{8}$$

The θ_{nd} in equation (8) represents the latent ability of the person as a function of the dimension of ability mapped onto item i. The mapping then would indicate that item i was related to only latent component d, and the value of δ_{ik}, commonly called a "step difficulty", would indicate whether it was relatively easier or harder for a police officer to be classified as achieving category k-1 compared to k.

The general MRCML formulation for the probability of a response vector, $\mathbf{x}$, is (Cathleen, 2005):

$$P(\mathbf{x}_n;\boldsymbol{\delta}\,|\,\boldsymbol{\theta}_n)=\frac{\exp\left[\mathbf{x}_n'\left(\mathbf{B}\boldsymbol{\theta}_n-\mathbf{A}\boldsymbol{\delta}\right)\right]}{\sum_{\mathbf{z}\in\Omega}\exp\left[\mathbf{z}'\left(\mathbf{B}\boldsymbol{\theta}_n-\mathbf{A}\boldsymbol{\delta}\right)\right]} \tag{9}$$

where,

The individuals' positions in the *D*-dimensional latent space are described by the *D* by 1 column vector $\boldsymbol{\theta}=(\theta_1,\theta_2,...,\theta_D)$, $\boldsymbol{\delta}$ is the vector of calibrated item parameters and Ω is the set of all possible response vectors. We use **z** to denote a vector coming from the full set of response vectors while **x** denotes the one of interest. The matrices **A** and **B** are known as the scoring and design matrices respectively. The scoring matrix B allows the description of the score that is assigned to each response category *k* on each of the *D* component latent traits. The design matrix A is used to specify the linear combinations of the *D* ingredient parameters $\boldsymbol{\delta}$ to describe the behavior of the response categories to each item.

3. SAMPLE DESCRIPTION

A sample consisting of 358 policemen was randomly selected (126 from north Taiwan, 118 from the middle of Taiwan, 114 from south Taiwan); 92.2% (n = 330) of them are male and 7.8% (n = 28) are female. The age of the participant officers ranged from 22 to 55 years with an average of 37.45. Ages are further categorized into three groups in this study, including 108 young participants (age<36), 149 mid-age participants ($36\le$ age <46), and 101 senior participants ($46\le \text{age}\le 55$). The sample consisted of 297 (82.96 %) police officers (sergeants included) and 61 (17.04%) branch captains. In terms of age, gender, and rank, the characteristics of participants are not found to significantly deviate from those of population at $\alpha=0.05$.

4. MODEL DIAGNOSIS

4.1 Model selection

The multidimensional Rasch model is a nested model, if the variances associated with additional random effects are 0, it reduces to a unidimensional model (Rijmen and Briggs, 2004). The deviance index (G2) is equal to –2×ln(Likelihood) (Myers, et al., 2006). This fitness statistic follows a chi-square distribution, and a lower G2 indicates better fit. Using a likelihood ratio (LR) test, the relative fit of the unidimensional and multidimensional models can be compared. The LR statistic is equal to the difference of deviance indexes between these two models. It has been shown by Verbeke and Molenberghs (1997) that the asymptotic null distribution of this LR statistic is an equally weighted mixture of two chi-square distributions with a degree of freedom that equals the difference in the number of parameters estimated by the two models. The deviance for the multidimensional model was smaller than the unidimensional model. The LR statistic was significant at 0.01 level ($\chi^2=736$, df =14), which confirmed that in this investigation, the multidimensional model fits the data significantly better than the unidimensional model.

4.2 Analysis approach selection

Given the limited relevant literature, it is too premature to conclude the multidimensional approach is better than the consecutive approach. In order to verify which is better, we compared these two approaches before conducting Rasch analysis. Between the multidimensional Rasch model and consecutive Rasch model, two major measurements are commonly used to diagnose model precision and fitness: the reliability index and Akaike's Information Criterion (AIC).

A. Reliability

The reliability index (R_p) represents the percentage of reproducible observed response variance (Wright and Masters, 1982).

$$R_p = \frac{SA_p^2}{SD_p^2} \tag{10}$$

The denominator is the total person variability, which demonstrates how much people differ on the measure of interest. The numerator is the adjusted person variability, which is obtained by subtracting error variance from total variance. The adjusted person variability represents the reproducible part of this variability; that is, the amount of variance that can be reproduced by the Rasch model. This reliability index helps us examine whether or not the model is convincing and the material is replicable. A higher ratio indicates a higher reliability.

Table 2. Reliabilities comparison

Approach	ADM	ASC	APD	APHD	AAE
Consecutive	0.833	0.835	0.805	0.857	0.757
Multidimensional	0.897	0.867	0.831	0.941	0.844

Table 2 showed the reliabilities for consecutive and multidimensional approaches. Using multidimensional approach, the reliability for each dimension was higher than that obtained from the consecutive approach. This evidence demonstrated that the multidimensional approach yielded more precise measurement than did the consecutive approach.

