Selecting Testlet Features With Predictive Value for the Testlet Effect

Paap, Muirne C. S.; He, Qiwei; Veldkamp, Bernard P.

Published in:
SAGE Open

DOI:
10.1177/2158244015581860

SAGE Open
April-June 2015: 1–12
© The Author(s) 2015
DOI: 10.1177/2158244015581860
sgo.sagepub.com
SAGE

Selecting Testlet Features With Predictive Value for the Testlet Effect: An Empirical Study

Muirne C. S. Paap[1], Qiwei He[1], and Bernard P. Veldkamp[1]

Abstract

High-stakes tests often consist of sets of questions (i.e., items) grouped around a common stimulus. Such groupings of items are often called *testlets*. A basic assumption of item response theory (IRT), the mathematical model commonly used in the analysis of test data, is that individual items are independent of one another. The potential dependency among items within a testlet is often ignored in practice. In this study, a technique called tree-based regression (TBR) was applied to identify key features of stimuli that could properly predict the dependence structure of testlet data for the Analytical Reasoning section of a high-stakes test. Relevant features identified included Percentage of "If" Clauses, Number of Entities, Theme/Topic, and Predicate Propositional Density; the testlet effect was smallest for stimuli that contained 31% or fewer "if" clauses, contained 9.8% or fewer verbs, and had Media or Animals as the main theme. This study illustrates the merits of TBR in the analysis of test data.

Keywords

tree-based regression, testlet response models, item response theory, high-stakes testing

Introduction

Item response theory (IRT) models have been increasing in popularity in many different fields (e.g., Behrend, Sharek, Meade, & Wiebe, 2011; de Jong, Steenkamp, & Veldkamp, 2009; Egberink, Meijer, & Veldkamp, 2010; Goetz et al., 2011; Jette et al., 2012; Liu, Hedeker, & Mermelstein, 2013; Mokkink, Knol, van Nispen, & Kramer, 2010; Paap et al., 2011; Reise & Waller, 2009; Stump, Husman, & Brem, 2012). They were first introduced in the field of educational measurement, and in this field, IRT is now the standard approach to analyze test data. Several assumptions underlie IRT models: unidimensionality, a specific parametric shape of the Item Characteristic Curve (ICC), and local independence (LID). LID implies that the relationship between the item responses is solely a function of the latent trait. This assumption does not receive as much attention in applied studies as do the other assumptions. Often researchers merely assume that this assumption holds, even though there is a significant body of literature describing how to detect and estimate the degree of local dependence (LD) (e.g., Chen & Thissen, 1997; Douglas, Kim, Habing, & Gao, 1998; Ip, 2001; Rosenbaum, 1984; Stout et al., 1996). The reason researchers do this is clear: Assuming LID holds allows the use of straightforward IRT analyses. If one wants to explicitly model the breach of LID, more complex IRT models are needed.

Yet in many practical applications where LID is assumed, it may be violated. One important example is when several items are grouped around common stimuli, such as text passages, tables, graphs, movie fragments, or other pieces of information. Such groups of items are generally referred to as item sets or testlets (Wainer & Kiely, 1987), and the dependence among the items in an item set has been referred to as *passage dependence* (Yen, 1993). The use of testlets is quite common in large-scale testing programs in educational measurement. Several explanations for testlet use have been offered, such as time efficiency (Wainer, Bradlow, & Du, 2000) and cost constraints (Bradlow, Wainer, & Wang, 1999). Ignoring the common stimulus that contextualizes the items will violate the assumption of LID. Some examinees might misread the stimulus, might not like the topic, might have a particular expertise on the subject covered in the stimulus, and so on. Ignoring this breach of LID can lead to overestimates of measurement precision and underestimates

[1]University of Twente, Enschede, The Netherlands

Corresponding Author:
Muirne C. S. Paap, Faculty of Behavioural, Management, & Social Sciences, Department of Research Methodology, Measurement, and Data-Analysis, University of Twente, Building Cubicus, Drienerlolaan 5, 7522NB Enschede, The Netherlands.
Email: m.c.s.paap@utwente.nl

of the standard error of the ability estimates, as well as misestimation of item parameters (Wainer & Wang, 2000; Yen, 1993).

