Local item independence is one assumption made under item response models. Local item independence is achieved when the probability of answering an item correctly is unaffected by the probability of answering other items correctly, conditional on the person and item parameters (Hambleton and Swaminathan, 1985; Embretson and Reise, 2000). Item response models may not be robust to the violation of the local item independence assumption. Previous studies (e.g., Ackerman, 1987; Chen and Thissen, 1997; Spray and Ackerman, 1987; Tuerlinckx and De Boeck,
2001a and 2001b; Yen, 1984) indicated that the presence of local item dependence (LID) had negative impact, such as on person and item parameter estimation, and test equating.

Factors that cause LID vary. For example, Yen (1993), Ferrara, Huynh, and Baghi (1997), and Ferrara, Huynh, and Michaels (1999) found that locally dependent and independent item clusters were differentiated by contextual characteristics, specifically, content and response requirements. For instance, in reading tests, items may be text dependent, and some items may share the same reading passage. In math content clusters, questions in the same cluster may depend on the same data table. If this is the case, the responses to items in the same item cluster would rely on how well the examinee understands the common stimuli. Thissen, Bender, Chen, Hayashi, and Wiesen (1992) explored the causes of LID by classifying LID into underlying local dependence and surface local dependence. The underlying local dependence model is similar to what Yen (1993), Ferara et al. (1997) and Ferara et al. (1999) assumed, as described above. The surface local dependence, on the other hand, considered test situations where examinees respond to a pair of items very similarly. For example, when a test is speeded, some items are not reached by many examinees. If this happens, examinees’ responses to these not-reached items are naturally very similar. Hoskens and De Boeck (1997) also classified LID into two categories. One refers to order dependency in which item order causes the dependency. The other is combination dependency that is similar to what the above-mentioned authors assumed in which the dependency is the “gestalt” that a set of items can form. Although there are different factors that can cause LID, a contextual effect on a cluster of items by common stimuli, such as a reading passage, is a common source of LID. Such an item cluster is often referred to as a testlet (e.g., Thissen, Steinberg, and Mooney, 1989; Wainer and Kiely, 1987, Wang, Bradlow, and Wainer, 2002; Wainer and Wang, 2000, 2001).

Studies related to LID have used different indices and methods to model LID. Yen (1984), for example, proposed the $Q_3$ statistic representing the correlation between two item scores controlling for the ability level. Chen and Thissen (1997) proposed four statistics to quantify LID, where two of them can be used to detect LID and multidimensionality. Ferrara, Huynh, and Baghi (1997) presented an approach to identify LID based on raw test scores. It checks the magnitude of inter-item correlations for examinees at different intervals on the test score scale. Douglas, Kim, Habing, and Gao (1998) investigated the local dependence of item
pairs using conditional covariance functions. Wang, Cheng, and Wilson (2005) used a multidimensional item response model to detect specific forms of LID. Hoskens and De Boeck (1997) and Tuerlinckx and De Boeck (1999) modeled item main effects and item interaction effects to account for LID. They proposed two models for LID: a constant interaction model and a dimension-dependent interaction model. Ip (2002) set up the reproducible and nonreproducible local dependence kernels to model LID based on conditional distributions describing multiple item responses as a function of ability without assuming local independence. Other studies (e.g., Bradlow, Wainer, and Wang, 1999; Du, 1998; Wainer and Wang, 2000; Wang, Bradlow, and Wainer, 2002) developed models to incorporate a parameter into unidimensional item response models, which indicates the interaction between person and item cluster.

This paper proposes a model that examines LID from the multilevel modeling perspective, where it is assumed that LID is present because of the contextual effects of the items’ association with a common stimulus, such as in testlets.

