Assessing Testlet Effect, Impact, Differential Testlet, and Item Functioning Using Cross-Classified Multilevel Measurement Modeling

Hamdollah Ravand¹

Abstract
The present study used the two-level testlet response model (MMMT-2) to assess impact, differential item functioning (DIF), and differential testlet functioning (DTLF) in a reading comprehension test. The data came from 21,641 applicants into English Masters’ programs at Iranian state universities. Testlet effects were estimated, and items and testlets that were functioning differentially for test takers of different genders and majors were identified. Also parameter estimates obtained under MMMT-2 and those obtained under the two-level hierarchical generalized linear model (HGLM-2) were compared. The results indicated that ability estimates obtained under the two models were significantly different at the lower and upper ends of the ability distribution. In addition, it was found that ignoring local item dependence (LID) would result in overestimation of the precision of the ability estimates. As for the difficulty of the items, the estimates obtained under the two models were almost the same, but standard errors were significantly different.

Keywords
DIF, DTLF, impact, MMMT-2, multilevel models

Item measures may be affected by person grouping factors such as gender, L1 background, and ethnic background, among others, as well as by item grouping factors such as common input or stimulus, common response format, and item chaining, to name but a few. In either case, item difficulty is affected by a factor irrelevant to the main construct; hence, construct validity of the test composed of the items is threatened. The effect of person grouping factors can be studied through impact and differential item functioning (DIF) analysis, and the effect of item grouping factors can be captured by studying testlet effect.

Fair assessment requires invariance of measures across different samples within the same population. Males and females of comparable abilities, belonging to the same population, for example, are expected to have equal probabilities of giving a correct response to any given DIF-free item. Furthermore, all types of statistical analyses in social sciences hinge upon the assumption of independence of observations. People or things that are nested within a hierarchy tend to perform in a more similar way than people or things in other clusters, which results in local person dependence (LPD). For example, the answers that students in a given classroom or school give to a set of test items may be consistently more similar to those of the students in other classes or schools. Violation of the local person independence assumption leads to underestimation of standard errors (SEs), which in turn results in spurious rejection of null hypotheses (Hox, 2010). Multilevel models (MLMs) have been designed to handle the interdependencies among the data points. Interrelatedness among a set of items, referred to as local item dependence (LID), can also pose some problems when it is neglected. Possible causes of LID include a shared passage, response format (open-ended vs. multiple choice), speededness, and practice effect. The concept of testlet has been proposed to capture LID (Wainer & Kiely, 1987). Research has shown that neglecting testlet effect results in inaccurate estimation of reliability, person ability, and item difficulty (Lee, 2004; Wainer & Lukhele, 1997).

Different models have been proposed to study DIF (e.g., logistic regression, Mantel Hanzel, item response theory [IRT]–based models, and multiple indicators multiple causes [MIMIC] models). However, when local item independence is violated, DIF detection may be affected (Bolt, 2002). Researchers have taken one of the following approaches to

¹Vali-e-Asr University of Rafsanjan, Iran

Corresponding Author:
Hamdollah Ravand, Vali-e-Asr University of Rafsanjan, Rafsanjan, Iran.
Email: ravand@vu.ac.ir

Creative Commons CC BY: This article is distributed under the terms of the Creative Commons Attribution 3.0 License (http://www.creativecommons.org/licenses/by/3.0/) which permits any use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access page (http://www.uk.sagepub.com/aboutus/openaccess.htm).
address LID: (a) Testlet data have been fitted to score-based polytomous IRT models such as the graded response model (Samejima, 1969), polytomous logistic regression (Zumbo, 1999), or polytomous SIBTEST (Penfield & Lam, 2000). In these polytomous item response models, each testlet with \( m \) questions is treated as an item with the total score of the items within each testlet ranging from 0 to \( m \). (b) An item-based testlet response theory (TRT) model such as the two-parameter logistic TRT (2PL-TRT; Bradlow, Wainer, & Wang, 1999), three-parameter logistic TRT (3PL-TRT; Wainer, Bradlow, & Du, 2000), and the Rasch testlet model or one-parameter logistic TRT (1PL-TRT; W.-C. Wang & Wilson, 2005) has been employed. (c) An item-based multilevel TRT model such as the three-level testlet response theory model (MMMT-3; Jiao, Wang, & Kamata, 2005) or the two-level cross-classified testlet response theory model (MMMT-2; Beretvas & Walker, 2012) has been employed.

Score-based approaches have been criticized on the following grounds: (a) They do not take into account the exact response patterns of test takers to individual items within a testlet; therefore, a lot of information would be lost (Eckes, in press); and (b) applying polytomous IRT models to capture testlet effect has been reported to lead to biased parameter estimates and substantial overestimation of reliability and test information values (Thissen, Steinberg, & Mooney, 1989; Wainer, 1995).