B. The AIC test

The second method is on the basis of comparison of Akaike's Information Criterion (AIC) (Akaike, 1974). The Akaike's Information Criterion offers a relative measure of the information lost when a given model is used to describe reality. It describes the tradeoff between bias and variance in model construction. Given a data set, several competing models may be ranked according to their AIC, and the one having the lowest AIC is the best. The general form for calculating AIC is

$$AIC = -2\ln(Likelihood) + 2K \tag{11}$$

where K is the number of parameters in the model, including (a) variance parameters, (b) the item parameters, (c) the step parameters[1], and (d) the covariance parameters for

[1] The number of step parameters that have been estimated for each item is one less than the number of

multidimensional Rasch model.

The figures in Table 3 provided the evidence that the multidimensional approach fits the data better, as the AIC for the multidimensional approach is lower than that for consecutive approach. Both criteria confirmed the advantage of the multidimensional approach, so in this study the multidimensional approach was applied to address further investigations.

Table 3. AIC comparison

Approach	No. of parameters	G^2	AIC
Consecutive	125	20066	20316
Multidimensional	135	19359	19629

Note: G^2 is equal to $-2\times\ln(Likelihood)$.

5. FINAL SCALE

5.1 Evaluation of fit to the Rasch model

The fit of the Rasch model, which can be expressed by either the form of responses observed for each candidate on all items (person fit) or the form for each item on all persons (item fit), is the degree of match between the form of observed responses and the modeled expectations. Indices of fit statistics estimate the extent to which responses show adherence to the modeled expectations. Two kinds of fit statistics, namely infit and outfit statistics can be expressed in two forms: unstandardized mean squares and standardized *t* values, and are used to determine how well the data meet the requirements of the model. Infit and outfit statistics are sensitive to violations of assumptions of the Rasch model, such as unexpected variation in response forms (e.g., a person with little ability endorses a severe symptom) or unequal slopes (e.g., item discrimination) across item response functions (Bond and Fox, 2001). Infit statistics, which are more sensitive to irregular inner forms, have more emphasis on unexpected responses near a person's or item's measure; outfit statistics, tending to be influenced by off-target observations, place more emphasis on unexpected responses far from a person's or item's measure. Compared with the outfit statistics, that are more sensitive to the influence of outlying scores, the infit *t* statistics give more weight to on-target performances.

As the estimation of fit is concerned, the calculation of a response residual (Y_{ni}) for each item *i* of person *n* is encountered. In other words, how far does the actual response (X_{ni}) deviate from Rasch model expectation (E_{ni}) (Wright and Master, 1982)?

$$Y_{ni} = X_{ni} - E_{ni}, \text{ and} \tag{11}$$

$$E_{ni} = \sum_{k=0}^{K-1} k(P_{nik}) \tag{12}$$

thresholds for the item (Wu, et al., 2007).

where K is the available response categories for observations and P_{nik} is the probability of person n being observed in category k on item i.
The outfit statistic is an unweighted average of the standardized residual (Z_{ni}) variance across both person and item as following equation:

$$\text{Means-square Outfit} = (\sum_{n=1}^{N} Z_{ni}^2) / N \tag{13}$$

Table 4. Estimates of model fit statistics for 20-item scale

Item	δ_i	Outfit		Infit	
		MNSQ*	*t* statistic	MNSQ	*t* statistic
1	0.292	1.07	1.0	1.08	1.0
2	-1.057	0.93	-1.0	0.93	-1.0
3	0.238	1.00	0.0	1.01	0.1
4	0.528	0.98	-0.3	0.99	-0.1
5	0.298	0.86	-1.9	0.89	-1.6
6	**-0.829**	**1.21**	**2.6**	**1.21**	**2.6**
7	0.368	1.03	0.5	1.05	0.7
8	0.162	0.86	-1.9	0.86	-1.9
9	0.194	0.91	-1.2	0.93	-0.9
10	0.379	0.93	-0.9	0.98	-0.2
11	-0.146	1.04	0.5	1.05	0.7
12	-0.312	0.97	-0.4	1.01	0.1
13	-0.116	1.11	1.5	1.14	1.8
14	-0.041	1.10	1.3	1.11	1.5
15	0.157	0.86	-1.9	0.86	-1.9
16	-0.116	1.01	0.1	1.03	0.5
17	-0.189	0.94	-0.8	1.02	0.3
18	-0.737	0.93	-1.0	0.98	-0.3
19	0.481	1.11	1.5	1.13	1.6
20	0.445	1.10	1.3	1.13	1.6