**Theoretical Framework**

*Hierarchical generalized linear model*

The hierarchical linear model (HLM; Bryk and Raudenbush, 1992) permits an appropriate modeling of nested or hierarchical data structures. According to de Leeuw (1992), HLM relaxes two basic assumptions made in traditional linear model analyses, namely, homoscedasticity, and independence of observations. Individuals in the same group are more similar than individuals in different groups. Thus, students in different classrooms can be assumed independent, but students in the same classroom share something in common. From the variance component model perspective, individual components are all independent and group components are independent between groups but correlated within groups. Some groups may be more homogeneous than other groups. Then, the variance of the group components differs. The same logic can be applied to the relationship between test items and their associated item clusters.

The hierarchical generalized linear model (HGLM) is an extension of the generalized linear model (McCullagh and Nelder, 1989) to hierarchical data. It deals with cases where the assumption of normality at level-1 is not reasonable, such as when the outcome variable is dichotomous. Several multilevel modeling software programs can be used to estimate HGLM
parameters. In the current study, HLM6 software (Raudenbush, Bryk, Cheong, and Congdon, 2004) was used.

Several studies (e.g., Adams, Wilson, and Wu, 1997; Fox and Glas, 1998; Kamata, 1999, 2001) have successfully combined item response model and multilevel modeling. Among them, Kamata (1999, 2001) demonstrated a formulation of the Rasch-equivalent two-level HGLM and its three-level extension. It treats the person ability parameter as a random effect. The two-level HGLM is extended to a three-level model that allows modeling the variation of students’ performance across groups such as classrooms and schools, as well as the interaction effect of person and group characteristic variables (Kamata, 2001). The three-level HGLM is formulated with level-1 as the item-level model, level-2 as the person-level model, and level-3 as the group-level model. In this three-level HGLM model formulation, items are fully crossed with the persons, and persons are nested within groups. The hierarchy of Kamata’s three-level HGLM model is shown in Figure 1.

This study proposes a LID model due to item cluster contextual effects. Then, the hierarchy structure becomes such that items are nested within item clusters, and then item clusters are fully crossed with the persons as shown in Figure 2. Here, the clustering of persons within groups is not modeled.

*Figure 1.* The Hierarchy of Kamata’s Three-level HGLM, in which Items (Level-One) are nested within Persons (Level-Two) clustered within Sites (Level-Three).
Bayesian random effects model for testlet

Bradlow, Wainer, and Wang (1999) proposed a Bayesian random effects model for testlets. Their study specified a frequently encountered assessment situation when a set of items is associated with a single common stimulus, such as a common reading passage. Their model for testlets is an extension of a two-parameter probit item response model. According to Bradlow et al., a latent score \( t_{ij} \) is

\[
t_{ij} = a_i (\theta_j - b_i - \gamma_{jd(i)}) + \varepsilon_{ij},
\]

where \( a_i, b_i, \gamma_{jd(i)}, \theta_j \) are item discrimination, item difficulty, a person-specific testlet effect, and examinee’s ability, respectively. \( \varepsilon_{ij} \) is a unit normal variate indicating the randomness in response \( y_{ij} \) for examinee \( j \) on item \( i \), with

\[
y_{ij} = \begin{cases} 
1 & \text{if } t_{ij} > 0 \\
0 & \text{otherwise.} 
\end{cases}
\]

\( \gamma_{jd(i)} \) is the same for all items within a testlet for a particular examinee \( i \) and is independent of ability and item parameters with \( \gamma_{jd(i)} \sim N(0, \sigma^2_\gamma) \). Therefore, the magnitude of LID is represented by \( \sigma^2_\gamma \).

A multilevel parameterization of Bradlow et al. (1999) model is reformulated here using the HGLM. The model is a simplified version in

Figure 2. The Hierarchy of the Proposed Three-level HGLM to Model Local Item Dependence, in which Items (Level-One) are nested within Item Clusters (Level-Two) within Persons (Level-Three).
which the item discrimination is assumed constant (set to a value of one) for all items and a logistic model is used, instead of a probit model.

