anales de psicología, 2017, vol. 33, nº 3 (october), 530-537
http://dx.doi.org/10.6018/analesps.33.3.238621

ISSN print edition: 0212-9728. ISSN web edition (http://revistas.um.es/analesps): 1695-2294

An investigation of enhancement of ability evaluation by using a nested logit model for multiple-choice items

Liu Tour[1], Wang Mengcheng[2], and Xin Tao[3*]

1 School of Education Science, Tianjin Normal University, Tianjin (China).
2 Center for Psychometric and Latent Variable Modeling, Guangzhou University, Guangzhou (China).
3 Collaborative Innovation Center of Assessment toward Basic Education Quality (CICA-BEQ), Bejing Normal University, Beijing (China).

Título: Una investigación de la mejora de la capacidad de evaluación mediante el uso de un modelo logit anidado para items de elección múltiple.
Resumen: Los items de elección múltiple se han usado ampliamente en tests psicológicos y educativos. Este estudio investiga si los items de elección múltiple tiene ventajas sobre los items dicotómicos o sobre la evaluación de rasgo latente. Un modelo de respuesta al item, con un modelo logit anidado, logístico 2-parámetros (2PL-NKM), fue usado para ajustar los datos de elección múltiple. Los estudios de simulación y empíricos indicaron que la precisión y la estabilidad de la estimación de capacidad mejoró usando el modelo de elección múltiple en contraposición al modelo dicotómico, debido a la mayor información incluida en los items distractores de la elección múltiple. Pero la precisión y la capacidad de estimación mostró pequeñas diferencias en items de cuatro elecciones, cinco y seis elecciones. Además, el modelo 2PL-NLM puede extraer más información respondientes de bajo nivel que de los de alto nivel, debido a que tienen conductas de elección con más distractores. En el estudio empírico, los respondientes en diferentes niveles de rasgo fueron atraídos por diferentes distractrores del Test de Vocabulario chino en el primer grado, usando trazos cambiantes en la probabilidad de distractor a partir de 2PL-NLM. Esto sugiere que las respuestas de los estudiantes a diferentes niveles puede reflejar un proceso evolutivo de vocabulario en los estudiantes.
Palabras clave: items de elección múltiple; modelo logit anidado; información distractora; capacidad de evaluación.

Abstract. Multiple-choice items are wildly used in psychological and educational test. The present study investigated that if a multiple-choice item have an advantage over a dichotomous item on ability or latent trait evaluation. An item response model, 2-parameter logistic nested logit model (2PL-NLM), was used to fit the multiple-choice data. Both simulation study and empirical study indicated that the accuracy and the stability of ability estimation were enhanced by using multiple-choice model rather than dichotomous model, because more information was included in multiple-choice items' distractors. But the accuracy of ability estimation showed little differences in four-choice items, five-choice items and six-choice items. Moreover, 2PL-NLM could extract more information from low-level respondents than from high-level ones, because they had more distractor chosen behaviors. In the empirical study, respondents at different trait levels would be attracted by different distractors from the Chinese Vocabulary Test for Grade 1 by using the changing traces of distractor probabilities calculated from 2PL-NLM. It is suggested that the responses of students at different levels could reflect the students' vocabulary development process.
Key words: multiple-choice item; nested logit model; distractor information; ability evaluation.

Introduction

Multiple-choice items are widely used in cognitive tests, aptitude tests, educational tests and some intelligence tests, since Frederick J. Kelly's introduction in 1914. A multiple-choice item usually includes one correct option and several incorrect options (distractors). The correct option makes a multiple-choice item have an objective scoring standard as well as a dichotomous item, rather than an open-ended item or an essay-type item. While the several incorrect options may distract respondents and decrease the guessing behaviors of respondents. Multiple-choice item may have some advantages over dichotomous item, but writing a multiple-choice item is more complicated than writing a dichotomous item. Also it will increase the cognitive loads of respondents. Consequently, if multiple-choice items can not make estimation of ability (latent trait) more accurate, it will be economical to use dichotomous items. Or if distractor can not provide more information about respondents, it will be better to credit multiple-choice item as binary item.

*** Correspondence address [Dirección para correspondencia]:**
Xin Tao. Collaborative Innovation Center of Assessment toward Basic Education Quality, Beijing Normal University, Beijing, (P. R. China).
E-mail: mikebonita@sina.cn

Distractor Information

Including several distractors is a distinctive feature of a multiple-choice item, compared with a dichotomous item. Various approaches have been proposed by many researchers to extract information from distractors of multiple-choice items. For one thing, some strategies of multiple-choice item construction were taken into consideration (Haladyna& Downing, 1989; Haladyna, Downing, & Rodriguez, 2002; Tamir, 1971, 1989). Briggs, Alonzo, Schwab, and Wilson (2006) suggested constructing multiple-choice items with ordered options which could seek to more diagnostic information. Liu, Lee, and Linn (2011) believed a multiple-choice item needed some explanatory components as a new tier, following a typical multiple-choice item. In addition, the optimal number of item options had been discussed by Haladyna and Dowing (1993). For another thing, technical treatments were good ways to mine the potential information from distractors without constructing a new test. To assign different weights to options was one way (Davis & Fifer, 1959). To create an augmented data matrix transformed from a raw response matrix by using some special scoring rules was another (Luecht, 2007). And also some indices, such as distractor selection ratio and point-biserial correlation, based on certain statistical models were useful (Attali & Fraenkel, 2000; Love, 1997).

