Investigating differential options functioning using multinomial logistic regression

Minjeong Park and Amery D. Wu*

Department of ECPS,
University of British Columbia Scarfe Building,
2125 Main Mall, Vancouver, BC V6T 1Z4, Canada
Email: minjeong.park@alumni.ubc.ca
Email: amery.wu@ubc.ca
*Corresponding author

Abstract: This paper focuses on an investigation of differential functioning in all options of multiple-choice items, referred to as differential options functioning (DOF). DOF is defined as an investigation into whether respondents from different groups (i.e. subpopulations), with equal levels of the attribute being measured (i.e. skill, ability, etc.) would have different probabilities to select the options. This paper aims to reconceptualise previous approaches to DOF in terms of its terminologies, purposes and uses. This paper further proposes a set of simple and integrated procedures for investigating DOF based on the well-known theory of multinomial logistic regression. A real data demonstration is provided to guide the application of the proposed method. The demonstration compares two test-language groups (English vs. French) on the four options of 15 multiple-choice items in booklet 13 of the 2011 Progress in International Reading Literacy Study.

Keywords: comprehensive differential item functioning; differential alternative analysis; differential distractor analysis; differential item functioning; differential options functioning; item responding; measurement bias; measurement invariance; multinomial logistic regression; multiple-choice question; option characteristic curves.

Reference to this paper should be made as follows: Park, M. and Wu, A.D. (2017) 'Investigating differential options functioning using multinomial logistic regression', *Int. J. Quantitative Research in Education*, Vol. 4, Nos. 1/2, pp.94–119.

Biographical notes: Minjeong Park is a graduate student in the Program of Measurement Evaluation and Research Methodology at the University of British Columbia.

Dr Amery Wu is a Faculty member at the Department of Educational and Counselling Psychology, and Special Education at the University of British Columbia. Her area of research interest is psychometrics and quantitative research methodology.

This paper is a revised and expanded version of a paper entitled 'Detecting Differential Option Functioning Using Multinomial Logistic Regression in SPSS', presented at *April 2017 at the annual meeting of the American Educational Research Association Division D, Measurement and Research Methodology*, San Antonio, TX.

1 Introduction

Multiple-choice items require an individual to respond by selecting from a pre-specified set of options. This item format is widely used in achievement/aptitude tests where examinees are required to choose a correct answer. It is also common in questionnaires/surveys where respondents are asked to choose from a set of phrases/statements to elicit their preference, inclination or valuing. In the latter context, the response options can be either ordered (e.g. a Likert-type rating scale) and result in ordinal response data or unordered (e.g. forced-choice or ipsative format) and result in nominal data. With its wide popularity and utility, multiple-choice items are often subject to an examination of lack of measurement invariance (a potential measurement bias) due to differential responding (e.g. how distinct groups understand the content or format of the items differently).

This paper advocates and facilitates the investigation of differential functioning of *all* response options, hereafter, referred to as *differential options functioning* (DOF). We define DOF as an investigation of whether respondents from different groups, with equal levels of the attribute being measured, would have different probabilities to select the options. The concept of DOF has been introduced using different terminologies such as differential distractor functioning (DDF), differential alternative functioning and comprehensive differential item functioning (CDIF) in the literature. Although the basic concept is very similar, we use the new term DOF to espouse the versatility of studying options in various research contexts (we will explain our reasons for using this term in more detail later).

Upmost, we need to point out that DOF is different from the concept of differential item functioning (DIF). Statistically, DIF investigates group difference in the probabilities of selecting one keyed option based on *dichotomised* binary data, conditioning on the attribute being measured, whereas DOF investigates group difference in the probabilities of selecting all options without dichotomising the data. When studying DIF in achievement/aptitude test items, the correct answer is often considered as a keyed option and coded as 1 and the rest of the incorrect answers are collapsed and coded as 0. Correspondingly, DIF only focuses on the keyed option and does not look into each option individually. In contrast, DOF investigates all options by looking at the unequal conditional probabilities in selecting each of the options chosen by the respondents. This feature of DOF helps to understand individuals' original selection of the options, which is not doable with a DIF study.

Specifically, this paper aims to fulfil the following three purposes in order to advocate and facilitate DOF studies:

1 Reconceptualise the previous approach to DOF in terms of its terminologies, purposes and uses.

2 Propose a simple and integrated analytical method for studying DOF based on the known statistical theory of multinomial logistic regression.

3 Through demonstration with a real data example, provide a proof of concept and a guide for application of the proposed DOF method.

2 Literature review

Compared to DIF, there are much fewer methodological and applied works pertaining to how groups of individuals with equal ability/attribute being measured, may select different response options. Our literature review found 19 related works that appeared intermittently from 1980 to 2015 (Abedi et al., 2008; Banks, 2006, 2009; Barton and Huynh, 2003; Bolt et al., 2001; Dorans et al., 1992; Green et al., 1989; Kato et al., 2009; Marshall, 1983; Middleton and Laitusis, 2007; Penfield, 2008, 2010; Schmitt and Bleistein, 1987; Schmitt and Dorans, 1990; Suh and Bolt, 2011; Suh and Talley, 2015; Thissen et al., 1993; Veale and Foreman, 1983; Westers and Kelderman, 1991). Among these works, the concept of DOF was introduced with various terminologies and a range of methods have been proposed for DOF analysis.

In an early study, Veale and Foreman (1983) introduced the concept of analysing examinees' incorrect responses for assessing cultural bias. Their approach simply compared the observed proportions of examinees who selected each of the incorrect options between cultural groups. As a more applied approach, log-linear models have been employed for studying distractors in various contexts such as gender, disabilities and culture (Banks, 2006, 2009; Barton and Huynh, 2003; Green et al., 1989; Marshall, 1983). In particular, Green et al. (1989) defined DDF as analysing the incorrect item responses (i.e. distractors) and investigated DDF based on the log-linear approach. Their approach involved a three-way contingency table stratified by the ability levels.

A group of researchers (Dorans et al., 1992; Middleton and Laitusis, 2007; Schmitt and Bleistein, 1987; Schmitt and Dorans, 1990) adopted a descriptive standardisation approach, which assessed differential functioning by computing weighted differences between the groups. Dorans et al. (1992) introduced CDIF for evaluating all response options including the keyed, omitted and not reached.

Later on, distractors were also studied under the framework of item response theory (IRT). Thissen et al. (1993) discussed differential functioning at the option level and referred to it as *differential alternative functioning* by testing the group difference in the response curves using a likelihood ratio test. Similarly, Westers and Kelderman (1991) studied differential functioning in distractors by comparing the fit of different latent class analysis (LCA) models. As an extension, Bolt et al. (2001) studied distractors using a mixture IRT of nominal response model (Bock, 1972) and Suh and Bolt (2011) introduced an application of a nested logit IRT model for studying distractors. Penfield (2008, 2010) proposed a new approach of odds ratio based on a revised nominal response model.

Methods based on logistic regression began to appear almost at the same time as IRT-based methods. Abedi et al. (2008) employed multi-step binary logistic regression (similar to Swarminathan and Rogers, 1990) for studying differential functioning in the incorrect responses. However, they only analysed the most common distractor on each item. Extending the approach of Abedi et al. (2008), Kato et al. (2009) applied multinomial logistic regression to analyse differential functioning in all options.

A recent work by Suh and Talley (2015) compared the three approaches for detecting DDF (log-linear approach, IRT-based approaches and an odds ratio approach). In the Appendix, we summarise all the different methods to date.

Our literature review also looked at the purposes of DOF studies. Nine out of 19 works discussed DOF in the context of DIF (Banks, 2009; Dorans et al., 1992; Penfield, 2010; Schmitt and Bleistein, 1987; Schmitt and Dorans, 1990; Suh and Bolt,

2011; Suh and Talley, 2015; Thissen et al., 1993; Westers and Kelderman, 1991). These researchers focused their DOF investigation on identifying causes of DIF. With this focus, DOF was seen as a secondary analysis after an item is detected as DIF. The other 10 studies did not discuss DOF in the context of DIF. Five of them treated the option as the main character that was subject to an investigation of measurement invariance (Abedi et al., 2008; Banks, 2006; Barton and Huynh, 2003; Green et al., 1989; Kato et al., 2009). These works focused their purpose of DOF on examining whether options function in the same manner for all groups. The remaining five studies mentioned other uses of DOF (Bolt et al., 2001; Marshall, 1983; Middleton and Laitusis, 2007; Penfield, 2008; Veale and Foreman, 1983). Their purposes included, for instance, recognising distinct groups' perceptions of items, understanding how stimuli attracts or repels distinct groups, identifying cognitive (mis)steps and reviewing items.