Neither polytomous IRT models nor TRT models permit simultaneous modeling of both LID and LPD commonly encountered in educational assessment settings. In social and behavioral sciences, typically, students are nested within schools, and schools are in turn nested within neighborhoods. Ignoring LPD might have consequences (e.g., biased parameter estimates) as serious as those of ignoring LID. An advantage of item-based multilevel TRT models such as MMMT-3 and MMMT-2 is that besides taking LID into account, they can also take into account dependencies as a result of higher clustering of data. If data are item scores on a test consisting of testlets administered to test takers in a school, adding a fourth level to MMMT-3 (e.g., Jiao, Kamata, Wang, & Jia, 2012) or a third level to MMMT-2 can capture the nested structure of the data. Another clear benefit of MMMT-2 and -3 is that they permit simultaneous assessment of impact and DIF by adding person predictors to the model. But it is not possible to assess differential testlet functioning (DTLF) with MMMT-3. Instead, MMMT-2 by benefiting from the flexibilities of cross-classified multilevel modeling can simultaneously test for impact, DIF, and DTLF. In this formulation rather than having a separate level for testlet, person and testlet are modeled at Level 2 using dummy-coded testlet indicator variables, as explained below.

Abstuse technicalities of multilevel measurement models (MLMs) have prevented educational practitioners from applying these models where they are more appropriate. The present study is an attempt to demonstrate the application of MMM-2 to second-language data and discuss the implications of ignoring the hierarchical structure of the data. To this end, an accessible presentation of MMMT-2 is presented. Then, the data are analyzed, and the implications of ignoring LID are discussed.

### MMM-2

In MMM-2, items at Level 1 are cross-classified by testlets and persons at Level 2. For a test of \( q \) items and \( m \) testlets, the log-odds of a correct response to item \( i \) for person \( j \) at Level 1 is expressed in Equation 1:

\[
\log \left( \frac{P_{ij}}{1 - P_{ij}} \right) = \pi_{0j} + \pi_{1j}X_{ij} + \cdots + \pi_{(q-1)j}X_{(q-1)ij} + \pi_{qj}X_{q+ij} + \cdots + \pi_{mqj}X_{mq+1j} + \pi_{TjTmij} + \pi_{T0j} + \pi_{TmjTmij},
\]

where \( X \) and \( T \) are item and testlet indicators, respectively. \( X \) and \( T \) are dummy-coded indicators with “−1” for the relevant testlet and item and 0 otherwise. To identify the model, one dummy-coded item for each testlet is dropped. According to Beretvas and Walker (2008), for each testlet of \( q \) items, there are \((q-1)\) dummy-coded item indicators; there is one reference item per testlet, and for the \( m \) testlets, there are \( m \) testlet indicators; there is no reference testlet.

The Level 2 model with fixed item and testlet effects can be presented as in Equation 2:

\[
\begin{align*}
\pi_{0j} & = \theta_{0j} \\
\pi_{1j} & = \beta_10 \\
& \vdots \\
\pi_{(q-1)j} & = \beta_{(q-1)0} \\
\pi_{q+1j} & = \beta_{q+10} \\
& \vdots \\
\pi_{mq-1j} & = \beta_{mq-10} \\
\pi_{T0j} & = \beta_{T00} \\
& \vdots \\
\pi_{Tmj} & = \beta_{Tmj0},
\end{align*}
\]

where \( \beta_{q0} \) to \( \beta_{(q-1)0} \) represent the difficulties of Items \( q \) to \( q-1 \) in the first testlet, \( \beta_{(q+1)0} \) to \( \beta_{(mq-1)0} \) represent the difficulties of Items \( q+1 \) to \( mq-1 \) in Testlet \( m \) (\( q-1 \) and \( mq-1 \) indicate that for each testlet, one item is excluded from the model as the reference item), and \( \beta_{T00} \) to \( \beta_{Tmj0} \) represent the fixed effects of Testlets 1 to \( m \). Therefore, the probability that test taker \( j \) answers Item \( q \) in Testlet \( d \) correctly is as follows:

\[
P_{qjd} = \frac{1}{1 + \exp(-\eta_{qd})} = \frac{1}{1 + \exp\{-[\theta_{0j} - (\beta_{q0} + \beta_{T0d})]\}}.
\]

Algebraically, Equation 3 is equivalent to the Rasch model. Person ability parameter, \( \theta_{0j} \), corresponds to \( \theta_j \). In MMMT-2, two aspects of item difficulty are modeled (Beretvas &
Walker, 2012): difficulty as a result of the item effect, $\beta_{00}$, and item-specific testlet difficulty, $\beta_{T00}$. The item difficulty of the reference items in each testlet is equal to the relevant testlet’s difficulty. A notable difference in the conceptualization of testlet effect in MMMT-3 and MMMT-2 is that in MMM-3, testlet effect is person-specific, that is, it contributes to person ability, but in MMMT-2, it is item-specific, that is, it contributes to the difficulty of the items in the respective testlets.