The information extracted from distractor can be used as auxiliary information in many ways. In person-fit area, Wollack (1997), Drasgow, Levine, and Williams (1985) were used distractor information to detect aberrant response behaviors. Kim (2006) found that linking procedure was improved when distractor information was added in. Roediger and Marsh (2005) thought that distractors could bring in some psychological consequences. Besides, the differential functioning relative to distractors, namely differential distractor functioning (DDF), could also provide some explanation for detecting measurement bias (Green, Crone, & Folk, 1989; Penfield, 2011; Suh & Bolt, 2011; Suh & Talley, 2015).

Multiple-Choice Item Modeling

To achieve accurate ability evaluation is an ultimate goal of all educational and psychological tests. However, test score is not a good indicator to support that distractor information can enhance ability evaluation. Sigel (1963) found no relationship between error patterns and respondents' scores. Jacob and Vandeventer 's (1970) study showed that types of error and total score were related. But they still did not found the distractor information could improve the ability evaluation. Along with the development of item response theory (IRT), information of respondents' score is accurate to item-level in contrast with test-level in classical test theory (CTT). IRT models are good ways to figure out the relationship between the distractor chosen behaviors and respondents' abilities (Levine & Drasgow, 1983; Thissen, 1976).

Generally, there are two kinds of IRT models can fit multiple-choice data. The multiple-choice data, in most instance, are transformed into binary data and fitted dichotomous IRT models for the sake of convenience. The distractor information is totally ignored when dichotomous IRT models (e.g., 2-parameter logistic model, 2PLM) are used. The second kind of IRT models were polytomous IRT models. In consideration of order of options in multiple-choice items, there are mainly two ways of modeling when polytomous models were used. One is transforming unordered categories into ordered ones and fitting ordered polytomous models, for example GPCM (Muraki, 1992). The other is fitting unordered polytomous models like Bock's (1982) nominal response model (NRM) and Thissen and Steinberg' s (1984) multiple-choice model (MCM) which was general model of NRM. The former way requires item options to be ordered. But multiple-choice items are unordered in many cases. The latter way is more flexible, and both NRM and MCM had been used to analysis empirical multiple-choice tests (Sadler, 1998; Thissen, Steinberg, & Fitzpatrick, 1989). However, a limitation of NRM and MCM is the same level of all options to an item, which means that a respondent is possible to choose any option of an item every time he responses. Practically, high-ability respondents may choose the correct options directly without a glance of the distractors, while low-ability respondents may be attracted by several distractors. Therefore distractors in multiple-choice tests show a collapsibility property. As a result, Suh and Bolt (2010) proposed a framework of nested logit models (NLMs) for multiple-choice items. The 2-parameter logistic version of NLMs (2PL-NLM) can be demonstrated as

$$P(u_{ij}=1|\theta_j)=\frac{\exp(\beta_i+\alpha_i\theta_j)}{1+\exp(\beta_i+\alpha_i\theta_j)}, \quad (1).$$

$$P(u_{ij}=0,d_{ijv}=1|\theta_j)=[1-\frac{\exp(\beta_i+\alpha_i\theta_j)}{1+\exp(\beta_i+\alpha_i\theta_j)}][\frac{\exp(\xi_{iv}+\lambda_{iv}\theta_j)}{\sum_{k=1}^{m}\exp(\xi_{iv}+\lambda_{iv}\theta_j)}], \quad (2).$$

where α_i and β_i are the slop parameter and the difficulty parameter. Equation (1), the 2PLM term, defines the probability that a respondent of θ_j chooses the correct option on item *i*. A respondent whose ability (θ) exceeds the item difficulty (β) will have a higher probability of correct response ($P(u_{ij}=1|\theta_j)$). Meanwhile, a respondent whose ability can not reach the item difficulty will have a higher probability of incorrect response ($P(u_{ij}=0|\theta_j)=1-P(u_{ij}=1|\theta_j)$). The second term of the equation (2) is the NRM, nested in the 2PLM, describes a propensity toward each distractor category *v* conditional upon an incorrect response. That means the probability of incorrect response can be further separated into three probabilities in a four-choice item. The distractor "difficulties" (ξ_{iv}) determine which distractor will be probably chosen. Naturally, NLMs present a better approximation to individuals' response behaviors on a multiple-choice item.