Although previous research has introduced several methods for studying DOF, there were some challenges in the accessibility and applicability of these methods. Most of the existing methods entail a mix of different statistical techniques, complicated computations and/or knowledge in new statistical software. Straightforward and integrated methods are not readily available to applied researchers. This may be one of the reasons why the study of DOF has not been well received in practice. Furthermore, the application of DOF was, by and large, limited to post hoc measurement invariance investigations in previous research. DOF has often been conducted as a method for identifying causes of DIF, a secondary analysis, or to examine measurement invariance at the option level once item level DIF is detected. Moreover, these applications of DOF were typically limited to achievement/aptitude tests. Although some studies discussed other possible uses of studying DOF, the potential uses of DOF were briefly mentioned or vaguely alluded to. As a result, many potential uses of DOF are not fully identified and may be neglected. Hence, difficulties with methods and lack of awareness of uses may keep researchers from studying DOF. To address these issues, this paper aims not only to reconceptualise the investigation of DOF but also to propose a new and accessible analytical method.

3 Our conceptualisation of DOF

In this paper, we attest that DOF should not simply be considered as a post hoc subsidiary study for identifying causes of DIF. Rather, DOF should be considered as a stand-alone, self-sufficient investigation which can reveal in-depth information and have versatile uses that a DIF study cannot accomplish.

3.1 The term differential options functioning (DOF)

Our first effort in promoting DOF as an independent study from DIF lends to a suggestion of using the term DOF rather than those already existing in the literature. We propose the term DOF for the following reasons based on our reflection on the literature review. First, the term CDIF used by Dorans et al. (1992) characterises the utility of studying options as a secondary analysis for identifying causes of DIF. This term obscures other possible uses for studying options. Second, the term DDF used by Green et al. (1989) is only suitable for achievement/aptitude types of measures where a specific option is considered as the correct answer and the others as 'distractors'. With items in questionnaires/surveys where the respondents are asked to choose a statement of their top

choice, no particular option should be considered as the correct answer or the distractor. In addition, the term differential 'distractor' functioning implies that the correct option is not susceptible to differential functioning; this clearly is not the case. The correct option should also be subject to investigation. Moreover, the term *differential alternative functioning* used by Thissen et al. (1993) implies that the provided response categories are 'alternatives' to one another. However, this term is often not applicable when the categories are given as mutually competing or contrasting statements for eliciting a respondent's preference, inclination or valuing in a survey/questionnaire. In contrast, our suggested term, DOF is more generic, neutral and flexible for wide types of response formats designed for different measurement contexts (measurement or assessment for different attributes, purposes and uses).

3.2 Potential uses of DOF

In previous research, DOF has been often conducted to identify causes of DIF or to examine measurement invariance in achievement/aptitude tests. However, there are many other advantages to studying individuals' original response options in various measurement contexts. In particular, this paper discusses its potential uses in the two contexts: achievement/aptitude tests and questionnaires/surveys.

In achievement/aptitude tests, one option is selected as the 'correct' answer by the test developer and the remaining options are often written to represent possible misunderstanding, lack of knowledge or missteps to reach the correct answer. In this context, DOF can be applied to understand how distinct groups make mistakes differently in answering the item while accounting for the ability being measured. This kind of application is useful to identify teaching and learning gaps between groups in understanding the knowledge or skill an item intends to assess. Say, for example, that male and female students are equally capable in math, but female students have higher probability in choosing an option written to test a common mistake. This may signal that there is some teaching or learning gaps between male and female students. By looking at the probability of choosing each of the incorrect options (compared to the correct option specified as the keyed option), one can see where each group is more likely or less likely to make mistakes. If a group has a higher probability to choose a certain incorrect option compared to the other groups (conditioning on ability), it indicates that the group is more likely to make a mistake in that option for some reason. Also, DOF can provide information about how test-takers with certain ability level (e.g. low or high ability) respond to each option by examining the probabilities of selecting options against their ability level (a situation of nonuniform DOF, which we will explain later). This can be presented by visualising the DOF curves.

In questionnaires/surveys research where none of the options are considered as the 'correct' answer, options often represent different levels of endorsement (e.g. a rating scale; *Strongly disagree, Disagree, Agree, Strongly agree*) or different choices (e.g. ipsative or forced-choice item format; 'Vanilla', 'Strawberry', 'Chocolate'). In this type of measure, DOF is useful for investigating different response patterns between groups by looking at how groups prefer or avoid options controlling for respondents' preference or attitude being elicited by the options. In this context, the keyed (reference) option is arbitrary and can be specified by the researchers depending on their specific research focus. When no particular option should be treated as the reference option, one can choose the option that is most frequently chosen by respondents.

DOF is particularly useful when different response patterns are suspected between groups. For instance, in an item asking about perception of body weights, it might be suspected that females are more likely to answer that they are overweight compared to males. This may happen in certain cultures where females are more self-conscious about their weight due to social expectation of body image. This response pattern can be investigated using DOF by looking into whether and how males and females, controlling for their body mass index (BMI) may tend to choose the options representing overweight (or non-overweight). Different response patterns between groups of individuals can also occur when an item has different *types* of choices such as personality types. In such cases, certain groups may prefer or avoid some choices due to social or cultural norms. In the same way, the response patterns of groups can be investigated by looking into how the groups are more likely to choose each option when they have the same (scores for) types of personality.

Furthermore, DOF can help understand response tendencies (e.g. response style and response set) such as avoidance of the extreme options and preference to the neutral option. For instance, one can investigate how different ethnic groups respond to the extreme options by comparing the probabilities of choosing the extreme options between the groups.

3.3 Relationship between DIF and DOF

The relationship between DIF and DOF was not clearly discussed in the literature. Rather, it is commonly assumed that DOF occurs when DIF is found. That is, DOF and DIF are believed to be equivalent phenomenon that provides the same information. However, this may not necessarily be true. In the following, we present four logical statements to explain four possible population relationships between DOF and DIF. Note that these statements are presented assuming we know the population conditional probabilities of choosing the options (holding constant the attribute being measured). Also, we assume that there are four options for an item and option A is the keyed option with a population probability of choosing option $A = p$.

Statement 1: If DIF occurs, then DOF occurs. This statement is true. When DIF occurs, the probabilities of choosing option $A(p)$ are different between the groups, see Table 1, for example where $p_1 = 0.25 \neq p_2 = 0.40$, hence the aggregate probabilities of choosing the non-keyed options $(1 - p)$ will be different between the two groups, $(1 - p)_1 = 0.75 \neq (1 - p)_2 = 0.60$. It is impossible that two groups have exactly same probabilities in all of the other options. This is the logic underlying the belief that DOF is a secondary analysis for causes of DIF.

Table 1 Example of population relationship for Statement 1

Probability of choosing an option conditioning on ability (trait)			
Options	*Group 1*	*Group 2*	*Population DIF/DOF status*
A (key)	0.25	0.40	DIF (favouring group 2)
B	0.25	0.25	No DOF
C	0.30	0.30	No DOF
D	0.20	0.15	DOF (favouring group 1)

Statement 2: If DOF occurs, then DIF occurs. This statement is false. A counter-example to this statement is presented in Table 2. When DOF occurs, the probabilities of selecting the non-keyed options between two groups are different. However, DIF may not occur because the probabilities for the keyed option can still be the same between two groups (see $p_1 = p_2 = 0.25$ for an example in Table 2).

Table 2 Example of population relationship for Statement 2

Probability of choosing an option conditioning on ability (trait)			
Options	*Group 1*	*Group 2*	*Population DIF/DOF status*
A (key)	0.25	0.25	No DIF
B	0.30	0.25	DOF (favouring group 1)
C	0.25	0.40	DOF (favouring group 2)
D	0.20	0.10	DOF (favouring group 1)

Statement 3: If DOF does not occur, then DIF does not occur. This statement is true. If none of the non-keyed options B, C and D show DOF, their probabilities are the same across the two groups. The aggregate probabilities for the non-keyed options $(1 - p)$ will be the same between the two groups, see $(1 - p)_1 = (1 - p)_2 = 0.55$ in Table 3, for example. Therefore, the probabilities for the keyed option (p) will be the same between the two groups as well, $p_1 = p_2 = 0.45$, indicating no DIF.

Table 3 Example of population relationship for Statement 3

Probability of choosing an option conditioning on ability (trait)			
Options	*Group 1*	*Group 2*	*Population DIF/DOF status*
A (key)	0.45	0.45	No DIF
B	0.20	0.20	No DOF
C	0.10	0.10	No DOF
D	0.25	0.25	No DOF

Statement 4: If DIF does not occur, then DOF does not occur. This statement is false. A counter-example to this statement is presented below. When DIF does not occur, it only indicates that the probabilities of choosing the keyed option (p) are the same between the two groups, see $p_1 = p_2 = 0.25$ in Table 4 for example. It does not require the probabilities for the non-keyed options $(1 - p)$ to distribute in the same way between the two groups. DOF can occur by showing unequal probabilities between two groups in selecting the non-keyed options.