Wainer et al. (2000) illustrate the effects of violation of LID by comparing the measurement precision of a one-item test and a 20-item test consisting of the same item administered 20 times. Not taking the dependency between the items into account would result in the same unbiased estimate of ability but in a standard error of the proficiency estimate four times smaller. LID is not so much related to the estimates themselves but to their precision. For example, violation of LID might not result in a different ordering of the ability estimates of a group of candidates, but it will affect estimates of the precision with which the ability levels have been measured.

Several approaches for handling LD have been proposed over the years (see Tuerlinckx & De Boeck, 2004). Bradlow et al. (1999) were among the first in proposing to model the testlet effect explicitly by extending the basic IRT model and introducing a new parameter to IRT models that accounts for the random effect within persons across items that belong to the same testlet. This testlet parameter $\gamma_{nt(i)}$, where *n, t*, and *i* are indices of respondents, testlets, and items, respectively, represents a random effect that exerts its influence through its variance (its sum over examinees within any testlet is zero); the larger the variance $\sigma^2_{t(i)}$, the larger the amount of LD among the items *i* within the testlet *t* (Wainer & Wang, 2000) and the smaller the precision of the parameter estimates. More specifically, it has been shown that ignoring the testlet effect if the value for σ^2_t is close to 1.00 leads to bias in the estimation of the discrimination and difficulty parameters (Glas, Wainer, & Bradlow, 2000). When the dependence is not properly modeled, the amount of information in the test might be overestimated.

Several procedures for estimating testlet response models have been developed and applications of testlet response theory (TRT) studied (Glas et al., 2000; Wainer, Bradlow, & Wang, 2007). In the three-parameter normal ogive (3PNO) testlet model, the probability of a correct response, denoted by $Y_{ni} = 1$, is given by

$$P\left(Y_{ni}=1\right)=P_{ni}=c_i+\left(1-c_i\right)\Phi\left(\tau_{ni}\right), \tag{1}$$

where f is the cumulative standard normal distribution, that is

$$\Phi\left(s\right)=\left(2\pi\right)^{-1/2}\int_{-\infty}^{s}\exp\left(-\frac{t^2}{2}\right)dt, \tag{2}$$

and

$$\tau_{ni}=a_i\theta_n-b_i+\gamma_{nt(i)}. \tag{3}$$

The discrimination parameter of item *i* is denoted by a_i, its difficulty parameter by b_i, c_i denotes the guessing parameter of item *i*, $\gamma_{nt(i)}$ is the random testlet effect for person *n* on testlet *t(i)*, and τ_{ni} is the probit of P_{ni}. If the guessing parameters are set to zero, the model given by Equation 1 turns into the 2PNO testlet response model, and if, in addition, the discrimination parameters are set equal to one, the model turns into the 1PNO testlet response model.

A question left unaddressed until relatively recently is whether features of the testlets can help *predict* the testlet effect. A potential explanation could be that the testlet effect is usually seen as a nuisance parameter. After all, if one can eliminate the "noise" caused by the testlet effect from the model by using a testlet parameter, why would one need to identify the underlying mechanism? The answer to this question is straightforward: because in daily practice, many high-stakes tests are equated with "regular" IRT models, even though the test in question might contain testlets. The overall aim of this study is to demonstrate how to predict the size of the testlet effect using testlet features; we expect this approach would be especially relevant for test designers.

To introduce covariates (testlet features) to the testlet IRT model, Wainer et al. (2007) proposed introducing covariates into the testlet part of the IRT model by using a log-normal prior

$$\log\left(\sigma^2_{t(i)}\right)\sim N\left(X'_{\gamma t(i)}\beta_\gamma,\lambda^2\right), \tag{4}$$

where $\sigma^2_{t(i)}$ indicates the strength of the testlet effect (with a larger value indicating a greater proportion of the total variance in test scores that is attributable to the given testlet), X' is a vector of covariates describing the testlet, and β_γ constitutes a set of covariate slopes. Covariates of interest in their report included the number of words in the testlet, the subject area of the testlet, the number of items in the testlet, or the location of the testlet in the overall test.