Random testlet effects can be modeled as presented in 
Equation 4:

$$\pi_{T1j} = \beta_{T10} + u_{T1j}$$
$$\pi_{T2j} = \beta_{T20} + u_{T2j}$$
$$\vdots$$
$$\pi_{Tmj} = \beta_{Tm0} + u_{Tmj},$$

where $u_{T1j}$ to $u_{Tmj}$ represent random testlet effects.

Based on fit indices of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) or significance of the testlet effect variance component, researchers can decide whether a testlet with random effects fits better or one with fixed effect. If the random effects for one or more testlets are statistically significant, person-level predictors can be incorporated into the relevant Level 2 equations to account for the variance. Inclusion of Level 2 categorical person predictors in the testlet effect equations is a test of DTLF. As Beretvas and Walker (2012) demonstrated, one of the benefits of MMMT-2 is that it can simultaneously test for impact, DIF, and DTLF.

A MMMT-2 that tests for gender impact, DIF, and DTLF can be expressed as shown in Equation 5:

$$\beta_{0j} = \gamma_{00}Female_j + h_{0j}$$
$$\beta_{ij} = \gamma_{i0} + \gamma_{i1}Female_j$$
$$\vdots$$
$$\beta_{(mq-1)j} = \gamma_{(mq-1)0}$$
$$\beta_{T1j} = \gamma_{T10} + \gamma_{T11}Female_j + u_{T1j}$$
$$\vdots$$
$$\beta_{Tmj} = \gamma_{Tm0} + u_{Tmj},$$

where $\gamma_{10}$ to $\gamma_{(mq-1)0}$ represent the difficulty of Items 1 to $q$ in Testlet $m$, $\gamma_{10}$ to $\gamma_{Tm0}$ represent fixed testlet effects, $u_{T1j}$ and $u_{Tmj}$ represent random testlet effects, $\gamma_{01}$, $\gamma_{11}$, and $\gamma_{T11}$ provide measures of gender impact, DIF, and DTLF, respectively. As is evident, Equation 5 tests for impact, DIF, DTLF, and the random effect of testlets simultaneously. If, after the inclusion of gender, the variance for the random effect of testlet is still significant, more person-level predictors can be included.

All DIF and DTLF effects are the result of a secondary factor or dimension assessed by an item or a cluster of items (Zumbo, 2007). The random testlet effect $u_{Tij}$ in MMMT-2 corresponds to the secondary dimension that the testlet might be measuring, which might be interpreted as construct-irrelevant or as a complementary dimension. In other words, the random testlet effect is the effect of testlet on the difficulty of the items within the testlet, which is due to the secondary dimension targeted by the testlet. However, the fixed testlet effect $\gamma_{T00}$ can be interpreted as the direct contribution of the common stimulus to the difficulty of an item on the primary dimension being measured by the item. In other words, the fixed testlet effect represents the effect of the testlet on the difficulty of the items, which is due to the primary dimension intended to be assessed by the testlet. Therefore, as Beretvas and Walker (2012) argued, “Addition of a fixed effect for each testlet permits more specific decomposition of a testlet’s contribution to an item’s functioning in addition to an assessment of the person’s testlet ability” (p. 208).

Research Questions

The present study aimed to analyze testlet effect, impact, DIF, and DTLF in the reading comprehension section of the University Entrance Examination (UEE) for applicants into English Masters’ programs at state universities in Iran. This part of the test measured test takers’ ability to comprehend and respond to written academic passages. There were three reading passages with 7, 6, and 7 items, respectively. The reading texts, along with the items associated with them, defined the three testlets to be analyzed.

The research questions in the present study were as follows:

Research Question 1: To what extent do the reading passages show testlet effects?
Research Question 2: Do the reading comprehension items display gender and major impact and DIF?
Research Question 3: Do the reading comprehension passages display gender and major DTLF?
Research Question 4: How do the parameter estimates obtained under MMMT-2 compare with the ones obtained using HGLM-2?

Method

Participant

The data for the present study came from all the applicants (21,641, 71.3% female and 26.8% male) into English Masters’ programs at state universities in Iran. UEE, as a sole selection criterion, is a very stringent test that screens the applicants into English Teaching, English Literature, and Translation Studies programs at the MA level in Iran. The UEE is administered once a year in February. The data came from the 2012 version of the test. The participants were Iranian nationals and mainly holding a BA degree in English Literature and Translation Studies (88%). About 12% of the participants held a BA degree other than English.
Instrument and Procedures

The UEE measures General English (GE) and content knowledge of the applicants. The GE section consists of 10 grammar items, 20 vocabulary items, 10 cloze items, and 20 reading comprehension items. The content knowledge part consists of three separate parts with 60 items each. Each part is intended for the applicants to English Teaching, English Literature, or Translation Studies programs. All the questions are multiple choice, and test takers have to complete both the GE and the content knowledge parts in 120 min, 60 min each. For the purpose of the present study, only the reading comprehension data of the GE part were used. There were three reading passages in the 2012 version of the test. The first passage was a relatively long passage of 720 words on natural selection theory, followed by seven questions. The second passage, about 524 words, discussed wastes and the precautions taken by governments against its harmful effects and was followed by six comprehension questions. The third passage, 584 words long, discussed the changes in the family structure and function over the centuries, followed by seven questions.