Table 4 Example of population relationship for Statement 4

Probability of choosing an option conditioning on ability (trait)			
Options	*Group 1*	*Group 2*	*Population DIF/DOF status*
A (key)	0.25	0.25	No DIF
B	0.30	0.50	DOF (favouring group 2)
C	0.25	0.10	DOF (favouring group 1)
D	0.20	0.15	DOF (favouring group 1)

From the above explanations for the population relationship between DIF and DOF, we can draw the following conclusion. First, the existence of DIF indicates the existence of DOF (Statement 1). In this case, a DIF study can be an overall evaluation of the existence of DOF. However, a DIF study cannot detect which option has DOF. Second, the existence of DOF is not a sign of DIF (Statement 2). A DOF study does not provide sufficient evidence for the existence of DIF. Third, the absence of DOF indicates the absence of DIF (Statement 3). If an item does not show DOF, then the item is non-DIF, either. That is, a DOF study can be used to assess the absence of DIF. Lastly, the absence of DIF does not mean that there will be no DOF (Statement 4). Hence, a DIF study does not guarantee the absence of DOF.

4 The proposed analytical method

Despite its many potential uses, the idea of conducting DOF investigation was not well received among applied researchers. In addition to the lack of awareness of the potential uses of DOF, we believe that this is partly due to a lack of an integrated and straightforward method that can be conducted using popular statistical software packages such as *SAS*, *SPSS* or *R*. In fact, various statistical techniques proposed for studying DOF are not easily accessible because these methods involve a mix of different techniques, additional complicated computations and/or learning new statistical software, all of which can be a barrier to the interested researchers who are not measurement professionals or psychometricians.

To facilitate the study of DOF, this paper proposes a set of integrated and straightforward procedures for studying DOF. The procedures are suggested based on the known theories of multinomial logistic regression (Agresti, 2012; Hosmer et al., 2013; Menard, 2002) as well as existing methods for DIF based on logistic regression (Swaminathan and Rogers, 1990; Zumbo, 1999). The proposed method is very accessible and does not involve a steep learning curve. All procedures can be easily conducted in popular statistical software packages and the results can be directly obtained from the basic outputs of these packages.

Note that this paper is not the first to apply multinomial logistic regression in the study of DOF. Kato et al. (2009) applied multinomial logistic regression and studied DOF based on a likelihood ratio test with the effect sizes of pseudo R^2 differences at the item level and MADs between the groups' response characteristic curves at the option level. However, the proposed method in this paper is based on different procedures and rules. Later in Section 6, we have introduced and demonstrated our proposed method, we will discuss more about how our method differs from their works and what new contributions our method brings to the literature. Furthermore, the proposed procedures take on the odds ratio as the effect size measure as several DOF methods suggested (Abedi et al., 2008; Banks, 2006, 2009; Penfield, 2008, 2010).

The next section provides a real data demonstration that serves as a proof of concept and guide for application of our proposed method. The step-by step procedures will be explained in the demonstration. The data were retrieved from the 2011 Progress in International Reading Literacy Study (PIRLS) of Canada. All analyses were conducted using maximum likelihood estimation in SPSS.

5 Demonstration of the proposed method

5.1 Participants and measure

Progress in International Reading Literacy Study 2011 international database provides students' original responses to the items assessing students' reading ability. Booklet 13 was used for this demonstration. The dataset contains students' actual choices from the four options of 15 multiple-choice items (out of a total of 30 items). One of the four options in each item was keyed as the correct answer by PIRLS experts and test developers. The sample includes 4,805 fourth grade students in Canada (50.7% males and 49.3% females). The students took either the English version (72.4%, coded as 0) or the French version (27.6%, coded as 1) of the reading assessment. The two test-language groups were treated as the grouping variable in the present DOF demonstration.

5.2 The proposed analytical procedures

The set of four response options of each item was analysed with three multinomial logistic regression models given as,

$$\text{Model 1}: \ log\frac{P(Y=j|T)}{P(Y=k|T)} = a_j + b_1 T \tag{1}$$

$$\text{Model 2}: log\frac{P(Y=j|T,G)}{P(Y=k|T,G)} = a_j + b_1 T + b_2 G \tag{2}$$

$$\text{Model 3}: \ log\frac{P(Y=j|T,G)}{P(Y=k|T,G)} = a_j + b_1 T + b_2 G + b_3 (T*G) \tag{3}$$

where $j = 1 \ldots J$ denotes the categories of the available options, k denotes the reference (key) category (in this demo, the correct option), T is the rest total score (the sum of the item scores of the entire test excluding the score of the item being studied for DOF), G is the grouping variable; in this demo, the English-language is the reference group and the French-language is the focal group), $T*G$ is the product of the rest total score and the grouping variable, indicating the interaction between the two variables.

On the left-hand side of Eqs. (1)–(3), the logits (hence the probabilities) of selecting each non-keyed option j (vs. selecting the keyed option k) are contrasted and modelled as linear models of the predictors on the right-hand side of the equations. That is, for each item with J options ($J = 4$ in this demo), there are $J - 1$ paired contrasts being modelled simultaneously. Simply put, a multinomial logistic regression for J response categories can be seen as a set of simultaneous $J - 1$ binary logistic regressions.

Model 1 in Eq. (1) can be considered as the baseline model which includes only the variable T, the rest total scores. The rest total scores were standardised and served as a proxy of students' true level of reading ability and controlled for examining DOF. Model 2 adds the grouping variable G to examine *uniform DOF*. Uniform DOF is a scenario where the groups' influence on the option selection is modelled as a *constant shift* (a decrease or increase) in the logit. This constant shift is indicated by the estimate of b_2 for G of Model 2. Model 3 includes an additional interaction term of T*G and

examines the existence of *non-uniform DOF*. *Non-uniform DOF* is a scenario where the group's difference in the logit (i.e. b_2) is further influenced by students' ability *T*. The presence of the moderated group influence is indicated by the estimate of b_3 for *T*G* of Model 3. To further explain the non-uniform DOF, Model 3 in Eq. (3) is rearranged into Eq. (3a). In (3a), the overall influence of the grouping variable *G* is expressed as a constant shift of b_2 plus the moderating function of b_3T that characterises the non-uniform DOF:

$$\text{Model 3a}: log\frac{P(Y=j|T,G)}{P(Y=k|T,G)} = \alpha_j + b_1T + (b_2 + b_3T)G \tag{3a}$$

5.3 Stage-wise approach

The statistical inferences of DOF are determined in two stages based on the results of Models 1–3. Stage 1 detects the presence of DOF *simultaneously* for the $J-1$ non-keyed options (vs. the keyed option) at the *item level*. This step is to inspect whether at least one option shows DOF, either uniform or non-uniform. Stage 2 detects DOF at the *option level* for each of the individual $J-1$ options to inspect which and how option(s) functions differently for the groups, if item-level DOF is detected in stage 1.

5.3.1 Stage 1: Simultaneous DOF detection at the item level

At stage one, the presence of uniform and non-uniform DOF is tested simultaneously for the $J-1$ contrasts of the non-keyed options (vs. the keyed option) by two likelihood ratio tests (LRT). Each LRT tests the −2 log likelihood difference ($-2LL\Delta$) between two nested models. Both $-2LL\Delta$ values are tested against χ^2 distributions with degrees of freedom = $df\Delta$ (difference in degrees of freedom between the two models). The first LRT detects the non-uniform DOF by testing the $-2LL\Delta$ between Model 2 and Model 3. The second LRT detects the uniform DOF by testing the $-2LL\Delta$ between Model 1 and Model 2. Together, the results of the two LRTs will inform one of the three exclusive conclusions about an item: (a) at least one non-uniform DOF (denoted as non-uniform^{1or+}) if the first LRT is significant, (b) at least one uniform DOF (denoted as uniform^{1or+}) if the second LRT is significant, but not the first, and (c) no DOF if neither of the two LRTs is significant.

5.3.2 Stage 2: Individual DOF detection at the option level

The presence of DOF detected at the item level in Stage 1 only determines whether *at least* one of the options is detected as DOF (non-uniform^{1or+} or uniform^{1or+}). Stage 2 examines DOF at the option level to find out which and how each of the individual $J-1$ options functions differently for the groups. Note that *only* items that are found to be non-uniform^{1or+} or uniform^{1or+} in Stage 1 will be subject to examination of option-level DOF in Stage 2. To understand the Stage 2 procedures, recall that DOF is examined for $J-1$ pairs of logits, contrasting the $J-1$ non-keyed options to the keyed option *k* (3 pairs in the demo). Each logit is expressed as a linear model of the predictors in Eq. (1)–(3). The predictors' regression coefficients for each of the $J-1$ paired contrasts will be examined for statistical inference (except for the total score that is regarded as a

controlled variable). Moreover, the option-level DOF procedures are further broken down by whether an item is detected as non-uniform^{1or+} or uniform^{1or+} in Stage 1.