* MNSQ is the fit statistic in the form of mean square.

where Z_{ni} is obtained by standardizing the residual that is the deviation of actual response from modeled expectations, and N is the number of items/persons. The standardized residual, Z_{ni}, is calculated:

$$Z_{ni} = Y_{ni} / (W_{ni})^{1/2}, \tag{14}$$

where $W_{ni} = \sum_{k=o}^{K} (k - E_{ni})^2 P_{nik}$ is the variance of X_{ni}. These standard residuals are squared and summed to form a chi-square statistic:

$$\chi^2 = \sum_{n=1}^{N} Z_{ni}^2 \tag{15}$$

However, infit statistics are generally preferred as they are weighted locally and, thus, are less susceptible to outlier influences (Bond and Fox, 2001). Residuals are weighted by their individual variance (W_{ni}) to diminish the impact of off-targeted improbable responses:

$$\text{Means-square Infit} = \sum_{n=1}^{N} W_{ni} Z_{ni}^2 / \sum_{n=1}^{N} W_{ni} \tag{16}$$

Both infit and outfit statistics can be expressed as the form of mean square with an expected value of unity and reported by the standardized form as a *t* statistic as well. Items for which the fit mean square statistic (MNSQ) is smaller than 0.8 or larger than 1.2 (Linacre and Wright, 1994; Wang et al., 2004), and for which the fit *t* statistic is smaller than -2.0 or larger than 2.0 (Smith, 1992) are considered to be a poor fit. To process the collected data, the Acer ConQuest was applied to this study. The marginal maximum likelihood estimation proposed by Bock and Aitkin (1981) is estimated using the EM algorithm. According to the estimation results of multidimensional approach (as shown in Table 4), the statistics of infit mean square (MNSQ) and *t* statistics for all individual items fell within the expected range of 1 ± 0.2 and ± 2, respectively, except for Item 6. Thus, Item 6 is removed from further analyses.

5.2 Differential item functioning analysis

Subsequent to removing Item 6, the differential item functioning (DIF) analysis was applied to evaluate scale stability across samples. The ability perception measures estimated by the Rasch model provide the opportunity to compare the differences among respondent groups. This comparison ensures that different group characteristics do not affect interpretation of the total measures of ability and thus equal comparisons can be made. The estimation of differential item functioning (DIF) involves comparing analyses conducted separately within each group (Holland and Wainer, 1993). When developing new tests, items displaying DIF have long been recognized as a potential source of bias in person measurement. Furthermore, since gender has long been recognized as the most important factor for developing an unbiased scale (David, et al., 2005, Mackintosh, 2006), DIF analyses were applied to examine whether differences in item difficulty exists between male and female police officers.

Wu et al. (2007) proposed that the between-groups differences in item difficulties larger than 0.2 logit somewhat presents the evidence of DIF. However, Cohen (1977) showed that the criterion for explaining significant differences between groups was 0.5 logit, at least in the difference of item difficulty estimates. This argument was consistent with Wright and Douglas' (1975) suggestion that differences in item parameters less than 0.5 logit had little effect on the accuracy of tests. Moreover, Smith (2004) and Wang et al. (2006) pointed out that differences smaller than 0.5 logit are hard to be detected. Thus, 0.5 logit is

commonly used as an acceptable criterion of removing individual items for remarkable DIF. However, investigation indicated that the differences between male and female police officers for all items were less than 0.5 logit, which implied the effects of DIF for the 19 items were not substantively important and they could be retained.

5.3 Final scale evaluation

Table 5. The correlations matrix

	ADM	ASC	APD	APHD	AATE
ADM	1	0.581	0.730	0.604	0.702
ASC	**0.645**	1	0.699	0.673	0.430
APD	**0.805**	**0.641**	1	0.638	0.777
APHD	**0.508**	**0.570**	**0.539**	1	0.506
AAE	**0.672**	**0.640**	**0.667**	**0.660**	1

Note: The 19-item scale correlations are shown below the diagonal; the 20-item scale correlations are shown above the diagonal.

The procedure for the above analyses of fit statistics and DIF suggest that Item 6 was a removable item. In order to identify if it is without losing crucial information because of removing Item 6, by comparing the final 19-item scale with the designed 20-item scale, we evaluated the goodness of fit of the final scale. The reliabilities of the designed scale for each dimension were 0.897, 0.867, 0.831, 0.941 and 0.844, whereas those for the final scale were 0.893, 0.872, 0.911, 0.905, and 0.893, respectively.