The level-1 model is set up by expressing the log-odds of person \( j \) answering item \( i \) in item cluster \( m \) using a linear regression equation that includes an intercept term. Level-1 model is

\[
\log \left( \frac{p_{imj}}{1 - p_{imj}} \right) = \eta_{imj} = \beta_0 + \sum_{q=1}^{k-1} \beta_{qmj} X_{qimj},
\]

(2)

where \( p_{imj} \) is the probability that person \( j \) answers item \( i \) in item cluster \( m \) correctly. \( X_{qimj} \) is the \( q \)th dummy variable for person \( j \), with value 1 when \( q = i \), and 0 when \( q \neq i \), for item \( i \) within item cluster \( m \). The coefficient \( \beta_{0mj} \) is an intercept term, and \( \beta_{qmj} \) is a coefficient associated with \( X_{qimj} \), where \( q = 1, \ldots, k-1 \). To achieve full rank for the design matrix of the model, one of the dummy variables in the equation is dropped. In this study, the dummy variable for the last item is dropped. The individual item effect \( \beta_{qmj} \) is the difference from the overall effect \( \beta_{0mj} \). Then, the probability that person \( j \) answers item \( i \) within item cluster \( m \) correctly is expressed as

\[
p_{imj} = \frac{1}{1 + \exp(-\eta_{imj})}.
\]

The level-2 model models the item cluster level effect. It is

\[
\beta_{0mj} = \gamma_{00j} + \nu_{0mj}, \quad \text{and} \quad \beta_{qmj} = \gamma_{q0j},
\]

(3)

where \( q = 1, \ldots, k-1 \), \( \gamma_{00j} \) is the fixed effect of the level-1 intercept, and \( \nu_{0mj} \) is a random effect of the level-1 intercept. The random effect \( \nu_{0mj} \) can be conceptualized as an interaction between item cluster and ability. This is analogous to a person-specific item cluster effect \( \gamma_{jkd} \) in Bradlow et al. (1999) formulation (see Equation 1). In Equation 3, \( \gamma_{q0j} \) is the item-specific effect for item with the \( q \)th dummy variable. Also, it is assumed that \( \nu_{0mj} \sim N(0, \sigma_u^2) \), and \( \sigma_u^2 \) is analogous to \( \sigma^2 \) in Bradlow et al., which provides the magnitude of LID.

The level-3 models person-level effects. It decomposes \( \gamma_{00j} \) at the level-2 model (Equation 3) into a fixed part and a random part. The random part is the person effect, which is equivalent to the person’s ability. The effects for the items remain fixed, although they don’t have to, in the level-3 model. Therefore, the level-3 model is

\[
\gamma_{00j} = \pi_{000} + \psi_{00j}, \quad \text{and} \quad \gamma_{q0j} = \pi_{q00}
\]

(4)
where \( q = 0, \ldots, k - 1, w_{00q} \sim N(0, \sigma_w^2) \). \( \sigma_w^2 \) is the variance of the ability distribution. This current study is investigating the simplest version of this model.

The proposed model can be applied to testing situations, in which a test consists of clusters of several items that are associated with a common stimulus. Two assumptions are made studying this model. One assumption is that there is no interdependence between item clusters, and the other is that there is no dependence between items that are associated with different item clusters.

**Method**

The proposed multilevel method of modeling LID was evaluated by analyzing simulated data sets. Data sets were randomly generated by assuming different degrees of LID, based on the modified Bayesian random effects model for testlets (see Equation 5) of Bradlow et al. (1999). Then, the proposed three-level HGLM (HGLM3) was fitted to estimate parameters. Also, the same data sets were analyzed by the Rasch-equivalent two-level HGLM (HGLM2), in which item clusters were not modeled and instead level-2 represented the person-level effects.