5.3.2.1 Items detected as non-uniform^{1or+} DOF in Stage 1: Reporting Model 3

If an item is detected as non-uniform^{1or+} in Stage 1, it signals that *at least* one of the $J-1$ options functions *non-uniformly*. The *Model 3* regression coefficients will be tested for statistical significance by the *Wald* χ^2 test. The two *Wald* χ^2 tests for b_2 of *G* and b_3 of *T*G,* predicting each $J-1$ contrast in Model 3, will reach one of the following *three exclusive* conclusions. (1) As long as the regression coefficient b_3 of *T*G* is statistically significant, it is concluded that the *j*th option functions non-uniformly for the groups. (2) If the regression coefficient b_3 of *T*G* is statistically non-significant but the regression coefficient b_2 of *G* is statistically significant, it is concluded that the *j*th option functions uniformly for the groups. (3) If neither the coefficient b_2 nor b_3 is statistically significant, it is concluded that the *j*th option does not function differentially for the groups. Note that testing and reporting the statistical significance of the *Model 2* regression coefficients are not needed because they are irrelevant when an item is detected as non-uniform^{1or+} in Stage 1.

5.3.2.2 Items detected as uniform^{1or+} DOF in Stage 1: Reporting Model 2

If an item is detected as uniform^{1or+} in Stage 1, it signals that *at least* one of the $J-1$ options functions *uniformly*. The *Model 2* regression coefficients will be tested for statistical significance by the *Wald* χ^2 test. The *Wald* χ^2 test for b_2 of *G,* predicting each $J-1$ contrast in Model 2, will reach one of *two* conclusions. (1) If b_2 is statistically significant, it is concluded that the *j*th option functions uniformly for the groups. (2) If b_2 is statistically non-significant, it is concluded that the *j*th option does not function differentially for the groups. Note that testing and reporting the statistical significance of the *Model 3* regression coefficients are not needed because they are irrelevant when an item is detected as uniform^{1or+} in Stage 1.

5.3.2.3 Effect size and option characteristic curves

For all Stage 2 detection of DOF, the odds ratio (OR) will aid the interpretation for both non-uniform and uniform DOF. The odds ratio of selecting the *j*th option (vs. the keyed option) comparing the focal group to the reference group are reported as a measure of size for DOF. A population odds ratio equal to one indicates DOF is absent. For uniform DOF, the odds ratio can be simply interpreted as a constant magnitude of DOF because the direction and the size of group difference are constant in the logit form for the entire continuum of the total score. Nonetheless, for non-uniform DOF, the odds ratio should be interpreted at each level of total score because the direction and/or the magnitude of DOF can vary across the continuum of the total score. The interpretation of odds ratio will be illustrated in Section 5.

We also tentatively take Ferguson's (2009) odds ratio 2, 3 and 4 as the cut-offs for small, medium and large DOF. Please note that Ferguson's (2009) cut-offs were suggested for logistic regression in general. We only suggest considering them as a rough

guide for interpreting the size of DOF. Researchers must consider multiple factors including the groups being compared, the attribute being measured, and the purpose of studying DOF when determining the cut-off for interpreting their own study results.

The *option characteristic curves (OCCs)* will help to visualise whether and how an option functions differentially between groups (see Figure 1). That is, these curves show whether each option shows no DOF, uniform DOF, or non-uniform DOF. Furthermore, the OCCs can help to visualise the magnitude of uniform DOF (the area between curves of the groups). They can also help to examine the pattern of non-uniform DOF; the moderating effect of ability (total) on DOF.

Figure 1 Option Characteristic Curves (OCCs) showing which and how the non-keyed options (vs. the keyed option) function differentially

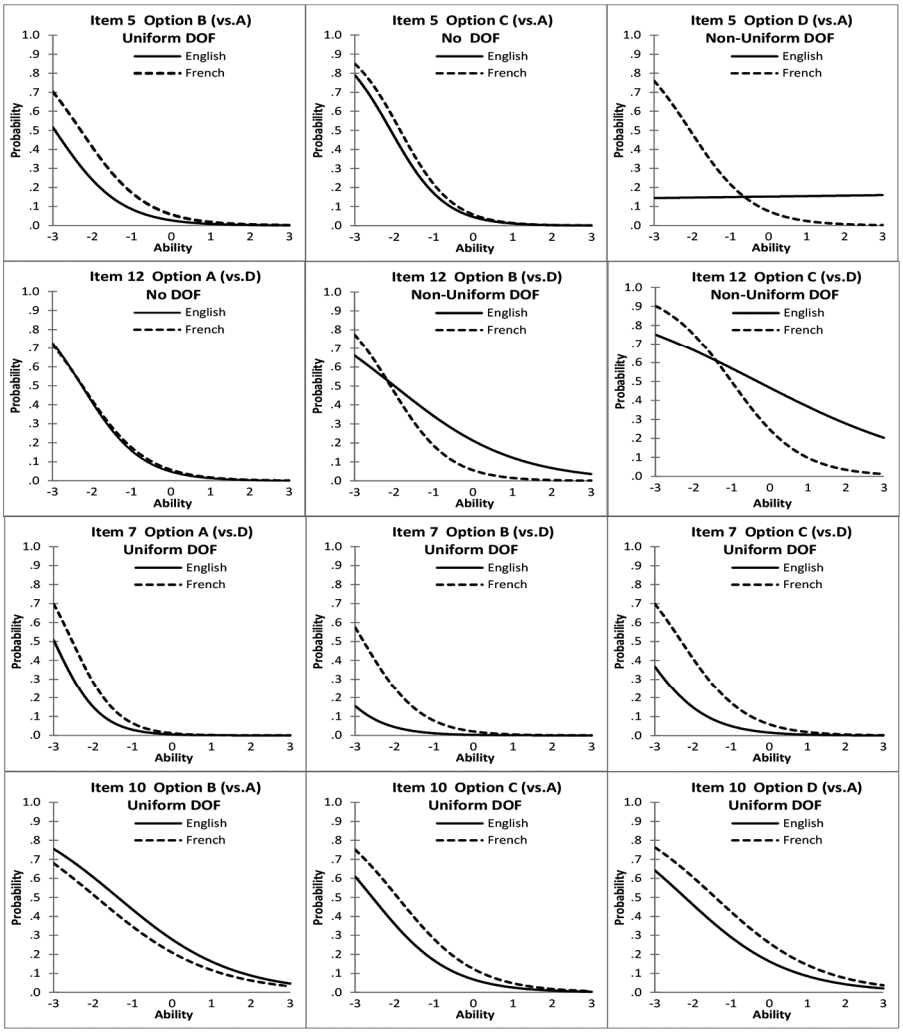
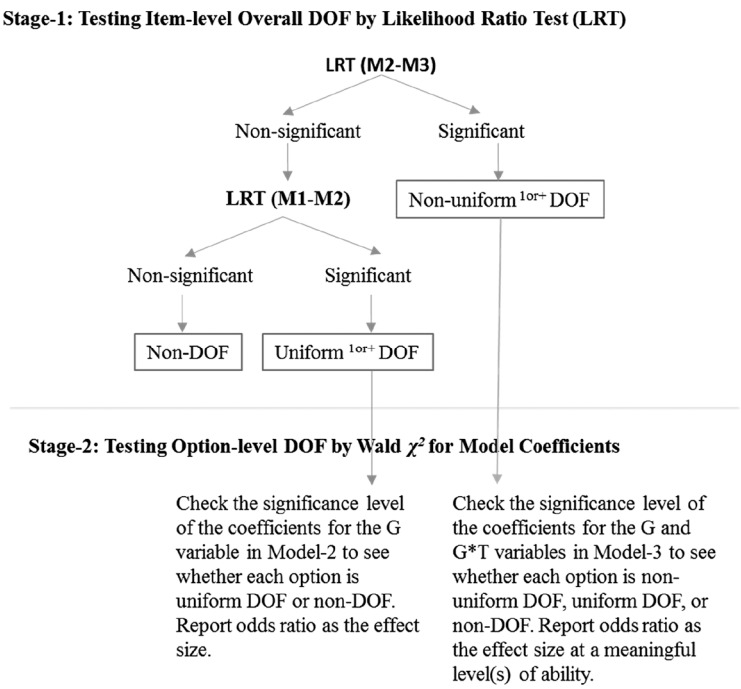


Figure 2 summarises our proposed two-stage procedures for studying DOF. The next section demonstrates the results based on the procedures in Figure 2 with the PIRLS data. All of the results were obtained using SPSS 16.0 and the OCCs were graphed in Excel.

Figure 2 Procedures for the proposed method of multinomial logistic regression for DOF



6 Results

6.1 Stage 1: Simultaneous DOF detection at the item level

The results for the LRT of Stage 1 are presented in Table 5. All 15 multiple-choice items were detected as showing either non-uniform^{1or+} or uniform^{1or+} DOF at the item level. That is, all items had at least one option that functioned differently for the two test-language groups. Six items (Items 2, 5, 9, 12, 13 and 15) were tested positive for non-uniform^{1or+} (the LRT for the non-uniform DOF was significant). The other nine items were tested positive for uniform^{1or+} DOF (the LRT for non-uniform^{1or+} was non-significant but was significant for uniform^{1or+}).