After removing Item 6, the average reliabilities increased from 0.876 to 0.895 and only two of five dimensions became smaller. This evidence confirmed that the reliabilities of latent constructs were not harmed by this item elimination. Moreover, as shown in Table 5, six of ten correlations between dimensions for the final scale were smaller than those for the designed scale. Nevertheless, those decreases were marginal (< 0.1), which implied that in applying the multidimensional Rasch model Item 6 can be removed without meaningfully decreasing correlations among latent straits.

Table 6 is the estimate result for the final scale. The fit statistics including infit and outfit for all items were within the acceptable range, and the *t* statistics all were insignificant, which indicated that the final 19-item scale was consistent with the assumptions of the Rasch model. Although the last form of the scale was brief, it estimates the true relations among underlying constructs. All the above evidence points to the fact that the shortened final 19-item scale possessed good attributes to assess police officers' perceived abilities to enforce red-light running.

Table 6. Final estimates of model fit statistics

Construct / Item	θ [b]	Person with enough ability (%)[c]	δ_i [d]	Outfit		Infit	
				MNSQ	*t* statistic	MNSQ	*t* statistic
AMD	0.494	**62.35**					
2		81.8	-0.684	0.92	-1.0	0.95	-0.6

1		70.2	-0.176	0.91	-1.2	0.97	-0.4
4		48.7	0.413	1.07	1.0	1.08	1.0
3		48.7	0.447	1.1	1.3	1.09	1.2
ASC	0.249	**56.8**					
5		88.5	-0.987	0.92	-1.1	0.94	-0.9
7		44.3	0.364	0.96	-0.5	0.96	-0.5
8		37.7	0.624	0.99	-0.1	1.00	0.0
APD	0.619	**65.0**					
13		71.8	-0.322	1.05	0.7	1.07	0.9
12		71.8	-0.174	0.93	-0.9	0.95	-0.6
11		66.7	-0.129	0.97	-0.3	0.96	-0.5
9		60.7	0.215	1.09	1.2	1.10	1.2
10		53.8	0.410	0.93	-0.9	0.96	-0.5
APHD	-0.201	**46.5**					
14		47.2	-0.029	0.92	-1.1	0.95	-0.6
15		41.2	0.211	1.04	0.5	1.09	1.1
16		51.2	-0.182	0.89	-1.4	0.91	-1.2
AAE	0.141	**47.7**					
17		71.7	-0.710	1.04	0.6	1.05	0.7
18		42.5	0.174	1.14	1.8	1.10	1.4
20		42.5	0.215	1.04	0.6	1.02	0.3
19		34.2	0.321	1.13	1.6	1.13	1.6

[a] This statistic is the average enforcement ability of all participating police for latent constructs.
[b] The bold is the average percentage across items in the same construct.
[c] The mean value of δ_i is 0 logit.

6. RESULTS

As stated, the Rasch model suggests that the answer to an item can be explained by two parameters: the difficulty of the item and the ability of the person. First, in terms of item parameters, according to the definition of item difficulty, the lower the item difficulty, the easier the task will be accomplished. In other words, large δ_i represents greater difficulty than does small δ_i. The coefficient δ_i measures the difficulty of an item, which cannot be reached by traditional measurement. The magnitude of δ_i parameter can be used to measures the relative distance of performance between items, so that we can directly apply this knowledge to related applications, such as arranging a training curriculum.

Thus, among the items of the AMD construct, the most difficult item was Item 3 followed by Item 4, while Items 2 was the easiest item. It indicated that recording the escaping red-light running vehicle's distinguishing characteristics was a difficult task. As to the constructs of ASC, APD, APHD and AAE, the most difficult items were Items 8, 10, 15

and 19, respectively. On the contrary, the easiest items were Items 5, 13, 16 and 18, respectively. These results suggested that intercepting escaping vehicle (Item 8), emotional intelligence (Item 10), sore waist (Item 15) and weather condition (Item 19) were the four serious problems experienced by police officers when conducting red-light running.