The true values of the person ability parameter were randomly generated from a standard normal distribution, where the mean = 0 and the standard deviation = 1. Sample size was fixed at 1000. The number of item clusters was fixed at 6, where each item cluster contained 5 items. The item difficulty parameters for the five items in each item cluster were fixed at five values of \(-1, -0.5, 0, 0.5, \) and 1. Based on Bradlow et al. (1999) as presented in Equation 5, LID parameter values were simulated from a normal distribution \( N(0, \sigma_u^2) \) by specifying different levels of LID variance \( \sigma_u^2 \). When there was no LID, \( \sigma_u^2 = 0 \). The magnitude of LID was set at four levels by specifying \( \sigma_u^2 \) at 0, 0.5, 1.0, or 1.5. Then, the LID parameters, \( u_{(mj)} \), were randomly sampled from \( N(0, \sigma_u^2) \). The LID estimate \( u_{(mj)} \) was the same for all items within the

<table>
<thead>
<tr>
<th>Table 1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation Study Conditions</td>
</tr>
<tr>
<td>Sample Size</td>
</tr>
<tr>
<td>1000</td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
<tr>
<td></td>
</tr>
</tbody>
</table>
same item cluster for the same examinee. Therefore, the number of the LID
effects was the number of item clusters times the number of examinees. In
each simulation condition there were $6 \times 1000 = 6000$ LID effects. The four
simulation conditions are summarized in Table 1.

For each simulation condition, item responses were generated by incor-
porating the true ability, true item parameters, and LID parameters
into the modified testlet model of Bradlow et al.

$$p_{\text{imj}} = \frac{\exp[\beta_j - D_i - u_{\text{imj}}]}{1 + \exp[\beta_j - D_i - u_{\text{imj}}]},$$

where $\beta_j$ is the person ability, $D_i$ is the item difficulty, $u_{\text{imj}}$ is the LID
effect and $p_{\text{imj}}$ is the probability of a correct response. Once the item responses
were generated, two models, HGLM2, and the proposed HGLM3, were
fitted to estimate their parameters. HLM6 software (Raudenbush, Bryk,
Cheong, and Congdon, 2004) was used to estimate the two HGLM models’
parameters. In HLM6, parameters were estimated by a sixth order approxi-
mation to the likelihood for the model based on a Laplace transform.

Parameter estimates were examined by referencing them to the true
parameter values used for data generation. Also, HGLM2 and the proposed
HGLM3 were compared on their fixed effect estimates (item parameter
estimates) and estimates of person-level random effect variance (ability vari-
ance estimates). Since this study examined only one set of generated data
for each simulation condition, estimation errors were not evaluated.

**Results**

The level-2 model in the HGLM3 was the item-cluster level model,
where the LID effect was modeled. The variance of the level-2 random

Table 2

| Estimation of Level-2 Variance Components for HGLM3 (LID Magnitude Esti-
| mates) |
|---|---|---|
| Simulation Condition | Variance | Standard Error of Variance |
| $\sigma_{\text{u}}^2 = 0$ | 0.0005* | n/a** |
| $\sigma_{\text{u}}^2 = 0.25$ | 0.2873 | 0.0301 |
| $\sigma_{\text{u}}^2 = 1$ | 1.0919 | 0.0548 |
| $\sigma_{\text{u}}^2 = 2.25$ | 2.3925 | 0.0984 |

*The estimate was obtained using the PQL method [$\chi^2(df = 5000) = 4743.29$, $p > .500$].

**This solution did not converge and thus no standard error estimate was provided.
Local item dependence indicates the magnitude of the LID variation, namely, the magnitude of local item dependence. The estimates of the level-2 variance and its standard error for the four simulation conditions are summarized in Table 2. With the increase of the LID magnitude from condition S1 ($\sigma_u^2 = 0$) to S4 ($\sigma_u^2 = 2.25$), the estimate of the level-2 variance of random effects increased accordingly. This clearly reflected the LID magnitude in the simulated data. When the estimates are compared to the true values, some are very close to the true values, while some others are slightly different. Specifically, the level-2 variance estimates were reasonably close to the simulated LID magnitudes for $\sigma_u^2 = 0.0$, $\sigma_u^2 = 0.25$, and $\sigma_u^2 = 1.0$ conditions, while it was of a difference of 0.14 for $\sigma_u^2 = 2.25$ condition. Since only a single data set was generated for each studied condition, this might be just a sampling fluctuation. Note that the estimate for $\sigma_u^2 = 0$ condition (0.0005) was based on the PQL estimation, because the Laplace estimation failed to converge. It was probably the case because the true value of $\sigma_u^2$ was 0.