Table 5 Results for likelihood ratio tests for item-level simultaneous DOF

				Likelihood ratio test (dfΔ = 3)				
				Non-uniform (M2−M3)		*Uniform (M1−M2)*		
Item (key)	*Model*	*−2LL*	*df*	*−2LLΔ*	*p*	*−2LLΔ*	*p*	*DOF conclusion*
1 (D)	M1	760.386	3	0.693	0.875	**24.421**	**<0.001**	Uniform^{1or+}
	M2	735.965	6					
	M3	735.272	9					

Table 5 Results for likelihood ratio tests for item-level simultaneous DOF (continued)

Item (key)	*Model*	*−2LL*	*df*	*Likelihood ratio test (dfΔ = 3)*				*DOF conclusion*
				Non-uniform (M2−M3)		*Uniform (M1−M2)*		
				−2LLΔ	*p*	*−2LLΔ*	*p*	
2 (B)	M1	967.254	3	**23.486**	**<0.001**	**26.186**	**<0.001**	Non-Uniform[1or+]
	M2	941.068	6					
	M3	917.582	9					
3 (C)	M1	697.544	3	3.529	0.317	**15.952**	**0.001**	Uniform[1or+]
	M2	681.592	6					
	M3	678.063	9					
4 (C)	M1	605.613	3	4.830	0.185	**21.189**	**<0.001**	Uniform[1or+]
	M2	584.424	6					
	M3	579.594	9					
5 (A)	M1	903.986	3	**103.131**	**<0.001**	**39.550**	**<0.001**	Non-Uniform[1or+]
	M2	864.436	6					
	M3	761.305	9					
6 (A)	M1	859.426	3	5.262	0.154	**51.833**	**<0.001**	Uniform[1or+]
	M2	807.593	6					
	M3	802.331	9					
7 (D)	M1	652.577	3	5.800	0.121	**116.197**	**<0.001**	Uniform[1or+]
	M2	536.38	6					
	M3	530.58	9					
8 (C)	M1	771.601	3	3.940	0.268	**35.397**	**<0.001**	Uniform[1or+]
	M2	736.204	6					
	M3	732.264	9					
9 (B)	M1	862.649	3	**18.170**	**<0.001**	**25.751**	**<0.001**	Non-Uniform[1or+]
	M2	836.898	6					
	M3	818.728	9					
10 (A)	M1	1072.586	3	6.920	0.074	**98.992**	**<0.001**	Uniform[1or+]
	M2	973.594	6					
	M3	966.674	9					
11 (B)	M1	805.395	3	0.174	0.982	**42.279**	**<0.001**	Uniform[1or+]
	M2	763.116	6					

Table 5 Results for likelihood ratio tests for item-level simultaneous DOF (continued)

Item (key)	*Model*	*−2LL*	*df*	*Likelihood ratio test (dfΔ = 3)*				*DOF conclusion*
				Non-uniform (M2−M3)		*Uniform (M1−M2)*		
				−2LLΔ	*p*	*−2LLΔ*	*p*	
	M3	762.942	9					
12 (D)	M1	1151.615	3	**80.132**	**<0.001**	**184.581**	**<0.001**	Non-Uniform^{1or+}
	M2	967.034	6					
	M3	886.902	9					
13 (C)	M1	808.422	3	**8.748**	**0.033**	**20.753**	**<0.001**	Non-Uniform^{1or+}
	M2	787.669	6					
	M3	778.921	9					
14 (A)	M1	975.659	3	0.454	0.929	**16.014**	**0.001**	Uniform^{1or+}
	M2	959.645	6					
	M3	959.191	9					
15 (D)	M1	937.934	3	**11.391**	**0.010**	0.093	0.993	Non-Uniform^{1or+}
	M2	937.841	6					
	M3	926.450	9					

Note: −2LL: −2 times log likelihood; −2LLΔ: −2 times LL difference between two models; dfΔ: degrees of freedom difference between two models. Statistically likelihood ratio tests (LRT) at α = 0.05 are highlighted in bold. In the last column, an item was considered as showing at least one non-uniform DOF (non-uniform^{1or+}) if the LRT (M2−M3) was significant, as showing at least one uniform DOF (uniform^{1or+}) if the LRT (M1−M2) was significant but not the LRT (M2−M3), and as having no DOF if neither of the two LRTs was significant.

6.2 Stage 2: Individual DOF detection at the option level

6.2.1 Items detected as non-uniform^{1or+} DOF in Stage 1: Reporting Model 3

Table 6 reports the *Wald* χ^2 test for items that had at least one option detected as non-uniform^{1or+} in Stage 1. We chose two items, Item 5 and Item 12, to demonstrate the interpretation of the results.

Table 6 Wald χ^2 test results for regression coefficients in Model 3 (for individual options of non-uniform[1or+] DOF items)

Item (key)	*Option*	*Predictor*	*b*	*s. e.*	*Wald* χ^2	*p*	*Odds*	*DOF conclusion*
2 (B)	A	*T*	−0.567	0.113	25.009	<0.001	0.567	
	A	*G*	0.244	0.128	3.611	0.057	1.277	
	A	***T*G***	**−0.407**	**0.130**	**9.741**	**0.002**	**0.666**	**Non-uniform**
	C	*T*	−0.490	0.070	49.344	<0.001	0.613	
	C	***G***	**−0.284**	**0.077**	**13.614**	**<0.001**	**0.752**	
	C	***T*G***	**−0.391**	**0.085**	**21.337**	**<0.001**	**0.676**	**Non-uniform**
	D	*T*	−1.432	0.130	120.858	<0.001	0.239	
	D	*G*	−0.261	0.202	1.664	0.197	0.770	
	D	***T*G***	**−0.362**	**0.160**	**5.129**	**0.024**	**0.696**	**Non-uniform**
5 (A)	B	*T*	−1.066	0.185	33.305	<0.001	0.344	
	B	***G***	**0.767**	**0.243**	**9.938**	**0.002**	**2.153**	**Uniform**
	B	*T*G*	−0.237	0.204	1.353	0.245	0.789	
	C	*T*	−1.453	0.139	109.510	<0.001	0.234	No
	C	*G*	0.254	0.203	1.569	0.210	1.289	
	C	*T*G*	−0.051	0.163	0.100	0.752	0.950	
	D	*T*	0.020	0.097	0.044	0.833	1.021	
	D	***G***	**−0.777**	**0.117**	**44.230**	**<0.001**	**0.460**	
	D	***T*G***	**−1.235**	**0.124**	**98.959**	**<0.001**	**0.291**	**Non-uniform**
9 (B)	A	*T*	−1.080	0.114	89.273	<0.001	0.339	
	A	***G***	**−0.342**	**0.153**	**4.962**	**0.026**	**0.711**	**Uniform**
	A	*T*G*	0.000	0.137	0.000	1.000	1.000	
	C	*T*	−1.303	0.100	170.106	<0.001	0.272	
	C	***G***	**−0.328**	**0.134**	**5.972**	**0.015**	**0.720**	**Uniform**
	C	*T*G*	0.164	0.120	1.881	0.170	1.178	
	D	*T*	−0.336	0.126	7.087	0.008	0.715	
	D	***G***	**−0.391**	**0.142**	**7.571**	**0.006**	**0.677**	
	D	***T*G***	**−0.550**	**0.148**	**13.725**	**<0.001**	**0.577**	**Non-uniform**
12 (D)	A	*T*	−1.324	0.166	63.989	<0.001	0.266	No
	A	*G*	0.202	0.243	0.691	0.406	1.224	
	A	*T*G*	0.082	0.189	0.189	0.664	1.086	

Table 6 Wald χ^2 test results for regression coefficients in Model 3 (for individual options of non-uniform^{1or+} DOF items) (continued)

Item (key)	*Option*	*Predictor*	*b*	*s. e.*	*Wald* χ^2	*p*	*Odds*	*DOF conclusion*
	B	*T*	−0.659	0.096	47.403	<0.001	0.517	
	B	***G***	**−1.510**	**0.142**	**113.710**	**<0.001**	**0.221**	
	B	***T*G***	**−0.693**	**0.132**	**27.353**	**<0.001**	**0.500**	**Non-uniform**
	C	*T*	−0.412	0.069	36.061	<0.001	0.662	
	C	***G***	**−0.980**	**0.080**	**151.800**	**<0.001**	**0.375**	
	C	***T*G***	**−0.699**	**0.086**	**66.404**	**<0.001**	**0.497**	**Non-uniform**
13 (C)	A	*T*	−2.553	0.281	82.424	<0.001	0.078	
	A	***G***	**1.692**	**0.491**	**11.859**	**0.001**	**5.431**	
	A	***T*G***	**0.799**	**0.306**	**6.839**	**0.009**	**2.223**	**Non-uniform**
	B	*T*	−1.490	0.133	126.195	<0.001	0.225	No
	B	*G*	0.247	0.187	1.749	0.186	1.280	
	B	*T*G*	0.204	0.154	1.750	0.186	1.227	
	D	*T*	−0.409	0.076	29.187	<0.001	0.665	
	D	***G***	**0.275**	**0.079**	**12.062**	**0.001**	**1.317**	**Uniform**
	D	*T*G*	0.108	0.087	1.534	0.216	1.114	
15 (D)	A	*T*	−1.175	0.107	120.445	<0.001	0.309	
	A	*G*	−0.188	0.136	1.910	0.167	0.828	
	A	***T*G***	**−0.404**	**0.130**	**9.625**	**0.002**	**0.668**	**Non-uniform**
	B	*T*	−0.437	0.087	25.053	<0.001	0.646	
	B	*G*	−0.025	0.092	0.077	0.782	0.975	
	B	***T*G***	**−0.245**	**0.103**	**5.595**	**0.018**	**0.783**	**Non-uniform**
	C	*T*	−1.092	0.113	93.639	<0.001	0.336	No
	C	*G*	−0.001	0.139	0.000	0.996	0.999	
	C	*T*G*	−0.182	0.135	1.821	0.177	0.834	

Note: Options with significant regression coefficients of *G* or *T*G* are highlighted in bold. In the last column, as long as the regression coefficient of *T*G* was significant, the option was considered functioning non-uniformly; if the regression coefficient of *T*G* was non-significant but of *G* was significant, the option was considered functioning uniformly; if neither the coefficient of *G* nor of *T*G* was significant, the option was considered not functioning differentially.