2PL-NLM can be easily transformed to 3-parameter logistic nested logit model (3PL-NLM) which treats respondents' guessing behaviors by adding a "guessing" parameter. And also NLMs had been generalized to fit multidimensional multiple-choice tests (Bolt, Wollack, & Suh, 2012). In this study, the unidimensional 2PL-NLM was used.

Overview

Multiple-choice items may have more information than dichotomous items owing to distractors. However, in contrast with dichotomous items, they will also increase cognitive loads of respondents and item-writers, and they require

a complicated model like 2PL-NLM rather than a simple one like 2PLM, for a 2PL-NLM has four more item parameters (λ_1, λ_2, ξ_1, ξ_2) than a 2PLM in a four-choice item. If a multiple-choice items can not enhance the ability evaluation, it is doubtful whether a multiple-choice format, rather than a simple dichotomous format, is necessary in psychological and educational tests.

In this study, 2PL-NLM is used to assess the ability evaluation based on multiple-choice items, since NLMs are more appropriate for multiple-choice items theoretically and conceptually as described above. Therefore, the purpose of present study is to clarify whether or not multiple-choice items, instead of dichotomous ones, should be used to enhance the ability evaluation. More specifically, whether distractor information from multiple-choice items (1) can improve the ability (person parameter) estimation, and (2) can offer some psychological explanations to respondents' distractor chosen behaviors. The rest of this paper is organized as follows. Study 1, a simulation study focuses on the enhancement of ability estimation, when 2PL-NLM is used in multiple-choice tests under different conditions, by contrast with dichotomous item tests fitted 2PLM. Study 2, a real multiple-choice test is used to assess distractor chosen behaviors of respondents at different ability levels.

Method

Simulation Study

The purpose of simulation study is to investigate whether multiple-choice model can extract more distractor information to enhance ability estimation than dichotomous model do. For this purpose simulation study is composed of two parts. Part 1 is to describe whether multiple-choice model should be used or not in a multiple-choice test. And part 2 is to explore the optimal number of distractors should a multiple-choice item have.

Multiple-choice model is compared with a dichotomous model in part 1. 2PLM for dichotomous format and 2PL-NLM for multiple-choice format, rather than 3PLM and 3PL-NLM, are introduced for two reasons. First, eliminating the randomly guessing factor may present more pure enhancement by distractor information. Second, the empirical data used in this study showed a better fitness of 2PLM in a pilot analysis. The accuracy of ability estimation for the 2PLM and 2PL-NLM was evaluated for varying sample size (1000, 2000, and 4000 respondents) and test length (5-, 10-, 20-, 30-, 40-, and 50-item tests) conditions. The condition when the number of respondents is less than 1000 was not included. Because 2PL-NLM has much more item parameters than 2PLM. For example, There are 120 ((8-2)*20) item parameters to estimate in a 20-item test. While 2PLM only have 40 (2*20) item parameters. More item parameters needs larger sample size. Embretson and Reise (2000) recommended over 500 respondents when graded response model which has less item parameters than 2PL-NLM dose. 100 replications were executed for each combination of conditions. For each combination of conditions, Respondents' true ability were generated from θ~Normal (0, 1). On account of order of multiple-choice options, GPCM and NRM, presented ordered multiple-choice items and unordered multiple-choice items respectively, were used to generate responses. Item parameters were generated randomly from the following distributions: slope parameter a~Uniform (0.5, 2) and intercept parameters δ_v~Uniform (-2, 2) for the GPCM, followed by the imposition of constraints $\delta_1 < \delta_2 < \delta_3$ (Muraki, 1992), and slop parameter λ_v~Uniform (-2, 2) and intercept parameter ξ_v~Uniform (-2, 2) for the NRM, followed by the imposition of constraints $\sum_{v=1}^{mi}\lambda_{iv}=0$ and $\sum_{v=1}^{mi}\xi_{iv}=0$ (Suh& Bolt, 2010). Besides, all simulated items in this part included four options, namely one correct option and three distractors.

In part 2 of simulation study, a fully crossed design was implemented under following conditions: 3 (the number of distractors) × 4 (test length). Specifically, test length was examined at four levels: 5 items, 10 items, 20 items and 30 items. The number of distractors was examined at three levels: 3 distractors (4 options), 4 distractors (5 options), and 5 distractors (6 options). The test-length conditions were designed based on some results of part 1. And the distractor conditions were the common settings in a real test. Response data of 2000 person were generated from 2PL-NLM with slop parameter a_i~Uniform (0.5, 2) and difficulty parameter β_i~Uniform (-2, 2). The accuracy of ability estimation was evaluated under each condition.