Estimated item difficulties for each simulation condition are graphically presented in Figures 3a-d. The four figures present the comparison between the estimated item difficulty parameters using HGLM2 and HGLM3 for each item when different degrees of LID were simulated (Figure 3a for $\sigma_u^2 = 0.0$, Figure 3b for $\sigma_u^2 = 0.25$, Figure 3c for $\sigma_u^2 = 1.0$, and Figure 3d for $\sigma_u^2$.

![Figure 3](image_url). Estimated Item Parameters for HGLM2 and HGLM3.
Note that the PQL estimates were reported for the simulation condition with $\sigma^2_u = 0.0$, because there was no Laplace estimates for this condition. For the other three conditions, Laplace estimates were reported in the figure. In each of these figures, the horizontal axis represents the item number. Recall that there were 30 simulated items, where five items were clustered within each of six item clusters. So, the displayed item numbers on the horizontal axis (items 1, 6, 11, 16, 21, and 26) represent the first item in each of the six item clusters. The same set of hypothesized item difficulty values were assumed for the five items in each item cluster, in the same order of $-1, -0.5, 0, 0.5, \text{ and } 1$. The vertical axis represents the estimated item difficulty. The two clustered bars for each item represent the estimated item difficulty values from HGLM2 and HGLM3.

The results indicate that the item difficulty estimates using HGLM2 and HGLM3 were almost identical for all items in the $\sigma^2_u = 0$ condition. For the other conditions, the magnitude of the item difficulty estimates from HGLM2 were most often underestimated as compared with those from HGLM3 especially when the magnitude of LID was larger, such as in the $\sigma^2_u = 1.0$ and 2.25 conditions. Therefore, it is speculated that item difficulty estimates are affected in such a way that their true values are underestimated when the presence of LID is ignored. This is consistent with what was found in Snijders and Bosker (1999) about the effect of ignoring a source of variance. When ignoring a source of variance, the ignored variance will increase the error variance. Then, all effects will look smaller in comparison with the standardized error variance. Note that some difficulty estimates from HGLM3 were larger than hypothesized difficulties in their absolute values, while corresponding estimates from HGLM2 were closer to the hypothesized difficulties. However, this does not necessarily mean that HGLM2 provided better estimates for those cases, since the simulation results were based on only one random sample of item responses. In order to reveal such estimation errors, more extensive simulation studies with multiple replications will be needed.

Table 3 summarizes the mean absolute difference between estimated difficulties and hypothesized true difficulties. As mentioned above, it is not possible to evaluate the amount of estimation error for each parameter. Instead, our focus here is to summarize how parameter estimates are affected on average by referencing to the hypothesized true difficulties as the LID magnitude changes. When $\sigma^2_u = 0$, the mean absolute difference was about the same for HGLM2 and HGLM3. However, as the magnitude
of LID increases, the mean absolute difference increased more for HGLM2 than for HGLM3. This implies that HGLM3 estimates were less affected by the presence of LID and their magnitudes.

The estimates of the ability distribution variance components from HGLM2 and HGLM3 are summarized in Table 4. For HGLM2, this estimate is the variance component of the level-2 random effects, while it is the variance component of the level-3 random effects for HGLM3. When the ability parameters were generated for the simulated data, the standard normal distribution with a mean of 0 and a variance of 1 was assumed. Therefore, it was our anticipation that the estimates would be close to 1.0. For HGLM2, the estimated variance of ability distribution was very close to 1.0 when \( \sigma^2_u = 0 \). However, it was shrunken when the LID was present, and the degree of shrinkage increased as the LID magnitude increased. This is due to ignoring a source of variance (Snijders and Bosker, 1999). Practically, this would result in under-discrimination of examinee ability.