Item 5 was flagged as non-uniform^{1or+} DOF at the item level in Stage 1, $-2LL\Delta_{M2-M3} = 103.13$, $df\Delta = 3$, and $p < 0.05$, which indicates at least one option showed *non-uniform* DOF (see Item 5 in Table 5). To examine the individual options of Item 5 in Stage 2, the

regression coefficients of the two predictors of *G* and *T***G* for each option (vs. keyed option A) from *Model 3* were examined (see Item 5 in Table 6). For option B, only the regression coefficient of *G*, Test Language, was significant, $b = 0.78$, $Wald\ \chi^2(1) = 9.94$, $p = 0.002$, and odds ratio = 2.15, indicating that option B showed uniform DOF. The odds of choosing option B (vs. the keyed option A) were 2.15 times higher for the French-test group than for the English-test group. This suggests that the French-test group was more likely to choose option B. For option C, both the regression coefficients for *G* and *T***G* predictors were non-significant, indicating option C showed no DOF. Lastly, for option D, the regression coefficient of *T***G* variable was significant, $b = -1.24$, $Wald\ \chi^2(1) = 98.96$, and $p < 0.001$, and odds ratio = 0.29, indicating significant non-uniform DOF. Remember that the odds ratio in the case of non-uniform DOF needs to be interpreted at each level of the total score because the magnitude of DOF varies depending on the level of the total score. For demonstrative purpose, the odds ratio of non-uniform DOF was interpreted only at the average total score (i.e. standardised total score = 0). Thus, the odds of choosing option D (vs. the keyed option A) was 3.45 times higher for the English-test group than for the French-test group at the average total score (the odds ratio was inverted for ease of interpretation, i.e. 1/0.29 = 3.45). The three graphs in first row of Figure 1 display the OCCs for Item 5 and summarise which and how the options function differentially.

Item 12 was also flagged as non-uniform^{1or+} DOF at the item level, $-2LL\Delta_{M2-M3} = 80.13$, $df\Delta = 3$, and $p < 0.05$. Therefore, regression coefficients of the two predictors of *G* and *T***G* for each option (vs. the keyed option D) from *Model 3* were examined for DOF at the option level in Stage 2 (see Item 12 in Table 6). Option A showed no DOF because neither the regression coefficient of *G* nor of *T***G* was statistically significant. The regression coefficients of *T***G* for both options B and C were significant, therefore, options B and C showed non-uniform DOF. For option B, $b = -0.69$, $Wald\ \chi^2(1) = 27.35$, $p < 0.001$, and odds ratio = 0.50; for option C, $b = -0.70$, $Wald\ \chi^2(1) = 66.40$, $p < 0.001$, and odds ratio = 0.50. Both option B and C were more appealing to the English-test group than the French-test group. The odds were two times higher (1/0.5 = 2) when comparing the groups at the average total score. The three graphs in the second row of Figure 1 display the OCCs for Item 12 and summarise which and how the options function differentially.

6.2.2 Items detected as uniform^{1or+} DOF in Stage 1: Reporting model 2

Table 7 reports the $Wald\ \chi^2$ test for items that had at least one option detected as uniform^{1or+} in Stage 1. We chose two items, Item 7 and Item 10, to demonstrate the interpretation of the results.

Table 7 Wald χ^2 test results for regression coefficients in Model 2 (for individual options of Uniform[1or+] DOF items)

Item (key)	*Option*	*Predictor*	*b*	*s.e.*	*Wald* χ^2	*p*	*Odds*	*DOF conclusion*
1 (D)	A	*T*	−1.143	0.073	245.562	<0.001	0.319	No
	A	*G*	−0.192	0.139	1.927	0.165	0.825	
	B	*T*	−1.111	0.059	355.618	<0.001	0.329	
	B	***G***	**0.513**	**0.126**	**16.484**	**<0.001**	**1.670**	**Uniform**
	C	*T*	−1.192	0.099	145.511	<0.001	0.304	
	C	*G*	0.385	0.205	3.542	0.060	1.470	No
3 (C)	A	*T*	−0.780	0.048	264.926	<0.001	0.458	
	A	*G*	−0.155	0.097	2.538	0.111	0.857	No
	B	*T*	−1.961	0.220	79.209	<0.001	0.141	
	B	*G*	−0.328	0.334	0.960	0.327	0.721	No
	D	*T*	−1.422	0.078	332.276	<0.001	0.241	
	D	***G***	**−0.530**	**0.136**	**15.211**	**<0.001**	**0.589**	**Uniform**
4 (C)	A	*T*	−1.944	0.167	135.147	<0.001	0.143	
	A	***G***	**−0.626**	**0.245**	**6.559**	**0.001**	**0.535**	**Uniform**
	B	*T*	−1.810	0.099	335.739	<0.001	0.164	
	B	***G***	**−0.538**	**0.151**	**12.707**	**<0.001**	**0.584**	**Uniform**
	D	*T*	−1.729	0.100	300.553	<0.001	0.177	
	D	***G***	**−0.463**	**0.157**	**8.728**	**0.003**	**0.630**	**Uniform**
6 (A)	B	*T*	−1.458	0.072	415.785	<0.001	0.233	
	B	***G***	**−0.601**	**0.127**	**22.470**	**<0.001**	**0.548**	**Uniform**
	C	*T*	−1.072	0.048	491.590	<0.001	0.342	
	C	***G***	**−0.584**	**0.088**	**43.706**	**<0.001**	**0.558**	**Uniform**
	D	*T*	−1.583	0.090	307.788	<0.001	0.205	
	D	***G***	**−0.378**	**0.162**	**5.437**	**0.020**	**0.685**	**Uniform**
7 (D)	A	*T*	−1.755	0.131	178.400	<0.001	0.173	
	A	***G***	**0.804**	**0.239**	**11.311**	**0.001**	**2.234**	**Uniform**
	B	*T*	−1.385	0.117	140.158	<0.001	0.250	
	B	***G***	**1.988**	**0.357**	**30.957**	**<0.001**	**7.304**	**Uniform**
	C	*T*	−1.198	0.072	273.525	<0.001	0.302	
	C	***G***	**1.380**	**0.182**	**57.169**	**<0.001**	**3.973**	**Uniform**
8 (C)	A	*T*	−0.951	0.062	236.080	<0.001	0.386	

Table 7 Wald χ^2 test results for regression coefficients in Model 2 (for individual options of Uniform^{1or+} DOF items) (continued)

Item (key)	*Option*	*Predictor*	*b*	*s.e.*	*Wald* χ^2	*p*	*Odds*	*DOF conclusion*
	A	***G***	**−0.701**	**0.118**	**35.304**	**<0.001**	**0.496**	**Uniform**
	B	*T*	−1.022	0.053	378.058	<0.001	0.360	
	B	***G***	**−0.219**	**0.105**	**4.406**	**0.036**	**0.803**	**Uniform**
	D	*T*	−1.362	0.138	97.780	<0.001	0.256	
	D	*G*	−0.165	0.258	0.406	0.524	0.848	No
10 (A)	B	*T*	−0.691	0.045	240.128	<0.001	0.501	
	B	***G***	**−0.376**	**0.088**	**18.173**	**<0.001**	**0.686**	**Uniform**
	C	*T*	−1.021	0.060	286.539	<0.001	0.360	
	C	***G***	**0.657**	**0.140**	**22.088**	**<0.001**	**1.930**	**Uniform**
	D	*T*	−0.737	0.044	277.148	<0.001	0.479	
	D	***G***	**0.587**	**0.101**	**33.843**	**<0.001**	**1.799**	**Uniform**
11 (B)	A	*T*	−1.354	0.068	395.165	<0.001	0.258	
	A	***G***	**0.630**	**0.139**	**20.620**	**<0.001**	**1.878**	**Uniform**
	C	*T*	−1.360	0.117	135.474	<0.001	0.257	
	C	*G*	0.279	0.224	1.547	0.214	1.322	No
	D	*T*	−0.815	0.044	337.059	<0.001	0.443	
	D	***G***	**0.507**	**0.096**	**28.106**	**<0.001**	**1.661**	**Uniform**
14 (A)	B	*T*	−1.028	0.059	306.369	<0.001	0.358	
	B	*G*	0.186	0.118	2.463	0.117	1.204	No
	C	*T*	−1.165	0.066	307.648	<0.001	0.312	
	C	***G***	**0.443**	**0.140**	**10.051**	**0.002**	**1.558**	**Uniform**
	D	*T*	−0.999	0.043	534.670	<0.001	0.368	
	D	***G***	**0.266**	**0.085**	**9.772**	**0.002**	**1.305**	**Uniform**

Note: Options with significant regression coefficients of *G* are highlighted in bold. In the last column, if the regression coefficient of *G* was significant, the option was considered functioning uniformly; if the regression coefficient of *G* was non-significant, the option was considered as not functioning differentially.