Data Analysis

To assess the accuracy of ability estimation, mean absolute bias (M_{bias}) and standard deviation of absolute bias (SD_{bias}) were used. These two indices are commonly used to assess bias of estimators in Statistics. They present the mean and the standard deviation of all differences between the estimated values and the true values ($|\hat{\theta}_j - \theta_j|$) (Hofmann, 2007; Walther & Moore, 2005). And they can be demonstrated as

$$M_{bias} = \frac{\sum_{j=1}^{N} |\hat{\theta}_j - \theta_j|}{N}, \quad (3).$$

$$SD_{bias} = \sqrt{\frac{\sum_{j}^{N} (|\hat{\theta}_j - \theta_j| - M_{bias})^2}{N}}, \quad (4).$$

where $\hat{\theta}$ is estimated from model. Small indices indicate that the estimating bias and the estimating variance are small. Both simulation study and empirical study were administrated in R Project version 3.1.2.

Empirical Study

A 32-item Chinese Vocabulary Test for Grade 1 (CVT-G1) data was analyzed. This test is one of a battery of Chinese Vocabulary Tests which contains twelve tests for twelve grades from Grade 1 of primary school to Grade 12 of high school in China. All these tests were constructed based on 2PLM and composed of five-option multiple-choice items (Cao, 1999). 1035 grade 1 students' responses were used to fit three models: 2PLM, NRM and 2PL-NLM in this study. The average of discrimination parameters from 2PLM is 1.050, with a range of 0.554 to 1.720, and the average of difficulty parameters is -0.571, with a range of -1.631 to 0.875.

The advantages of NLMs and an example of distractor analysis procedure will be discussed in this study. First, the performance of NLM on short tests was explored. 32-item CVT-G1 was used to construct another three versions of short tests: 24-item tests, 16-item tests and 8-item tests. Three short tests were constructed by randomly canceling items from the full CVT-G1 test (32 items). The ability parameters obtained from full test were treated as "true values" (θ), because estimating values from long test is supposed to be more accurate than short test theoretically. And the person parameters from short CVT-G1 tests were treated as target estimating values ($\hat{\theta}$). The randomly canceling procedure had repeated 30 times for each length version. Then the average of M_{bias}s and SD_{bias}s were calculated. In second part of this study, probabilities of distractor responses were used to extract the potential psychological meaning.

Results

Enhancement of Ability Estimation

Table 1 presents the biases (M_{bias}s) of ability estimation under each conditions. Estimating biases in GPCM and NRM are treated as baselines respectively. Theoretically, more items a test has, more accurate ability parameters ($\hat{\theta}$s) will be gained. Results from Table 1 show that the enhancement of ability estimation appears under each condition when 2PL-NLM is used. Moreover when the number of items is less than 30, the performance of 2PL-NLM is close to the basic model NRM and GPCM, while 2PLM presents greater bias estimating especially when GPCM is the basic model. In general, 2PL-NLM shows smaller estimating bias than 2PLM when either distractors are ordered or unordered. And 2PLM may lose more information of ability estimation under ordered condition. In addition, the biases rise when the number of items is over 30. That is because more item parameters require more respondents.

Table 1. Mean Bias ofPerson Parameter Estimation.

		Generated based on NRM			Generated Based on GPCM		
		NRM	2PL-NLM	2PLM	GPCM	2PL-NLM	2PLM
1000 respondents	5 items	0.399	0.418	0.478	0.404	0.410	0.734
	10 items	0.320	0.322	0.396	0.312	0.314	0.627
	20 items	0.237	0.247	0.287	0.236	0.246	0.426
	30 items	0.196	0.237	0.257	0.210	0.236	0.371
	40 items	0.180	0.237	0.235	0.205	0.239	0.344
	50 items	0.165	0.244	0.211	0.214	0.245	0.313
2000 respondents	5 items	0.406	0.422	0.634	0.409	0.409	0.734
	10 items	0.316	0.315	0.383	0.309	0.309	0.627
	20 items	0.228	0.247	0.301	0.233	0.241	0.408
	30 items	0.195	0.228	0.256	0.208	0.224	0.376
	40 items	0.175	0.227	0.233	0.200	0.225	0.333
	50 items	0.165	0.238	0.217	0.206	0.240	0.333
4000 respondents	5 items	0.408	0.411	0.500	0.409	0.411	0.601
	10 items	0.313	0.317	0.386	0.307	0.315	0.515
	20 items	0.231	0.246	0.297	0.232	0.241	0.402
	30 items	0.195	0.230	0.254	0.204	0.222	0.367
	40 items	0.173	0.227	0.227	0.199	0.224	0.341
	50 items	0.163	0.239	0.211	0.206	0.239	0.307

The M_{bias} can provide how accurate the estimation is, and the SD_{bias} can provide variation information of estimation. Figure 1 demonstrates the SD_{bias} under various test length conditions when 2000 persons were generated. The other two sample size conditions are not here since they show similar shapes. The decreasing tendency of all SD_{bias} curves show that more item information enhances the stability of estimation. Figure 1 also demonstrates that the distance between NLM curves and baselines (NRM curve, GPCM curve) are much smaller than the distance between 2PLM ones and baselines under 2000-person condition before 30-item condition. In other words, the ability estimation based on NLM is as stable as baseline, more stable than 2PLM. In addition, it is notable that the difference between the NLM and 2PLM decreases. This indicates NLM is losing its ad-

vantages of distractor information gradually, as the number of items increases.