Logit	AMD		ASC		APD		APHD		AAE	
	person	item	person	item	person	item	person	item	person	item
4										
							X			
					X		X			
3	X		X		X		X			
	X		X		XX		X			
	X		X		X		XX		X	
	XX		X		XXX		XXX		X	
	XX		XX		XXX		X		XX	
2	XXXX		XXX		XXXX		XX		XX	
	XXXXX		XXX		XXXXX		XXX		XXXX	
	XXXXX		XXX		XXXXXX		XXXXX		XXXX	
	XXXXXXX		XXXXXX		XXXXX		XXXX		XXX	
	XXXXXXXX		XXXXXXX		XXXXXXXXX		XXXX		XXXX	
1	XXXXXXXX		XXXXXXXX		XXXXXXXX		XXXX		XXXXXX	
	XXXXXXX		XXXXXXXXXX		XXXXXXX		XXXXXX		XXXXXXX	
	XXXXXXXX		XXXXXXXX	8	XXXXXXXX		XXXXXX		XXXXXXX	
	XXXXXXXX	3, 4	XXXXXXX	7	XXXXXXXX	10	XXXXXXX		XXXXXXXXX	19
	XXXXXXXX		XXXXXXXX		XXXXXXX	9	XXXXXXXX	15	XXXXXXXX	18, 20
0	XXXXXXXXX		XXXXXXXXX		XXXXXX	11	XXXXX	14	XXXXXXXX	
	XXXXXXXXXX	1	XXXXXXXXX		XXXXXXXXX	12, 13	XXXXXX	16	XXXXXXXXX	
	XXXXX		XXXXXXXX		XXXXXX		XXXXXX		XXXXXXXXX	
	XXXXXXXX	2	XXXXXX		XXXXX		XXXXXX		XXXXXXX	17
	XXXX		XXXXXXX		XXXX		XXXXX		XXXXXX	
-1	XXXX		XXX	5	XXXX		XXXXX		XXXX	
	XXX		XX		XX		XXXXXX		XXXXXX	
	XX		XXXX		X		XXX		XX	
	X		XX		X		XXXX		XX	
-2			X		X		XXX		XXX	
			X				XX		XX	
			X				XXX		X	
							XX			
							XX			
-3							XX			
							XX		X	
							X			
							X			
							X			
-4							X			

Note: Each 'X' represents 3.0 persons.

Figure 2. Person map of item for the respondents

Next, as person parameters concerned, Rasch analysis converts the original ordinal raw scores into the measurements on an interval scale and provides the opportunity to meaningfully compare officers' abilities conducting red-light running enforcement with item difficulties on a consistent basis. The mean value of all item difficulties for each component construct is anchored at 0 logit, thus the positive mean ability values of AMD

(0.494 logit), ASC (0.249 logit), and APD (0.619 logit) for all participant officers indicated they were confident in their mental abilities, stopping and chasing vehicle abilities, psychological abilities, and abilities under abnormal environment when conducting red-light running enforcement (see Table 6). On the contrary, the negative mean value (-0.201 logit) of APHD implied that traffic police officers did not have enough confidence in their physical abilities to conduct red-light running enforcement. In view of the AAE construct, the average enforcement ability value was positive (0.141 logit), but less than 50% of police were capable of performing the tasks. It can be categorized as the second difficult construct to achieve.

Figure 2 further plots the distributions of item difficulties and person abilities on red-light running enforcement for the five constructs. It provides a visual illustration to demonstrate the relative positions between a person's abilities and item difficulty. The vertical dash-line represents a scale with the logit unit, and the numbers on the right hand side of the logit scale are the corresponding items whose positions are arranged from low to high in order of item difficulties. Similarly, the distributions of persons' abilities are displayed on the left hand side of the logit scale. The officer located at a higher position suggests a higher ability on a certain construct to conduct red-light running enforcement.

When a police officer is located on the same level as a certain item/task, he/she will have a probability of 50% to accomplish this task well; when a police officer lies higher than a certain item/task, he/she will have a probability of more than 50% to handle this task well. Therefore, from the item-person maps for the five constructs in Figure 2, we can calculate the percentage of police with enough ability to complete the task, which is displayed on the third column of Table 6.

The second column of Table 6 shows the average enforcement ability of all participant police for certain constructs, which is calculated by averaging the person parameter estimate (θ_n) across all respondents. The estimates of item difficulty (δ_i) for each task are listed in the fourth column. The results of average enforcement ability showed that police had the highest APD, which was followed by the AMD. On average, more than 60% of police were capable of performing the tasks for these two constructs. However, there were some tasks not well controlled by police. For example, of four tasks measuring police AMD, Items 3 and 4 were possessed by less than half of capable police, indicating that most police are weak in detecting escape intentions and recording the escaping vehicle's distinguishing characteristics. In addition, the positive and higher *bi*s confirmed these two tasks were difficult. As to the APD, most police agreed that they can conduct the tasks correctly, thereby their execution was convincing. Yet, police needed to improve their Emotional Intelligence (EI), as they were weak in "EQ Management" (Item 10).

The worst performing were the APHD and the AAE constructs. On average, less than half of police possessed enough abilities for the tasks in these two constructs. All tasks for the APHD construct were poorly performed by police. The physical condition of Taiwan police was worrisome, and needed attention by the authorities. This may be a common circumstance for other countries, especially in safe and stable areas. In addition, the police perceived that their enforcement abilities were influenced by weather, temperature, and traffic conditions. Of them, enforcing during a cold winter (Item 19) was viewed as extremely difficult (0.321 logit) and only 34.2% of all participants had enough ability to accomplish it well. When conducting red-light running enforcement during extremely cold

winters police officers are often exposed to high winds which slow their actions and thereby decrease enforcement efficiency. Intercepting an escaping vehicle (Item 8) was the second most difficult (0.624 logits) of all tasks as only 37.7% of participant police officers had high ability to accomplish it well, so it needs expert police to conduct the difficult task of *intercepting an escaping vehicle safely*.