Item 7 was detected as uniform^{1or+} DOF at the item level in Stage 1, $-2LL\Delta_{M1-M2}$ = 116.20, $df\Delta$ = 3, and $p < 0.05$. To detect DOF at the option level in Stage 2, the regression coefficient of the predictor *G* for each option (vs. the keyed option D) from *Model 2* was examined. For all three non-keyed options, the regression coefficients of predictor of *G* were significant, indicating the presence of uniform DOF (see Item 7 in

Table 7). For option A, $b = 0.80$, *Wald* $\chi^2(1) = 11.31$, $p = 0.001$, and odds ratio = 2.23; for option B, $b = 1.99$, *Wald* $\chi^2(1) = 30.96$, $p < 0.001$, and odds ratio = 7.30. For option C, $b = 1.38$, *Wald* $\chi^2(1) = 57.17$, $p < 0.001$, and odds ratio = 3.97. The odds ratios indicate that the French group was more likely to choose all three non-keyed options. That is, all non-keyed responses were more appealing to the French group. The three graphs in third row of Figure 1 display the OCCs for Item 7 and summarise which and how the options function differentially.

Item 10 was also detected as uniform^{1or+} at the item level in Stage 1, $-2LL\Delta_{M1-M2} = 98.99$, $df\Delta = 3$, $p < 0.05$. For all non-keyed options of B, C and D, the regression coefficients of the predictor *G* in *Model 2* were significant indicating the presence of uniform DOF (see Item 10 in Table 7). For option B, $b = -0.38$, *Wald* $\chi^2(1) = 18.17$, $p < 0.001$, and odds ratio = 0.69. For option C, $b = 0.66$, *Wald* $\chi^2(1) = 22.09$, $p < 0.001$, and odds ratio = 1.93. For option D, $b = 0.59$, *Wald* $\chi^2(1) = 33.84$, $p < 0.001$, and odds ratio = 1.80. Options C and D were more appealing to the French-test group, but option B was more appealing to the English-test group. The three graphs in the last row of Figure 1 display the OCCs for Item 10 and summarise which and how the options function differentially.

Table 8 reports the DOF results which are summarised for the entire test. Of the total of 45 non-keyed options (3 for each of the 15 items), 10 options (22.2%) were detected as non-uniform DOF, 24 options (53.3%) were detected as uniform DOF, and 11 options (24.4%) were detected non-DOF when contrasted with the keyed options. Although it is not the focus of this paper, the DIF results based on likelihood ratio test of binary logistic regression were also reported in the last column of Table 8 for readers who are interested in the comparison.

Table 8 Test-level summary of DOF results

DOF				
Stage 1: Item-level	*Stage 2: Option-level*			
Non-uniform^{1or+}	*Non-uniform*	*Uniform*	*Non-DOF*	*DIF*
Item 2	3	0	0	Y
Item 5	1	1	1	Y
Item 9	1	2	0	Y
Item 12	2	0	1	Y
Item 13	1	1	1	Y
Item 15	2	0	1	N
Uniform^{1or+}				
Item 1	NA	1	2	N
Item 3	NA	1	2	Y
Item 4	NA	3	0	Y
Item 6	NA	3	0	Y
Item 7	NA	3	0	Y
Item 8	NA	2	1	Y

Table 8 Test-level summary of DOF results (continued)

DOF				
Stage 1: Item-level	*Stage 2: Option-level*			
Non-uniform^{1or+}	*Non-uniform*	*Uniform*	*Non-DOF*	*DIF*
Item 10	NA	3	0	N
Item 11	NA	2	1	Y
Item 14	NA	2	1	Y
Total # (%)	10 (22.2%)	24 (53.3%)	11 (24.4%)	12 (80%)

7 Discussion

Despite the potential versatility of DOF studies, previous research often limited DOF application to examining measurement invariance. However, there are many advantages to studying individuals' original responses when DOF is studied with different focuses extending beyond identifying causes of DIF. This paper explicates the potential uses of DOF as an independent investigation that can reveal in-depth and fruitful information by looking into individuals' original responses. To extend the currently limited application of DOF, we chose to use the more generic and neutral term, DOF. This term is more suitable for a broader range of measurement contexts and purposes for both achievement/aptitude tests and questionnaires/surveys.

To facilitate the application of DOF, we further proposed a set of integrated and straightforward procedures, based on maximum-likelihood multinomial logistic regression. There are three major advantageous features of this methods. First, the proposed method can be readily conducted by anyone who has experience in logistic regression using popular software packages. Second, although our method was demonstrated with only four response options and two groups, it can investigate a greater number of options and more than two groups without entailing any additional complexities. This method is capable of analysing multiple options and multiple groups simultaneously, which cannot be easily done with other methods. Finally, this method is comprehensive in terms of providing useful information such as statistical inference, effect size and OCC simultaneously that can enhance the understanding of DOF results.

Readers are reminded that logistic regression with categorical predictors, based on maximum likelihood estimation, can run into sparse data problems. The scenario of sparse data is considered problematic for typical utilities of logistic regression in substantive research where the focus is to predict or classify the outcome categories. However, when logistic regression is applied to investigate measurement invariance with an eye on examining how the response categories may function differently between groups, a sparse data scenario can be a manifestation that certain option does not work well for certain group because that option is never or hardly chosen by certain group. In other words, when the response categories themselves are the very subject of study, sparse data can reveal the possibility of DOF first hand.

We acknowledge the possibility that the stage 1 likelihood ratio test may suggest the presence of DOF at the item level but the stage 2 Wald χ^2 test detects no DOF at the option level. This scenario could happen because the Wald χ^2 test is based on a subset

of the sample (the subset who select the specific options being contrasted), hence has less statistical power than the likelihood ratio test, which is based on the entire sample (the likelihood of the entire data). Also, the logistic regression literature has documented that the Wald χ^2 test can lack of statistical power with sparse data (fewer than 10 event cases per predictor) (e.g. Peduzzi et al., 1996). In this case, we recommend that the likelihood ratio test rules over the Wald χ^2 test because for small to moderate sample sizes, the likelihood-ratio test is usually more reliable than the Wald χ^2 test (Agresti, 2012, p. 12). In this case, odds ratio for all options should be examined (despite the nonsignificant results of the Wald χ^2 test) to see which option deserves further attention.

Our proposed method suggests using both statistical inference (*p*-value) and odds ratio effect size *jointly* (neither precedes the other) for understanding DOF in Stage 2. This practice may encounter the following two scenarios: a nonsignificant Wald χ^2 test with a large observed odds ratio or a significant Wald χ^2 test with a small observed odds ratio. In both scenarios, we found that reviewing both the *p*-value and the effect size can inform mutually which option needs further attention. The former may happen when an option is rarely chosen by respondents (a sparse data problem, i.e. very small sample size for the contrasted options). If a suspiciously large effect size is observed, one should look into whether it is an unreliable result due to small sample size. The latter scenario indicates that the Wald χ^2 test detects a small effect. In this case, the researcher's discretion is needed to judge whether the small effect is too trivial to worth further attention.

We would like to point out that this paper is not the first to suggest the use of multinomial logistic regression and odds ratio for understanding DOF. As we noted in the literature review, Kato et al. (2009) introduced the use of multinomial logistic regression for DOF and Penfield (2008) was one of the first to introduce the use of odds ratio as the effect size of DOF (associated with a nominal response IRT model). However, our proposed method differs from theirs in many ways and lends to more streamlined and integrated procedures.