We will use an alternative approach. The key to this approach is that the testlet parameter is predicted from testlet features using a covariate model that is incorporated into the testlet model (Glas, 2012). In the present study, we will demonstrate the usefulness of tree-based regression (TBR) to select variables that significantly contribute to the prediction of the testlet effect. In a next phase of this study, the performance of the model proposed by Glas (2012) will be tested by incorporating the variables selected in this study into his model as covariates.

TBR has been used to model item difficulty estimated with IRT models, in an attempt to integrate elements from cognitive psychology and assessment; in this approach, it is assumed that the ability to answer an item involves one or several cognitive components (Gao & Rogers, 2011). In the present study, we will use TBR to identify those testlet features that best predict the testlet effect. TBR has several advantages over more

traditional methods such as linear regression analysis, among which are (a) fewer statistical assumptions because it is a nonparametric technique, (b) ease of interpretation by tracing the splitting rules down the branches of the tree, (c) optimizing the usage of categorical independent variables by merging redundant categories, (d) invariance to monotone transformations of independent variables, (e) ease of dealing with complex interactions, and (f) the ability to handle missing data (Gao & Rogers, 2011; Su et al., 2011). TBR has also been shown to outperform linear regression analysis with respect to prediction precision (Finch et al., 2011).

In summary, the aim of the present study is to illustrate how TBR can be used to identify relevant testlet features that help predict LD between items belonging to the same testlet. Our approach consists of three steps:

1. Obtaining testlet features using both text mining and manual scoring
2. Using TBR to select those features that best predict the variance of the testlet effect
3. Assessing whether different results are obtained when the two-parameter normal ogive (2PNO) model is used instead of the 3PNO model

Method

Stimuli

The responses of 49,256 respondents to 594 items nested within 100 total testlets (stimuli) administered on the Analytical Reasoning section of the Law School Admission Test were used. The items are designed to test the ability of the examinee to reason within a given set of circumstances. These circumstances are described in the stimulus (testlet-specific text passage). The stimulus contains information about a number of elements (people, places, objects, tasks, and so on) along with a set of conditions imposing a structure on the elements (e.g., ordering them, assigning elements of one set to elements of another set, and so on). The stimuli always permit more than one acceptable outcome satisfying all of the requirements in the stimulus text. Each student was presented with up to four stimuli.

Scoring Testlet Features

Testlet features can be extracted from the stimulus either manually or automatically. In this study, features are selected based partly on previous research (Drum, Calfee, & Cook, 1981; Embretson & Wetzel, 1987; Gorin & Embretson, 2006) and partly on new ideas. The first and last author of this article, separately, carried out manual feature extraction. For automated feature selection, a text-mining algorithm was applied.

The testlet feature variables used in this study can be divided into three categories: (a) variables describing the logical structure of the stimuli, (b) variables describing the themes contained in the stimuli, and (c) surface linguistic variables. Two raters (the first and last author of this article) independently coded the variables in Categories 1 and 2. In the case of incongruent scorings, a consensus was reached through a thorough discussion. A discussion log was kept for these stimuli. The surface linguistic features were generated by the second author using the software Python (Python Software Foundation, 2009).

Manually coded features. The following variables described the structure of the stimuli: Number of Features, Stimulus Type, Number of Entities, Number of Positions, Cardinality of Entities, Cardinality of Positions, Number of Entities Smaller/Larger Than Number of Positions, and Ordered Positions (Yes/No/Partially). The description of the variables can be found in Table 1. The following stimulus is an example of a testlet involving ordering entities:

> On one afternoon, an accountant will meet individually with each of exactly five clients—Reilly, Sanchez, Tang, Upton, and Yansky—and will also go to the gym alone for a workout. The accountant's workout and five meetings will each start at either 1:00, 2:00, 3:00, 4:00, 5:00, or 6:00.
>
> The following conditions must apply:
>
> The meeting with Sanchez is earlier than the workout.
>
> The workout is earlier than the meeting with Tang.
>
> The meeting with Yansky is either immediately before or immediately after the workout.
>
> The meeting with Upton is earlier than the meeting with Reilly.