Table 4

<table>
<thead>
<tr>
<th>LID Magnitude ((\sigma^2_u))</th>
<th>Simulation Condition</th>
<th>Mean Absolute Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>(0)</td>
<td>S1</td>
<td>0.0459</td>
</tr>
<tr>
<td>(0.25)</td>
<td>S2</td>
<td>0.0600</td>
</tr>
<tr>
<td>(1)</td>
<td>S3</td>
<td>0.0920</td>
</tr>
<tr>
<td>(2.25)</td>
<td>S4</td>
<td>0.1493</td>
</tr>
</tbody>
</table>

*The mean absolute difference was computed based on the item parameter estimates obtained from the PQL method since there was no Laplace estimate for item parameters under the no LID condition.

Table 3

<table>
<thead>
<tr>
<th>Mean Absolute Difference between the Estimated and True Item Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>LID Magnitude ((\sigma^2_u))</td>
</tr>
<tr>
<td>-------------------------------</td>
</tr>
<tr>
<td>(0)</td>
</tr>
<tr>
<td>(0.25)</td>
</tr>
<tr>
<td>(1)</td>
</tr>
<tr>
<td>(2.25)</td>
</tr>
</tbody>
</table>

*There were no estimates of these parameters using Laplace estimation due to the lack of convergence of the model.
ties, that is, the examinee ability estimates tend to cluster around its mean and the variability of the ability estimates would be smaller than its true values. On the other hand, the estimates of the ability distribution variance were nearly unaffected by the LID magnitudes for HGLM3. All estimates were very close to the hypothesized true value of 1.0. Note that we are not reporting the variance estimate for HGLM3 when $\sigma^2_u = 0$, because the Laplace estimations were not obtained in this simulation condition.

**Conclusion**

When LID is present, items within the same item cluster are more similar to each other in terms of each examinee’s item response patterns than with items in other item clusters. When the magnitude of LID becomes larger, similarity among items within the same item cluster will be stronger, and differences between items from different item clusters will be relatively greater. Accordingly, the interaction effect between person and item cluster will be stronger. This study demonstrated that the proposed three-level HGLM could capture the magnitude of LID in the simulated data sets. Also, this study demonstrated that the proposed three-level HGLM model produced smaller mean absolute differences from the hypothesized true values for item difficulty estimates than did the two-level HGLM that ignored item clusters. Furthermore, this study demonstrated that the estimate of the ability distribution variance was unaffected by the magnitude of LID with the proposed model, while it was shrunken as the LID magnitude became larger with a model that ignored item clusters and associated LID. This provides a particularly important observation, because it implies that we may under-differentiate examinees’ abilities or the estimates of examinees’ ability cluster around a smaller range than they should if we ignore a clustered item structure that is associated with LID. This study indicated that such a negative impact can be potentially avoided by the proposed 3-level HGLM approach.

Overall, this study was able to show the potential utility of the proposed three-level HGLM to model LID. However, this is only an initial step in the attempt to explore the utility of the HGLM to model LID. In this current study, only one data set was sampled for each simulation condition. Therefore, we were not able to evaluate the estimation error of parameters. Multiple data sets should be sampled to evaluate the accuracy and precision of parameter estimates. Also, this study assumed only one condition, in terms of the number of item clusters and the number of
items in each item cluster. However, we speculate that the number of item clusters and the number of items in each item cluster will affect the estimation accuracy and precision of parameters, especially for the level-2 variance component (i.e., LID magnitude). A systematic investigation of these factors is expected to be conducted in the future.

This current study illustrates the simplest scenario of the multilevel modeling of testlet effects. This model can be extended to include additional random effects or predictors. For example, group variables can be included into the model to perform DIF analyses. A higher level of hierarchy variables could also be included in the model by combining this simplest testlet multilevel model and Kamata’s three-level model as described in Figure 1 to examine additional effects from classrooms, schools, and states.

References