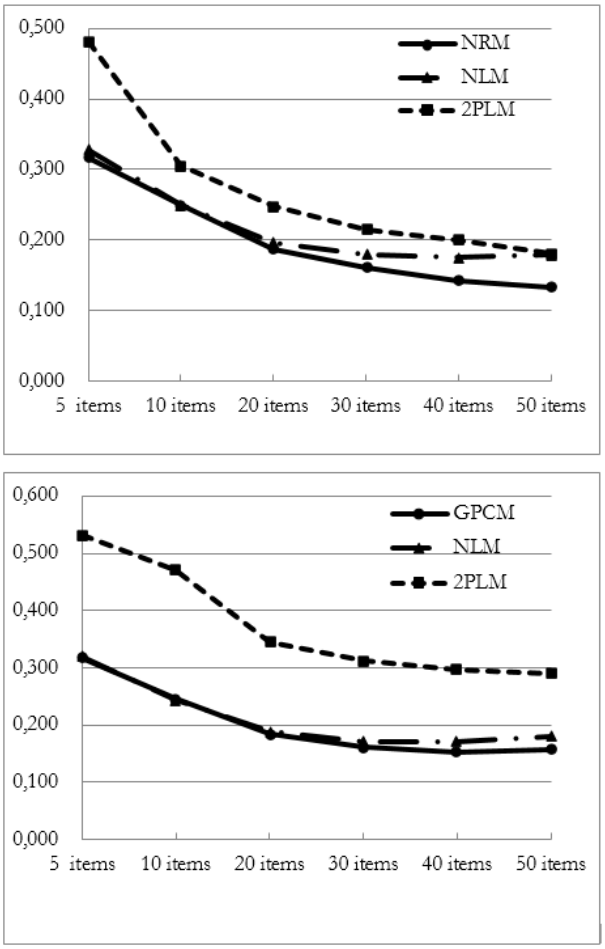

Figure1. Standard Deviation of Absolute Biases Under 2000 Respondents Condition.

Based on the results above the optimal number of distractors under four different test-length conditions had been explored. Three distractor conditions (three distractors, four distractors and five distractors) had been discussed, because too few (e.g., two) or too many (e.g., more than five) distractors are rarely used in reality. The results of estimating biases under different conditions are shown in Table 2. More distractors only promote slight enhancement of ability estimation in short tests. Considering that the challenges of more distractors in five- or six- choice items, four-choice items are accurate enough.

Table2. The Accuracy of Ability Estimation Under Different Distractor Settings.

	5 items		10 items		20 items		30 items	
	M_{bias}	SD_{bias}	M_{bias}	SD_{bias}	M_{bias}	SD_{bias}	M_{bias}	SD_{bias}
4 options (3 distractors)	0.436	0.342	0.334	0.265	0.254	0.201	0.227	0.178
5 options (4 distractors)	0.427	0.337	0.327	0.261	0.249	0.199	0.226	0.177
6 options (5 distractors)	0.423	0.334	0.321	0.258	0.248	0.197	0.226	0.175

Next, $\hat{\theta}$s were divided into five levels to look into the details. Date generated from NRM was used because of its generalizability. Four conditions (5-, 10-, 20-, and 30-item) were taken into consideration for reason of results shown above. The differences of M_{bias} between 2PLM and 2PL-NLM are shown in Table 3. Results show that there are very small differences of estimating bias at intermediate levels, larger differences at high levels, and the largest differences at the low levels. It can be inferred that respondents at low levels chose more distractors which could offer more information by using NLM.

Table3. Differences Between the 2PL-NLM Estimating Biases and 2PLM Estimating Biases.

θ levels	5 items	10 items	20 items	30 items
(-∞,-2)	0.437	0.503	0.430	0.384
(-2,-1)	0.148	0.078	0.041	0.032

(-1,1)	0.041	0.060	0.033	0.002
(1,2)	0.099	0.065	0.053	0.038
(2,+∞)	0.098	0.075	0.099	0.146

Ability Evaluation Based on Empirical Data

The simulation results have shown that distractor information can enhance the accuracy of person parameter estimation by using NLM. In other words, NLM will fit the short tests better than 2PLM. The averages of M_{bias}s and SD_{bias}s, presented the bias between the full-test estimation and short-test estimation, are shown in Table 4. The results illustrate that the M_{bias} based on 2PLM is as good as 2PL-NLM, SD_{bias} even a bit better by using 24-item test. But when shorter tests are used, both two indices based on 2PL-NLM are smaller than 2PLM. These results also prove that distractors can offer more information for ability estimation in short tests.