Data Analysis

To answer the research questions, MMMT-2 was estimated using SAS PROC GLIMMIX (SAS Institute, 2006). PROC GLIMMIX is designed for generalized linear mixed models’ estimation and is a flexible procedure commonly used for estimations in MMMs (Flom, McMahon, & Pouget, 2007).

Results

Testlet Effects

Testlet effects obtained under MMMT-2 and MMMT-3 are presented in Table 1. As it was previously alluded to, under MMMT-3, the testlet effect variance is assumed to be constant (fixed) across all the testlets. Therefore, SAS PROC GLIMMIX generated one testlet effect variance for all the three testlets under MMMT-3 and a separate testlet effect for each testlet under MMMT-2.

Remember that the testlet effect variance indicates the degree of LID among the items associated with a given testlet. When there is no LID, the testlet effect variance equals 0. The higher the variance, the more the items of a testlet are locally dependent. In the absence of any commonly accepted criteria for judging the magnitude of the testlet effect variance, it was decided to follow two kinds of evidence available in the literature: (a) evidence provided by simulation studies, wherein variances lower than 0.25 are considered negligibly small (Glas, Wainer, & Bradley, 2000; W.-C. Wang & Wilson, 2005; O. Zhang, Shen, & Cannady, 2010) and (b) reference to empirical studies, which considered testlet effect variances of 0.50 to 2.00 as substantial (X. Wang, Bradlow, & Wainer, 2002; B. Zhang, 2010). According to the above criteria, the omnibus variance effect of 0.27, obtained under MMMT-3, was somewhat slightly above the criterion suggested by simulation studies and still below the guideline proposed by empirical studies. The interesting point is that the MMMT-3 testlet effect is the average of the separate testlet effects generated by MMMT-2 for each testlet. According to Table 1, the variance for Testlet 2 is negligible, judged by criteria suggested by either simulation or empirical studies. The variances for Testlets 1 and 3 are above the criterion suggested by simulation studies and below the guideline suggested by empirical studies, hence considered medium.

Impact and DIF

Impact refers to “difference in person abilities as a function of some person-level predictors” (Beretvas, Cawthon, Lockhart, & Kaye, 2012). In the present study, two person predictors were added to the intercept model \( \beta_0j \) in Equation 5 to test for impact. For gender, male test takers were coded 0 and females were coded 1. As for major, English majors were coded 0 and non-English majors were coded 1. As MMMT-3 does not permit simultaneous impact, DIF, and DTLF tests, in what follows, the results generated by MMMT-2 formulation are provided. As Table 2 shows, the impact of major was not significantly different from 0. On the contrary, the impact of gender was significantly different from 0 (the estimate was at least twice its SE). The negative sign of the estimate indicated that, on average, females had significantly higher ability estimates than males (as females were coded 0).

DIF refers to significant difference in item difficulties across different groups in the same population, which are matched for ability. Differences in item difficulties and discriminations are referred to as uniform and non-uniform DIF, respectively. MMM formulations in general and MMMT-2
in particular permit only tests of uniform DIF. Table 2 presents major and gender DIF estimates for the items that have been flagged for gender and/or major DIF. The last item in each testlet was dropped as reference items (Items 7, 13, and 20 for Testlets 1, 2, and 3, respectively), and Item 1 was not flagged for either gender or major DIF.

As Table 2 shows, all the items flagged for gender DIF (items with DIF estimates twice their SEs), except for Item 12, have got negative signs, which indicate that they favored females (gender was coded 0 for females and 1 for males). From the 10 items flagged for major DIF, Items 2, 5, 6, and 8 with negative values favored non-English-language majors, and the items with positive values favored English majors (major was coded 0 for non-English-language majors and 1 for English majors).

**Sources of DTLF**

The concepts of DTLF refer to differential measurement properties of a cluster of items for different groups in the same population, conditioning on ability. As it was stated above, testlet effect, under MMMT-2, is decomposed into two: (a) random effect, which corresponds to the secondary or added dimension assessed by a testlet, and (b) fixed effect, which represents the contribution of the common stimulus to the difficulty of an item obtained from the primary dimension assessed by a testlet (Beretvas & Walker, 2012). In the MMMT-2 formulation, for identification purposes, one item is dropped per testlet as the reference item, the difficulty of which is the fixed effect of the respective testlet. The fixed effects of the testlets (testlet difficulties) were 0.86, 0.24, and −1.00 for Testlets 1, 2, and 3, respectively. As the estimates suggest, Testlet 1 was the hardest testlet followed by Testlets 2 and 3. In the present study, the flexibility of MMM-2 was employed to investigate DTLF, DIF, and impact simultaneously. Gender and major DTLF results are presented in Table 3.