The approach of Kato et al. (2009) to DOF (or, in their term, DDF) is part of the larger investigation of DIF. Although they also employed a likelihood ratio test as the method for simultaneous detection of DOF at the item-level (which they viewed as DIF detection), they did not consider the existence of non-uniform DOF. Only one single likelihood ratio test was conducted to detect uniform DOF at the item level. Also, instead of taking advantage of inferential statistics of the Wald χ^2 test and the natural effect size measure of odds ratio in the formal theories of logistic regression, their determination of DOF at the option level was based on a descriptive statistics of mean absolute difference (MAD) between two OCCs (as they called RCCs). Unfortunately, most popular statistical packages do not provide a measure of MAD; researchers have to compute them for all the options of all the items in order to conclude which particular option is DOF. In contrast, most popular statistical packages, if not all, automatically provide Wald χ^2 tests and odds ratios for multinomial logistic regression that can be readily used for DOF conclusion and effect size interpretation.

The approach of Penfield (2008) using the odds ratio as the effect size of DOF is in line with our method. However, his odds ratio measure was obtained based on a fairly complicated IRT-based nominal response model which may be unfamiliar to many

applied researchers. This approach was further complicated by a revision of a typical IRT nominal response model in order to produce interpretable odds ratios. In contrast, our method based on multinomial logistic regression is straightforward as long as the researchers have some basic knowledge and experience in logistic regression.

Purification of the total scores is a regular step of DIF procedures. Purification iteratively removes the items that are detected as DIF in the previous run from the summing of the total scores in the subsequent run of DIF. A purified total score is considered as a better approximation of an individual's true ability. The procedures of purification were not considered in our demonstration of the DOF procedures. This was because all items were detected to have at least one option functioning differently. In general, it is expected that the proportion of items that have at least one option showing DOF will be large. Removing any DOF items from summing the total scores may not serve the original purpose of purification. That is, a total score based on a small number of items can be an even poorer approximation of individuals' true ability. It is, however, viable to purify the total score for a DOF study by removing items showing large DIF.

Through simulation, future studies can explore the power and Type-I error of the likelihood ratio test at the item level and of the Wald's χ^2 square test at the option level. The design factors can include the effect size (odds ratio), sample size, number of options within an item, number of options that are DOF within an item, as well as the different combination of types of DOF (non-uniform, uniform and non-DOF) within an item.

In closing, we believe that the terminologies and analytical procedures laid out in this paper streamline and simplifies the investigation of differential responding between groups. With these new developments, we hope to facilitate more empirical studies of DOF and enhance in-depth understanding of item responding.

Acknowledgements

The research was funded in part by the 2016 Research Grants Award of Paragon Testing Enterprises, a solely own subsidiary of the University of British Columbia in Canada.

References

Abedi, J., Leon, S. and Kao, J.C. (2008) Examining differential distractor functioning in reading assessments for students with disabilities. CRESST Rept. No. 743, University of Minnesota, National Center for Research on Evaluation, Standards, and Student Testing (CRESST), Minneapolis.

Agresti, A. (2012) 'Logit models for multinomial responses', In *Categorical Data Analysis,* 3rd ed., John Wiley & Sons, Inc, Hoboken, NJ. pp. 293–329.

Banks, K. (2006) 'A comprehensive framework for evaluating hypotheses about cultural bias in educational testing', *Applied Measurement in Education,* Vol. 19, No. 2, pp.115–132.

Banks, K. (2009) 'Using DDF in a post hoc analysis to understand sources of DIF', *Educational Assessment,* Vol. 14, No. 2, pp.103–118.

Barton, K.E. and Huynh, H. (2003) 'Patterns of errors made by students with disabilities on a reading test with oral reading administration', *Educational and Psychological Measurement,* Vol. 63, No. 4, pp.602–614.

Bock, R.D. (1972) 'Estimating item parameters and latent ability when responses are scored in two or more nominal categories', *Psychometrika,* Vol. 37, No. 1, pp.29–51.

Bolt, D.M., Cohen, A.S. and Wollack, J.A. (2001) 'A mixture item response model for multiple-choice data', *Journal of Educational and Behavioral Statistics*, Vol. 26, No. 4, pp.381–409.

Dorans, N.J., Schmitt, A.P. and Bleistein, C.A. (1992) 'The standardization approach to assessing comprehensive differential item functioning', *Journal of Educational Measurement*, Vol. 29, No. 4, pp.309–319.

Ferguson, C.J. (2009) 'An effect size primer: A guide for clinicians and researchers', *Professional Psychology: Research and Practice,* Vol. 40, No. 5, pp.532–538.

Green, B.F., Crone, C.R. and Folk, V.G. (1989) 'A method for studying differential distractor functioning', *Journal of Educational Measurement*, Vol. 26, No. 2, pp.147–160.

Hosmer Jr., D.W., Lemeshow, S. and Sturdivant, R.X. (2013) 'Logistic regression models for multinomial and ordinal outcome', In *Applied Logistic Regression*, 3rd ed., Vol. 398, John Wiley & Sons, Inc., Hoboken, NJ, pp.269–289.

Kato, K., Moen, R.E. and Thurlow, M.L. (2009) 'Differentials of a State reading assessment: item Functioning, distractor functioning, and omission frequency for disability categories', *Educational Measurement: Issues and Practice*, Vol. 28, No. 2, pp.28–40.

Marshall, S.P. (1983) 'Sex differences in mathematical errors: An analysis of distracter choices', *Journal for Research in Mathematics Education,* Vol. 14, No. 5, p.325.

Menard, S. (2002) 'Polytomous logistic regression and alternatives to logistic regression', In *Applied Logistic Regression Analysis*, 2nd ed. Sage, Thousand Oaks, CA, pp.91–101.

Middleton, K. and Laitusis, C.C. (2007) 'Examining test items for differential distractor functioning among students with learning disabilities', *ETS Research Report Series,* Vol. 2, p.i-34.

Peduzzi, P., Concato, J., Kemper, E., Holford, T.R. and Feinstein, A.R. (1996) 'A simulation study of the number of events per variable in logistic regression analysis', *Journal of Clinical Epidemiology*, Vol. 49, No. 12, pp.1373–1379.

Penfield, R.D. (2008) 'An odds ratio approach for assessing differential distractor functioning effects under the nominal response model', *Journal of Educational Measurement*, Vol. 45, No. 3, pp.247–269.

Penfield, R.D. (2010) 'Modeling DIF effects using distractor-level invariance effects: Implications for understanding the causes of DIF', *Applied Psychological Measurement,* Vol. 34, No. 3, pp.151–165.

Schmitt, A.P. and Bleistein, C.A. (1987) *Factors Affecting Differential Item Functioning for Black Examinees on Scholastic Aptitude Test Analogy Items (RR-87-23)*, Educational Testing Service, Princeton, NJ.

Schmitt, A.P. and Dorans, N.J. (1990) 'Differential item functioning for minority examinees on the SAT', *Journal of Educational Measurement*, Vol. 27, pp.67–81.

Suh, Y. and Bolt, D.M. (2011) 'A nested logit approach for investigating distractors as causes of differential item functioning', *Journal of Educational Measurement*, Vol. 48, No. 2, pp.188–205.

Suh, Y. and 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*, Vol. 28, No. 1, pp.48–67.

Swaminathan, H. and Rogers, H.J. (1990) 'Detecting differential item functioning using logistic regression procedures', *Journal of Educational Measurement*, Vol, 27, No. 4, pp.361–370.

Thissen, D., Steinberg, L. and Wainer, H. (1993) 'Detection of differential item functioning using the parameters of item response models'. In Holland, P.W. and Wainer, H. (Eds.): *Differential Item Functioning*, Lawrence Erlbaum, Hillsdale, NJ, pp.67–113.

Veale, J. and Foreman, D. (1983) 'Assessing cultural bias using foil response data: Cultural variation', *Journal of Educational Measurement,* Vol. 20, No. 3, pp.249–258.

Westers, P. and Kelderman, H. (1991) 'Examining differential item functioning due to item difficulty and alternative attractiveness', *Psychometrika,* Vol. 57, No. 1, pp.107–118.

Zumbo, B.D. (1999). *A Handbook on the Theory and Methods of Differential Item Functioning (DIF)*, National Defense Headquarters, Ottawa.

Appendix: Methods used for detecting differential options functioning during 1980s–2010s

Paper/Report	*Obd %*	*Log linear*	*STD*	*IRT*	*Odds ratio*	*LCA*	*BLR*	*MLR*
Veal and Foreman (1983)	•							
Marshall (1983)		•						
Schmitt and Bleistein (1987)			•					
Green et al. (1989)		•						
Schmitt and Droans (1990)			•					
Westers and Kelderman (1991)						•		
Dorans et al. (1992)			•					
Thissen et al. (1993)				•				
Bolt et al. (2001)				•				
Barton and Huynh (2003)		•						
Bank (2006)		•						
Middleton and Laitusis (2007)			•					
Abedi et al. (2008)					•		•	
Penfield (2008)				•	•			
Bank (2009)		•			•			
Kato et al. (2009)								•
Penfield (2010)				•	•			
Suh and Bolt (2011)				•				

Note: Obd %: Observed proportion; STD: standardisation; LCA: latent class analysis; BLR: binary logistic regression; MLR: multinomial logistic