7. DISCUSSIONS

Factor analysis demonstrated that the ability of conducting red-light running enforcement was found to be comprised of five component constructs based on 20 design items. One of these was removed due to model fit issues without meaningfully decreasing correlation among the five dimensions (from 0.722 to 0.720). Applying the multidimensional approach of Rasch modeling to examine the issues of interest in this study is justified. The evidence of less error and higher reliability confirmed that the multidimensional Rasch model indeed develops a more precise measurement for our research. Although the final form of the scale is shorter, it estimates the true relations among underlying constructs and lends support to establishing invariance of the measure across groups. Moreover, the estimated parameters of ability and item difficulty can be further applied as follows.

First, clustering persons according to their abilities could provide relative information about excellent and unqualified persons, which is helpful for supervisors to assign duties to the right persons, as well as to design a suitable training program for those unqualified persons. The thresholds required for each construct were assumed to be the highest value of item difficulty among all items, which were respectively 0.447, 0.624, 0.41, 0.211 and -0.321 logits for the five constructs. Among 358 participating police officers, 223 (62.4%), 268 (74.9%), 214 (59.8%), 246 (69.7%) and 194 (54.2%) persons were qualified for conducting the five constructs' tasks respectively, and the other participants for the five constructs were classified as unqualified and who needed more education and training if they were to be allowed to conduct such missions.

Table 7. The estimated results of regression models

Model/Construct	AMD	ASC	APD	APHD	AAE
Constant	0.428 (2.143*****)	-0.545 (-2.798*****)	0.350 (0.615*****)	-1.592 (-6.327*****)	-1.037 (-5.526*****)
Gender (male=1, female=0)	0.231 (1.271)	0.299 (1.691)	0.273 (1.416)	0.363 (1.607)	0.284 (1.610)
Rank (captain=1, others=0)	0.520 (0.405)	0.165 (1.320)	0.073 (0.538)	-0.070 (0.431)	0.142 (1.137)
Aged 36-45	-0.506 (-4.413*****)	-0.446 (-3.988*****)	-0.637 (-5.236*****)	-1.227 (-8.509*****)	-0.696 (-5.526*****)
Aged 46-55	-0.560 (-4.382*****)	-0.537 (-4.315*****)	-0.822 (-6.070*****)	-1.616 (-10.054*****)	-0.774 (-6.239*****)
R^2	0.093	0.187	0.168	0.449	0.323
Adjusted R^2	0.083	0.178	0.159	0.443	0.316

Note: The numbers in parentheses are *t*-statistics.
*significant at 5% level

Next, the interval scale measurements of the five constructs obtained from Rasch

analysis provided a convenient opportunity to explore whether the police officers with different characteristics will have different abilities to conduct red-light running enforcement. Thus, five multiple regression models were applied to investigate the effect of gender, rank, and age of police officers on these abilities. The study results shown in Table 7 indicated no significant difference between the two rank groups or the two gender groups. Yet, the officers aged 36-55 had significantly lower abilities for all ability constructs than those aged under 36.

In this study, moderate to high correlations were found between each pair of the five ability constructs, as shown in Table 5. It implied that police officers qualified on each ability construct were also qualified with at least one other construct. Thus, only the 35.2% (n=126) of participant police officers had enough or the five abilities and were considered as the qualified to conduct red-light running enforcement, and those (12.3%, n=44) who failed to pass all five minimum required abilities were definitely not qualified to conduct red-light running enforcement. As to the remainders (52.5%) who failed to pass between one and four of the required abilities, they should be forced to take further training and education to enhance their abilities in which they are weak if they are allowed to conduct red-light running enforcement.

Age was highly correlated with the seniority of police officers in Taiwan and found to be the most influential factor to determine police officer ability to conduct red-light running enforcement. Among the police officers qualified for conducting red-light running enforcement, only 15.1% (n=19) and 37.3% (n=47) of them were over 45 and 36-45 years of age, respectively; however, 36.4% (n=16) and 50% (n=22) of those unqualified to conduct red-light running enforcement were over 45 and 36-45 years of age, respectively. It further indicated that young police officers were more appropriately qualified than older police officers to conduct red-light running enforcement.