Table 4. Person Parameter Estimating BiasesBetweenRandomly Constructed Short Tests and Full Test.

	24-item test		16-item test		8-item test	
	M_{bias}	SD_{bias}	M_{bias}	SD_{bias}	M_{bias}	SD_{bias}
2PLM	0.155	0.114	0.264	0.196	0.416	0.305
2PL-NLM	0.156	0.122	0.241	0.186	0.373	0.277

Except for the enhancement of estimation, the probabilities of responses on each distractor obtained from NLM are available. Two steps were carried out to explore the students' response behavior on distractors. Step 1, divide 1035 respondents into five levels by their estimating θ values as follow: level 1 $\theta \in (-\infty,-2)$, level 2 $\theta \in (-2,-1)$, level 3 $\theta \in (-1,-1)$, level 4 $\theta \in (1,2)$, level 5 $\theta \in (2,+\infty)$. Step 2, calculate the mean probabilities of respondents at each level on every distractor. These probabilities can reveal the degree of distractor attractiveness for different levels of respondents. Take item 11 for example (see Table 5). The stem of item 11 is "**赶快** (hurry up)". Respondents were required to choose the best interpretation to this phrase. If a respondent randomly chooses an option, the probability will be 0.2 for a five-choice item. So the probabilities over 0.2 were picked out. Distractor 1 is the most attractive distractor for level 1 respondents with a probability of 0.329, and distractor 3 is the most attractive distractor for level 3 and level 4 respondents with probabilities of 0.263 and 0.220. For level 2 respondents, the probabilities of distractor 1 and distractor 3 are proximate. It reflects that higher vocabulary ability respondents could be attracted by distractor 3 rather than distractor 1. Obviously, these probabilities form a changing trace of distractor responses. And a developmental psychological explanation can be given by analyzing the distractor contents along this changing trace. Distractor 1 shares the same first Chinese character with item stem. Distractor 3 is something about time as well as stem. This contents analysis reveals that respondents at low levels interpreted the item stem in terms of images, and respondents at higher levels began to understand the abstract meanings of words gradually. The procedure of analyzing response probability changing trace is meaningful, but it is impossible to explain all the changing traces item by item in this paper.

Table 5. Probabilities of Distractor Response to Item 11

Distractors (Chinese/English)	Level 1	Level 2	Level 3	Level 4	Level 5
Distractor 1**赶走** (drive away)	**0.329**	**0.192**	0.056	0.006	0.001
Distractor 2**奔跑** (run)	0.117	0.106	0.066	0.028	0.015
Distractor 3**花**时间 (take time)	0.093	**0.185**	**0.263**	**0.220**	**0.156**
Distractor 4说话**很快** (speak quickly)	0.034	0.037	0.030	0.017	0.011

Discussion and Conclusion

Model Selection

Model selection for multiple-choice items has been discussed for a long time. Some conclusions were indeed conflicting (Divgi, 1986; Henning, 1989). In practical assessment, a few researchers used various polytomous models for multiple-choice tests. Much more administrators used dichotomous models for them. However, neither polytomous models nor dichotomous models were proposed upon multiple-choice data. NLMs model the response behaviors in a multiple-choice test theoretically (Suh & Bolt, 2010), so it is worth investigating.

The simulation study of this research showed the conditions under which NLMs should be used rather than simple 2PLM in terms of the enhancement of ability estimation. Obviously, distractor information was effective when short tests were used, especially for the test below 30 items. Increasing the number of item would provide more information for both NLM and 2PLM. But when the number of item exceeded 40, a large number of NLM item parameters might bring in a negative effect. Therefore, it is suggested that a shorter multiple-choice test is acceptable by using NLMs. Yet if the multiple-choice test is too long, over 30 items for example, dichotomous models can offer accurate and stable estimation, and NLMs will not be recommended. With respect to the order of item options, NRM and GPCM were used to generate responses. The results showed that estimating biases of 2PL-NLM under GPCM condition were smaller than the ones under NRM condition. On the contrary, estimating biases of 2PLM under GPCM condition were larger than the ones under NRM condition. So if an ordered

multiple-choice item, in which options may represent the cognitive level of respondents, is used, recoding multiple-choice data into binary data will lose more useful information.

The empirical study results were similar with simulation study results. However, the estimation differences across three versions of tests (24-item test, 16-item test and 8-item test) between two models were smaller in empirical study than they were in simulation study. For one reason, the θs from real full test (32-item test) were not "true". For another reason, the CVT-G1 is an easy test (mean difficulty parameter equal to -0.571), and easy test means fewer distractor chosen behaviors.