As for major, all of the testlets were functioning differentially. Inspecting the right section of Table 3, the reader notes that the most difficult testlet (Testlet 1) and the least difficult one (Testlet 3) favored English majors, whereas the second testlet was easier for non-English majors than for English majors. Gender DTLF test, as shown in the left side of Table 3, revealed that Testlets 1 and 3 were functioning significantly differently for males and females (the estimates are twice as large as the respective SEs). Testlet 1 was easier for females, whereas Testlet 3 was slightly easier for males.

| Testlet | Gender DTLF | SE | Major DTLF | SE |
|---------|-------------|----|------------|----|
| 1       | −0.16       | 0.02| 0.04       | 0.01|
| 2       | −0.3        | 0.02| −0.03      | 0.01|
| 3       | 0.03        | 0.01| 0.03       | 0.01|

**Table 2. Gender and Major Impact and DIF Estimates.**

| Item | Gender DIF | SE | Major DIF | SE |
|------|------------|----|-----------|----|
| 2    | −0.3       | 0.05| −0.2      | 0.03|
| 3    | −0.2       | 0.05| −0.2      | 0.03|
| 4    | −0.2       | 0.05| −0.2      | 0.03|
| 5    | −0.2       | 0.05| 0.2       | 0.03|
| 6    | 0.1        | 0.05| 0.1       | 0.05|
| 7    | 0.5        | 0.05| 0.1       | 0.05|
| 8    | 0.07       | 0.03| 0.03      | 0.03|
| 9    | 0.07       | 0.03| 0.03      | 0.03|
| 10   | −0.3       | 0.05| −0.2      | 0.03|
| 11   | 0.2        | 0.06| 0.2       | 0.06|
| 12   | −0.2       | 0.05| 0.2       | 0.05|
| 13   | 0.03       | 0.01| 0.03      | 0.01|

Note. DIF = differential item functioning.
unobserved sources of DTLF are presented in Table 1 above. As the random effects of Testlet 2 were small, person-level predictors were added only to the testlet part of Equation 5 for Testlets 1 and 3. Addition of gender significantly reduced the testlet effect variances to 0.17 and 0.23 (compare the variances with those in Table 1). As the new variances suggest, addition of gender reduced testlet effects to below 0.25, which are negligibly small, judged either by the criterion suggested by simulation studies or by the guideline suggested by empirical studies. The amount of variation in testlet effects explained by gender can be measured by computing how much the testlet effect variance component has diminished between the model with unobserved and the model with observed (here, gender) source of differential functioning (Singer, 1998). Following Bryk and Raudenbush (1992), the amount of variance explained can be computed for each testlet as (testlet effect in the unobserved model − testlet effect in the model with predictors) / testlet effect in the unobserved model as follows:

Testlet 1 \( \frac{0.27 - 0.17}{0.27} = 0.37 \)

Testlet 3 \( \frac{0.40 - 0.23}{0.40} = 0.42 \)

These can be interpreted by stating that for the first and second testlets, 37% and 42% of the explainable variation, respectively, are explained by gender.

Addition of major as an observed source of DTLF also resulted in the reduction of testlet effect variances to 0.20 and 0.29 for Testlets 1 and 3, respectively. The amount of variance explained by major in each testlet is as follows:

Testlet 1 \( \frac{0.27 - 0.20}{0.27} = 0.25 \)

Testlet 3 \( \frac{0.40 - 0.29}{0.40} = 0.27 \)

The interpretation is that major accounts for 25% and 27% of the explainable variance in Testlets 1 and 3, respectively. The random effect of Testlet 3 is still above the criterion suggested by simulation studies, implying that there are still unobservable sources affecting the testlet’s variance, and more person predictors can be added to the model to account for the variance.

### Comparison of Parameter Estimates Obtained Under MMMT-2 and Two-Level HGLM Rasch Model

To investigate the effect of ignoring LID in the data, person ability and item difficulty estimates along with their SEs generated under MMMT-2 (Beretvas & Walker, 2012), which takes into account LID and the two-level HGLM Rasch model (Kamata, 2001), which ignores LID, were compared.

A paired-samples \( t \) test was run to compare the ability estimates generated under HGLM-2 and MMMT-2. The results showed that the mean ability estimates were not significantly different, \( t(21640) = 0.047, p = .93 \). Test takers were divided into three groups: low, high, and mid, according to their bachelor’s grade point averages (GPAs). A separate paired-samples \( t \) test was run in each group to compare the ability estimates produced by the two models. The results of the analyses are displayed in Table 4.