8. CONCLUSION

The Rasch measurement provides the difficulty of item and the ability of person to interpret the response of a participant to an item, and converts ordinal responses into an equal-interval logit scale. Next, the person ability contributes reasonable judgment to discriminate qualified persons from unqualified persons to conduct red light running enforcement as well as to compare perceived enforcement abilities for different groups. Thus, this study used the Rasch model to assess the perceived enforcement ability to conduct red-light running enforcement. The evidences demonstrated that the multidimensional Rasch approach developed a more appropriate questionnaire than unidimensional approach and consecutive Rasch approach did.

The results found that police had the highest ability on tasks that measure APD, which is followed by tasks that measure AMD and ASC. Taiwan police agreed that they can conduct the enforcement correctly; however, there were some tasks not well controlled by police. According to the study results and further analyses, we found that some missions for conducting red-light running enforcement were not well controlled by traffic police in Taiwan, especially in intercepting escaping vehicles (Item 8), emotional intelligence (Item 10), APHD, and enforcing under abnormal environment (Items 18, 19, 20). It indicated that some improvement is needed in order to conduct the mission more effectively and some strategies and programs were suggested as follows.

1. Continuing professional development programs to improve police officers' enforcement abilities:

According to the study results, only 46.5%, 47.7%, and 56.8% of traffic officers had enough APHD, AAE, and ASC, respectively, to be qualified to conduct red-light running enforcement in Taiwan. It implies that the effectiveness of the enforcement would not be significantly improved unless some continuing professional development programs could be implemented to enhance those abilities. These programs should focus on improving the critical weaknesses of police officer abilities in intercepting escaping vehicle, enforcing under extremely cold weather, and physical condition. Considering the unique occupational characteristics, police often conduct their duties under risk. Maintaining a good physical condition is not only for improving police enforcement performance, but to also help them deal with all kinds of unexpected troublesome or risky situations. In view of ASC, police officers should be educated and trained not only in how to stop, chase, and intercept DWI vehicles, but also in how to protect themselves and other road users.

2. Applying video equipments to promote enforcement performance

As we found, to perceive an intention to escape (Item 3) and to chase and intercept an escaped vehicle (Item 8) are difficult tasks. Considering the limitation of reading others' minds and the hardship of chasing fleeing violators, instead of chasing escaping violator we suggest that police set up video recording cameras in the enforcement area to collect evidence about escaping vehicles. As stated previously, the current physical condition of police is weak and enforcement performance is influenced by weather. If video is used to collect evidence of violations, red-light running enforcement will be more successful and less risky to conduct.

3. Designing an emotional intelligence education program

Excessive enforcement is harmful to the reputation of the police, and some cases even result in more serious consequences such as protests and riots. A lot of improper enforcement starts when police are exasperated by the challenges of the violators.

When facing an irritable violator, it is arduous to keep in good emotion all the time. To ask a violator to maintain a good attitude is difficult, so we would rather expect police officers to possess high EQ. Therefore, an education on the work-load pressure and the methods of accommodation is helpful for enhancing police EQ.

4. Efficient use of human resources

Young police officers are found to be more appropriate to conduct red-light running enforcement in this study. Thus, some incentive policies should be established to encourage young officers to devote themselves to traffic enforcement. Furthermore, unqualified officers should be educated and trained with priority in order to help them do the job well and protect themselves from being hurt if they are allowed to conduct traffic law enforcement.

Actually, the duty of a traffic police is multi-tasks. He/She has to treat any kind of violations, not only red light running, but also drunk driving, speeding, etc. Besides, it can be that a policeman with high ability on red light running enforcement has less ability on speeding and/or drunk driving, etc. Therefore, a further research that integrated in larger context in which the police would have abilities to carry out any required enforcement is truly needed. However, for speeding or/and drunk driving, it is necessary to use an instrument for

collecting violated evidences, and their dangerous levels are much higher than other enforcements. Therefore, we suggested that the items of questionnaires for these kinds of tasks should take these characteristics into account in further studies.

REFERENCES

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

Akaike, H. (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716-723.

Andrich, D. (1978) A rating formulation for ordered response categories. Psychometrika, 43, 357-374.

Bock, R. D., and Aitkin, M. (1981) Marginal maximum likelihood estimation of item parameters: an application of the EM algorithm. Psychometrika, 46, 443-449.

Bond, T. G., and Fox, C. M. (2001) Applying the Rasch Model: Fundamental Measurement in the Human Sciences. Mahway, NJ: Erlbaum.

Briggs D. C., and Wilson M. (2003) An introduction to multidimensional measurement using Rasch models. Journal of Applied Measurement, 4, 87–100.

Burkey, M. and Obeng, K. A (2004) Detailed Investigation of Crash Risks Reduction Resulting from Red Light Cameras in Small Urban Areas, Transportation Institute, North Carolina A&T University, Greensboro.