It is hard to tell if NLMs could offer more accurate estimation than NRM in this study. However, the NRM was proposed to model the nominal response rather than multiple-choice response. There are no correct option and distractors constructionally, so it is troublesome to explain whether an item is discriminating or not. Yet NLM inherited the advantage of 2PLM in this aspect. Two correlation coefficients of discrimination parameters (slope parameters) of NLM, NRM, and 2PLM were calculated. The correlation between NLM and 2PLM is 0.991, instead 0.647 between NRM and 2PLM. In a word, NLM can be used to guide the item construction and revision as conveniently as 2PLM because of the 2PL-term, and also can provide more distractor information to make better ability estimating.

Distractor Analysis

Distractor information can not only enhance the ability estimation, but also provide some psychological explanation. By analyzing the changing traces of distractor response probabilities together with distractor contents, some meaningful psychological inference could be drawn. A good multiple-choice item with good distractors can indicate that which trait level the respondents are, and it also can reflect some cognitive developmental information and thinking strategies.

Distractors are a kind of wrong options on earth, since high level respondents choose few. This could be concluded from results of Table 3. And it is unnecessary to add more options to a four-option item according to the results in Table 2. Multiple-choice item with three or four distractors is recommended in item writing. Furthermore, when test length over 30, distractor information can help little.

Limitation and Future Research

The conclusions resulted from this study were based on 2PL-NLM. That is, the guessing behaviors were ignored and the dimension of test was unique. In some cases, tests are multidimensional and respondents may use guessing strategies. Consequently, the performance of multiple-choice item on ability evaluation could be different from this study. And also explanation to the distractor chosen behaviors should be much more complex. So the more generalized multiple-choice model (e.g., 3PL-NLM) have to be discussed in those cases.

Less items and more accurate is an ideal aim of psychological assessment. On this point of view, making maximally use of item information to enhancing the estimating accuracy based on multiple-choice items by using NLMs is, to some extent, similar to computerized adaptive testing (CAT). And multiple-choice items are also popular in CAT. However, the conclusion from this study was based on paper-pencil test. So how to applying NLMs to CAT with multiple-choice items still need to be explored.

The changing trace of distractor response is a simple way to explain the response behaviors of respondents. But sometimes it is difficult to directly analyze distractors from a long test. Future research will focus on how to establish a more effective distractor analysis procedure to extract the explanatory information for practical application.

References

Attali, Y., & Fraenkel, T. (2000). The Point-Biserial as a Discrimination Index for Distractors in Multiple-Choice Items: Deficiencies in Usage and an alternative. *Journal of Educational Measurement*, 37(1), 77-86. doi: 10.1111/j.1745-3984.2000.tb01077.x

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika*, *37*, 29-51.doi: 10.1007/BF02291411

Bolt, D. M., Wollack, J. A., &Suh, Y. (2012). Application of a multidimensional nested logit model to multiple-choice test items. *Psychometrika*, *77*, 339-357.doi: 10.1007/S11336-012-9257-5

Briggs, D. C., Alonzo, A. C., Schwab, C., & Wilson, M. (2006). Diagnostic assessment with ordered multiple-choice items. *Educational Assessment*, *11*, 33-63.doi: 10.1207/s15326977ea1101_2

Cao, Y. W. (1999). Construction of vocabulary tests for junior school level. *Acta Psychologica Sinica*, *31*, 460-467.

Davis, F. B., & Fifer, G. (1959). The effect on test reliability and validity of scoring aptitude and achievement tests with weights for every choice. *Educational and Psychological Measurement*, *19*, 159-170.doi: 10.1177/001316445901900202

Divgi, D. R. (1986). Does the Rasch model really work for multiple choice items? Not if you look closely. *Journal of Educational Measurement*, *23*, 283-298.doi: 10.1111/j.1745-3984.1986.tb00251.x

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness Measurement with Polychotomous Item Response Models and Standardized Indices. *British Journal of Mathematical and Statistical Psychology*, *38*, 67-86. doi: 10.1111/j.2044-8317.1985.tb00817.x

Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for Psychologists*. Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Green, B. F., Crone, C. R., & Folk, V. G. (1989). A Method for Studying Differential Distractor Functioning. *Journal of Educational Measurement*, *26*, 147-160.doi: 10.1111/j.1745-3984.1989.tb00325.x

Haladyna, T. M., & Downing, S. M. (1989). A Taxonomy of Multiple-Choice Item-Writing Rules. *Applied Measurement in Education*, *2*, 37-50.doi: 10.1207/s15324818ame0201_3

Haladyna, T. M., & Downing, S. M. (1993). How Many Options is Enough for a Multiple-Choice Testing Item. *Educational and Psychological Measurement*, *53*, 999-1010.doi: 10.1177/0013164493053004013

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A Review of Multiple-Choice Item-Writing Guidelines for Classroom Assessment. *Applied Measurement in Education*, *15*, 309-333.doi:

10.1207/S15324818AME1503_5

Henning, G. (1989). Does the Rasch model really work for multiple-choice items? Take another look: a response to Divgi. *Journal of Educational Measurement*, *26*, 91-97.doi: 10.1111/j.1745-3984.1989.tb00321.x

Hofmann, K. P. (2007). *Psychology of Decision Making in Economics*, Business and Finance. Nova Publishers.