As is evident from Table 4, ability estimates generated by the two models were significantly different at the two ends of the ability distribution. At the lower end, there was a significant difference between the ability estimates produced by MMMT-2 \( (M = -22.262, SD = 0.780) \) and those generated by the HGLM-2 \( (M = -24.128, SD = 0.845) \) model, \( t(7353) = 24.025, p = 0.000 \). Estimates produced by MMMT-2 were higher than the estimates generated by HGLM-2, but the estimates by HGLM-2 showed more dispersion at this part of the ability distribution. In the middle of the ability distribution, the estimates generated by the two models were almost the same. At the upper end, the estimates produced by HGLM-2 \( (M = 24.983, SD = 0.883) \) were significantly higher than those produced by MMMT-2 \( (M = 23.070, SD = 0.815) \) formulation, \( t(7209) = -23.674, p = .000 \). As the sample size in the present study was very large and with large samples even very small differences can turn out to become statistically significant, effect size was also calculated. Interpreted according to the guidelines proposed by Cohen (1988; \( 0.01 = \) small effect, \( 0.06 = \) moderate effect, \( 0.14 = \) large effect), \( \eta^2 \) (one of the most commonly used effect size statistic) statistics in Table 4 show slightly above-medium effect sizes for the difference between the ability estimates generated by the two models, both at the lower and upper ends of the ability continuum.

The precision with which each of the two models estimated person abilities was investigated by comparing SEs of ability estimates. As Table 5 shows, the SEs of the ability estimates generated by MMMT-2 and HGLM-2 were significantly different at all levels (lower, mid, and upper end) of the ability distribution. \( \eta^2 \) values provided in the last column of Table 5 show very large effect sizes at all levels of the ability distribution.

### Table 4. Paired-Samples \( t \) Tests of Ability Estimates.

| GPA (binned) | \( M \) | \( t \) | \( df \) | Significance (two-tailed) | \( \eta^2 \) |
|-------------|-------|-------|-------|--------------------------|-------|
| Pair 1      | MMMT-2–HGLM-2 | 0.018659 | 24.02 | 7.353 | .000 | .07 |
| Pair 2      | MMMT-2–HGLM-2 | 0.000168 | 0.21 | 7.076 | .834 | —   |
| Pair 3      | MMMT-2–HGLM-2 | -0.019131 | -23.67 | 7.209 | .000 | .07 |

*Note. GPA = grade point average; HGLM-2 = two-level hierarchical generalized linear model; MMMT-2 = two-level cross-classified testlet response theory model.*
The difficulty estimates produced by MMMT-2 ($M = -0.26111100$, $SD = 0.979504495$) and HGLM-2 ($M = -0.26111100$, $SD = 0.979504495$) were exactly the same. But the precision with which difficulties were estimated by MMMT-2 ($M = 0.02268$, $SD = 0.0031$) and HGLM-2 ($M = 0.02316$, $SD = 0.0031$) formulations, $t(18) = -6.908, p = 0.000$, was significantly different. The eta squared statistics (.7) indicated a very large effect size.

### Summary of the Results and Discussion

The present study investigated testlet effects within the context of the reading comprehension section of the UEE for MA applicants into English programs at Iranian state universities. Drawing on the unique benefits of multilevel measurement modeling in general and MMMT-2 (Beretvas & Walker, 2012) in particular, the present study simultaneously assessed impact, DIF, and DTLF for males versus females and English majors versus non-English majors, and investigated the effect of ignoring item clustering on the estimates of item difficulty, person ability, and their associated SEs.

The first research question concerned the magnitude of testlet effect for each of the three testlets in the reading comprehension section of UEE. The result obtained under MMMT-3 indicated a medium testlet effect. According to the results generated by MMMT-2, Testlets 1 and 3 showed medium effects, and Testlet 2 showed negligible testlet effect. One of the interesting findings of the present study was that the omnibus testlet variance obtained under MMMT-3 was exactly the average of the three separate testlet effect variances generated by MMMT-2. As the variances for the random effects of the Testlets 1 and 3 were of medium size, gender and major were separately included in the model. Inclusion of gender accounted for 42% and 37% of the explainable variance in Testlets 1 and 3, respectively. Major explained 25% and 27% of the explainable variances in Testlets 1 and 3, respectively.

From among the TRT models capable of assessing testlet DIF, MMMT-2 uniquely permits assessing for both observed and unobserved (unmeasured) sources of DTLF. Along with the separate addition of gender and major as measured sources of DTLF, the random effects of testlets were added to the model, besides their fixed effects. Statistically non-significant random effect variances for Testlets 1 and 3 after addition of gender suggested that there were no latent (unobserved) factors causing testlets to function differentially. Major, separately added to the model, did not leave any significant testlet effect variance unexplained in Testlet 1 but left some statistically significant variance in Testlet 3. Therefore, deeper inspection of Testlet 3 was needed to assess what other unmeasured factors might have been the cause of DTLF to be included in the model.

Gender and major impact assessment indicated that, on average, females were significantly more able than males, but English and non-English majors did not show significant difference in their ability estimates. Nine items were flagged for gender DIF and 10 items were flagged for major DIF. To test for unobserved sources of DIF, besides fixed effect of the items, their random effects could have also been included in the model. To save time, it was decided not to employ this unique capability of MMMT-2 to test for unobserved sources of DIF in the items. Addition of each random effect would increase processing time exponentially. As for DTLF, all the testlets were flagged for major DTLF; Testlets 1 and 3 favored English majors, whereas Testlet 2 favored non-English majors. Only Testlet 1 showed gender DTLF, which favored females.