Cathleen A. K. (2005) Constructing Measurement Models for MRCML Estimation: A Primer for Using the BEAR Scoring Engine. 2005-04-02, Berkeley Evaluation & Assessment Research Center, Graduate School of Education University of California, Berkeley.

Chang, H. L., and Shih, C. K. (2007) How do the traffic police perceive their ability for red light running enforcement？- An application of the Rasch measure, Journal of the Eastern Asia Society for Transportation Studies, Vol. 7, 2623-2638.

Chang, H. L., and Shih, C. K. (2012) How Do The Traffic Police Perceive Their Ability for Red Light Running Enforcement?－An Application Of The Rasch Measure. Journal of the Eastern Asia Society for Transportation Studies, 7, 2623-2638.

Cohen, J. (1977) Statistical Power Analyses for the Behavioral Sciences. revised ed. Hillsdale, NJ: Lawrence.

David R. S., Christopher W. K., Roger L. G. and John, S. (2005) Isolating a primary dimension within the Cook-Medley hostility scale: a Rasch analysis. Personality and Individual Differences, 39, 21-23.

De Ayala, R. J. (1994) The influence of dimensionality on the graded response model. Applied Psychological Measurement, 18, 155 – 170.

Erke, A. (2009) Red light for red light cameras? A meta-analysis of the effects of red-light cameras on crashes. Accident Analysis and Prevention, 40 (1), pp. 167–173.

Holland, P. W. & Wainer, H. (1993) Differential Item Functioning. Hillsdale, NJ7 Lawrence Erlbaum.

Hsiung, P. C., Fang, C. T., Chang, Y. Y., Chen, M. Y, and Wang, J. D. (2005) Comparison of WHOQOL-BREF and SF-36 in patients with HIV infection. Qual Life Res; 14: 141-150.

Linacre, J. M., and Wright, B. D. (1994) Reasonable mean-square fit values. Rasch Measurement Transactions, 8, 370.

Mackintosh M.A., Earleywin M., and Michael E. D. (2006) Alcohol expectancies for social facilitation: A short form with decreased bias. Addictive Behaviors, 31, 1536-1546.

Masters, G. N. (1982) A Rasch model for partial credit scoring. Psychometrika, 47,

149-174.
Myers, N. D., Wolfe, E. W., Feltz , D. L., and Penfield, R. D. (2006) Identifying differential item functioning of rating scale items with the rasch model: an introduction and an application. Measurement in Physical Education and Exercise Science, 10 (4), 215-240.
Nattaporn, Y. (2004) Evaluation update of red light camera program in Fairfax County, Virgina. Master thesis, Science in Civil and Environmental Engineering. Blacksburg, Virginia.
Rasch, G. (1960). Probabilistic Models for Some Intelligence and Attainment Tests. Denmarks Paedagogiske Institute, Copenhagen.
Rijmen, F., and Briggs, D. C. (2004) Multiple person dimensions and latent item predictors. In: P. D. Boeck and M. Wilson (Eds.), Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach, pp.247-65, Springer, New York.
Shin K. and Washington, S. (2007) The impact of red light cameras on safety in Arizona. Accident Analysis & Prevention, 39 (6), 1212–1221.
Skevington, S. M., Lotfy, M., O'Connell, K. A. (2004) The World Health Organization's WHOQOL-BREF quality of life assessment: psychometric properties and results of the international field trial. A report from the WHOQOL group. Qual Life Res; 13: 299–310.
Smith, R. M. (1992) Application of Rasch Measurement. Sacramento: JAM Press.
Smith, R. M. (2004) Detecting item bias with the Rasch model. Journal of Applied Measurement, 5, 430-449.
Verbeke G., and Molenberghs G. (1997) Linear Mixed Models in Practice: A SAS-Oriented Approach, Lecture Notes in Statistics 126, Springer-Verlag, New York.
Wang, W., Wilson, M, and Adams, R. (1997) Rasch models for multidimension-ability between items and within items. In G. Engelhard and M. Wilson (Eds.), Objective Measurement: Theory into Practice, Vol. 4. Greeewich, CN: Ablex Publishing.
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 (1), 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. (1977) Solving measurement problems with the Rasch model. Journal of Educational Measurement, 14, 97-115.
Wright, B. D., and Douglas, G. (1975) Best test design and self-tailored testing. MESA Memorandum No. 19, Available at http://www.rasch.org/memo19.pdf.
Wright, B. D., and Masters, G. N. (1982) Rating Scale Analysis, MESA, Chicago.
Wu, M. L., Adams, R. J., and Wilson, M. R. and Haldane, S. A. (2007) ACER ConQuest: Generalized item response modeling software, Computer program version 2. Melbourne: ACER Press.