Jacobs, P. I., &Vandeventer, M. (1970). Information in wrong responses. *Psychological Reports*, *26*, 311-315.doi: 10.2466/pr0.1970.26.1.311

Kim, J. (2006). Using the Distractor Categories of Multiple-Choice Items to Improve IRT Linking. *Journal of Educational Measurement*, *43*, 193-213.doi: 10.1111/j.1745-3984.2006.00013.x

Levine, M. V., &Drasgow, F. (1983). The relation between incorrect option choice and estimated ability. *Educational and Psychological Measurement*, *43*, 675-685.doi: 10.1177/001316448304300301

Liu, O. L., Lee, H., & Linn, M. C. (2011). An investigation of explanation multiple-choice items in science assessment. *Educational Assessment*, *16*, 164-184.doi: 10.1080/10627197.2011.611702

Love, T. E. (1997). Distractor selection ratios. *Psychometrika*, *62*, 51-62.doi: 10.1007/BF02294780

Luecht, R. M. (2007). Using information from multiple-choice distractors to enhance cognitive-diagnostic score reporting. In J. P. Leighton &M. J. Gierl (Eds.),*Cognitive diagnostic assessment for education: Theory and practices*(pp. 319–340). Cambridge University Press.doi: 10.1017/CBO9780511611186

Muraki, E. (1992). A Generalized Partial Credit Model：Application of an EM Algorithm. *Applied Psychological Measurement*, *16*, 159-176.doi: 10.1002/j.2333-8504.1992.tb01436.x

Penfield, R. D. (2011). How are the Form and Magnitude of DIF Effects in Multiple-Choice Items Determined by Distractor-Level Invariance Effects？. *Educational And Psychological Measurement*, *71*, 54-67.doi: 10.1177/0013164410387340

Roediger III, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1155.doi: 10.1037/0278-7393.31.5.1155

Sadler, P. M. (1998). Psychometric models of student conceptions in science: Reconciling qualitative studies and distractor-driven assessment instruments. *Journal of Research in science Teaching*, *35*, 265-296.doi: 10.1002/(SICI)1098-2736(199803)35:3<265::AID-TEA3>3.0.CO;2-P

Sigel, I. E. (1963). How intelligence tests limit understanding of intelligence. *Merrill-Paker Quarterly,9*, 39-56.

Suh, Y., & Bolt, D. M. (2010). Nested logit models for multiple-choice item response data. *Psychometrika*, *75*, 454-473.doi: 10.1007/s11336-010-9163-7

Suh, Y., & Bolt, D. M. (2011). A Nested Logit Approach for Investigating Distractors as Cause of Different Item Functioning. *Journal of Educational Measurement*, *48*, 188-205.doi: 10.1111/j.1745-3984.2011.00139.x

Suh, Y., & Talley, A. E. (2015). An Empirical Comparison of DDF Detection Methods for Understanding the Causes of DIF in Multiple-Choice Items. *Applied Measurement in Education*, *28*, 48-67.doi: 10.1080/08957347.2014.973560

Tamir, P. (1971). An alternative approach to the construction of multiple choice test items. *Journal of Biological Education*, *5*, 305-307.doi: 10.1080/00219266.1971.9653728

Tamir, P. (1989). Some issues related to the use of justifications to multiple-choice answers. *Journal of Biological Education*, *23*, 285-292.doi: 10.1080/00219266.1989.9655083

Thissen, D. M. (1976). Information in wrong responses to the Raven Progressive Matrices. *Journal of Educational Measurement*, *13*, 201-214.doi: 10.1111/j.1745-3984.1976.tb00011.x

Thissen, D., & Steinberg, L. (1984). A Response Model for Multiple Choice Items. *Psychometrika*, *49*, 501-519.doi: 10.1007/BF02302588

Thissen, D., Steinberg, L., & Fitzpatrick, A. R. (1989). Multiple-Choice Models: The Distractors Are Also Part of the Item. *Journal of Educational Measurement*, *26*, 161-176.doi: 10.1111/j.1745-3984.1989.tb00326.x

Walther B. A., & Moore J. L. (2005). The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. *Ecography*, *28*, 815-829. doi: 10.1111/j.2005.0906-7590.04112.x

Wollack, J. A. (1997). A Nominal Response Model Approach for Detecting Answer Copying. *Applied Psychological Measurement*, *21*, 307-320.doi: 10.1177/01466216970214002

(Article received: 02-10-2015; revised: 09-03-2016; accepted: 16-05-2016)