To answer the last research question, parameter estimates obtained under the model taking LID into account (MMMT-2 model) and the model ignoring LID (HGLM-2) were compared. The results indicated that the ability estimates obtained under the two models were significantly different at the lower and upper ends of the ability distribution but they were not significantly different in the middle. The results also indicated that ignoring LID would result in overestimation of the precision of the ability estimates. As for the difficulty of the items, the estimates obtained under the two models were exactly the same but SEs were significantly different, suggesting that ignoring LID results in lower SEs, hence overestimation of the precision of item difficulty estimates.

The findings regarding ability estimates, their SEs, and the SEs of item difficulty estimates are in line with those of the previous studies, which have shown that ignoring LID would result in biased estimation of ability parameters at low and high ends of ability distribution (Ackerman, 1987; Tuerlinckx & De Boeck, 2001) and underestimation of SEs (Thissen et al., 1989; Wainer, Sireci, & Thissen, 1991; Yen, 1984, 1993). Ackerman (1987) and Tuerlinckx and De Boeck...
(2001) also found that violation of LID leads to inflated item difficulty estimates, a finding that was not corroborated by the present study.

Another interesting finding of the present study was the fact that all the three items in the first testlet flagged for major DIF favored non-English majors, but the testlet as a whole favored English majors. Had I employed a conventional differential bundle functioning (DBF) method, the DIF and DTLF in Testlet 1 would have canceled each other out as they displayed differential functioning in opposite directions. Unlike conventional DBF, the cross-classified multilevel testlet model (MMMT-2; Beretvas & Walker, 2012) partitions a testlet’s DBF into a part common to all items in a testlet (i.e., DIF) and the part that is unique to each item (i.e., DTLF).

Limitations and Suggestions for Further Research

In the present study, the author tested for DIF and DTLF. The cause of differential functioning of items and testlets was not focused upon. Future research can inspect the items qualitatively to find what qualities of the items or testlets might give test takers of one gender or major an upper hand over the other. Qualitative inspection of the subskills required for each item along with the topic of the reading passages might shed some light on differential item and testlet functioning. The present study investigated the effect of ignoring LID on person and item parameter estimates. Future research can (a) compare parameter estimates obtained under models with and without DIF and DTLF, (b) investigate the effect of ignoring LID on DIF and impact detection, (c) investigate the effect ignoring local person dependence on DIF and DTLF detection, and (d) investigate the effect of simultaneous ignoring of LPD and LID on DIF and DTLF detection. As to the criterion for selecting the reference items, previous studies have excluded the last item as the reference item (e.g., Jiao et al., 2005; Kamata, 2001). To achieve continuity with the literature, I dropped the last item in each testlet in the present study. Future research can study the effect of changing the reference item on parameter estimates, impact, DIF, and DTLF detection.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research and/or authorship of this article.

References

Ackerman, T. (1987). *The robustness of LOGIST and BILOG IRT estimation programs to violations of local independence* (Research Report 87-14). Retrieved from http://www.act.org/research/researchers/reports/pdf/ACT_RR87-14.pdf.

Beretvas, S. N., Cawthon, S., Lockhart, L., & Kaye, A. (2012). Assessing impact, DIF and DFF in accommodated item scores: A comparison of multilevel measurement model parameterizations. *Educational and Psychological Measurement, 72*, 754-773. doi:10.1177/0013164412440998

Beretvas, S. N., & Walker, C. M. (2008). Use of the multilevel measurement model for testlets for differential testlet functioning. Unpublished manuscript.

Beretvas, S. N., & Walker, C. M. (2012). Distinguishing differential testlet functioning from differential bundle functioning using the multilevel measurement model. *Educational and Psychological Measurement, 72*, 200-223. doi:10.1177/0013164411412768

Bolt, D. M. (2002). A Monte Carlo comparison of parametric and nonparametric polytomous DIF detection methods. *Applied Measurement in Education, 15*, 113-141. doi:10.1207/S15324818AME1502_01

Bradlow, E., Wainer, H., & Wang, X. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*, 153-168. doi:10.1007/BF02294533

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods—Advanced quantitative techniques in the social sciences*. Newbury Park, CA: SAGE.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum.

Eckes, T. (in press). Lokale Abhängigkeit von Items im TestDaF-Leseverstehen: Eine Testlet-Response-Analyse [Local item dependence in the TestDaF reading section: A testlet response analysis]. *Diagnostica*.

Flom, P., McMahon, I., & Pouget, E. R. (2007). Using PROC NLMIXED and PROC GLMMIX to analyze dyadic data with binary outcomes. *SAS Global Forum 2007*. Retrieved from www2.sas.com/proceedings/forum2007/179-2007.pdf

Glas, C.A.W., Wainer, H., & Bradlow (2000). MML and EAP estimates for the testlet response model. In W.J. van der Linden & C.A.W. Glas (Eds.), *Computer Adaptive Testing: Theory and Practice* (pp.271-287). Boston, MA: Kluwer-Nijhoff Publishing.

Hox, J. J. (2010). *Multilevel analysis: Techniques and applications*. New York, NY: Routledge.

Jiao, H., Kamata, A., Wang, S., & Jin, Y. (2012). A multilevel testlet model for dual local dependence. *Journal of Educational Measurement, 49*, 82-100. doi:10.1111/j.1745-3984.2011.00161.x

Jiao, H., Wang, S., & Kamata, A. (2005). Modeling local item dependence with the hierarchical generalized linear model. *Journal of Applied Measurement, 6*, 311-321.

Kamata, A. (2001). Item Analysis by the Hierarchical Generalized Linear Model. *Journal of Educational Measurement, 38*, 79-93. doi: 10.1111/j.1745-3984.2001.tb01177.x

Lee, Y.-W. (2004). Examining passage-related local item dependence (LID) and measurement construct using Q3 statistics in an EFL reading comprehension test. *Language Testing, 21*, 74-100. doi:10.1191/0265532204lt260oa

Penfield, R. D., & Lam, T. C. M. (2000). Assessing differential item functioning in performance assessment: Review and recommendations. *Educational Measurement: Issues and Practice, 19*(3), 5-15. doi:10.1111/j.1745-3992.2000.tb00033.x
Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Vol. 17). Richmond, VA: Psychometric Society.

SAS Institute Inc. (2006). SAS Version (9.2) [Computer Software]. Cary, NC: Author.

Singer, F. M. (1998). Relations Lineaires entre Solutions d’une Equation Differentielle (with E. Compoint) [Linear relationships between a differential equation solutions]. *Annales des Fac. des Science de Toulouse*, 6, 659-670.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement*, 26, 247-260. doi:10.1111/j.1745-3984.1989.tb00331.x

Tuerlinckx, F., & De Boeck, P. (2001). The effect of ignoring item interactions on the estimated discrimination parameters in item response theory. *Psychological Methods*, 6, 181-195.

Wainer, H. (1995). Precision and differential item functioning on a testlet-based test: The 1991 Law School Admission Test as an example. *Applied Measurement in Education*, 8, 157-186. doi:10.1207/s15324818ame0802_4

Wainer, H., Bradlow, E., & Du, Z. (2000). Testlet Response Theory: An analog for the 3PL model useful in Testlet-Based Adaptive Testing. In W. Linden & G. W. Glis (Eds.), *Computerized Adaptive Testing: Theory and practice* (pp. 245-269). Netherlands: Springer.

Wainer, H., & Kiely, G. L. (1987). Item Clusters and Computerized Adaptive Testing: A case for testlets. *Journal of Educational Measurement*, 24, 185-201. doi:10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., & Lukhele, R. (1997). How reliable are TOEFL scores? *Educational and Psychological Measurement*, 57, 741-758.

Wainer, H., Sireci, S. G., & Thissen, D. (1991). Differential testlet functioning: Definitions and detection. *Journal of Educational Measurement*, 28, 197-219. doi:10.1111/j.1745-3984.1991.tb00354.x

Wang, W.-C., & Wilson, M. (2005). Assessment of differential item functioning in testlet-based items using the Rasch Testlet Model. *Educational and Psychological Measurement*, 65, 549-576. doi:10.1177/0013164404268677

Wang, X., Bradlow, E. T., & Wainer, H. (2002). A general Bayesian Model for testlets: Theory and applications. *Applied Psychological Measurement*, 26, 109-128. doi:10.1177/0146621602026001007

Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8, 125-145. doi:10.1177/014662168400800201

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30, 187-213. doi:10.1111/j.1745-3984.1993.tb00423.x

Zhang, B. (2010). Assessing the accuracy and consistency of language proficiency classification under competing measurement models. *Language Testing*, 27, 119-140. doi:10.1177/0265532209347363

Zhang, O., Shen, L., & Cannady, M.(2010, April). *Polytomous IRT or testlet model: An evaluation of scoring models in small testlet size situations*. Paper presented at the Annual Meeting of the 15th International Objective Measurement Workshop, Boulder, CO, USA.

Zumbo, B. D. (1999). *A handbook on the theory and methods of Differential Item Functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (Ordinal) item scores*. Ottawa, Ontario, Canada: Directorate of Human Resources Research and Evaluation, Department of National Defense.

Zumbo, B. D. (2007). Three generations of DIF analyses: Considering where it has been, where it is now, and where it is going. *Language Assessment Quarterly*, 4, 223-233. doi:10.1080/15434300701375832

**Author Biography**

Hamdollah Ravand is an assistant professor at Vali-e-Asr University of Rafsanjan where he is teaching Masters’ courses of Research Methods in TEFL, Language Testing, Statistics & Computer, and Advanced Writing. His major areas of interest are Language Testing & Assessment, Cognitive Diagnostic Modeling, Structural Equation Modeling, Multilevel Modeling, and Item Response Theory.