In this example, there are two features: the entity variable "*x*" (i.e., the six appointments) and the position variable "*y*" (i.e., the six positions in the schedule). Both the Cardinality of Entities and the Cardinality of Positions is equal to "1" because the appointments can only be assigned once to a position in the schedule, and only one single appointment can be assigned to a position in the schedule.

One variable was used to describe the main theme of the stimulus. The following categories were used: B (Business), E (Education), R (Recreation), M (Media), A (Animals), T (Transport/Vehicle), N (Nature), and H (Health). In the example presented in the paragraph above, the theme would be scored as "B" (Business). Ideally, the theme of the stimulus would be retrieved from the description of the testlet contained in the item bank, assuming that a content classification was assigned to it when the stimulus was designed. In this case, however, such a classification was not available. Therefore, each stimulus was assigned to a category by hand.

Text mining. Text mining is a form of data mining; data mining is, as the name suggests, a data-driven approach. To apply data-mining techniques for finding critical features

Table 1. Overview of Independent Variables Used in the Regression Analyses, Divided in Three Categories.

Independent variables	Description/remarks
Structural variables	
Number of Features	Takes values 2, 3, or 4; 2 if it only contains information on one type of entity variable ("*x*") and one position variable ("*y*"), 3 if it contains information about two entity variables and one position variable or one entity variable and two position variables, and 4 if it contains information about two entity variables and two position variables.
Stimulus Type	Only scored if the number of features is 3 or 4; it specifies whether the stimulus is of type 1 (two or more *x*'s were assigned to one y), 2 (one *x* was assigned to two or more *y*'s) or 3 (*x* was assigned to *y*, which was assigned to a higher-order position variable "*z*").
Number of Entities	The number of entities, summed over all entity variables present in the stimulus; entities are defined as the units in the stimulus that had to be assigned to positions.
Number of Positions	The number of positions, summed over all position variables present in the stimulus.
Cardinality of Entities	Takes values "1" or "multiple." The cardinality of entities is "1" if they can only be assigned to a position once and multiple if they can be assigned more than once.
Cardinality of positions	Takes values "1" or "multiple." The cardinality of positions is "1" if only one entity can be assigned to a position and "multiple" if more than one entity can be assigned to a position.
Number of Entities Smaller/Larger Than Number of Positions	
Ordered Positions (Yes/No/Partially)	
Theme variable	
Theme/Topic	Variable used to describe the main theme of the stimulus. The following categories are used: B (Business), E (Education), R (Recreation), M (Media), A (Animals), T (Transport/Vehicle), N (Nature), P (Intrapersonal Relationships/Family), and H (Health).
Surface linguistic variables	
Word Token[a]	Length of the stimulus text, total number of words excluding punctuation.
Word Type[a]	Vocabulary size, total number of words excluding word repetition and punctuation.
Word Diversity	Word type divided by word token.
Average Characters[a]	Average number of letters used per word in the stimulus text.
Percentage of Negative Words	Percentage of "negative" words such as "no," "not," "neither," and so on. May increase the difficulty of a text.
Brown News Popularity	The popularity of verbs, nouns, adjectives, adverbs, and names in the Brown News Corpus. Note that the Brown News Corpus is often used as a reference database in natural language processing. It contains 100,554 words in total, of which 14,394 are unique. To calculate the Brown News Popularity variable, the Porter Stemmer algorithm was used to standardize each word.
Percentage of Content Words[a]	The number of verbs, nouns, adjectives, adverbs, and names divided by word token.
Modifier Propositional Density[a]	Number of adjectives divided by word token.
Predicate Propositional Density[a]	Number of verbs divided by word token.
Number of Sentences[a]	Number of sentences used in stimulus text.
Average Sentence Length[a]	Word token divided by number of sentences.
Percentage of "If" Clauses	In the AnalyticalReasoning stimuli, "if" clauses are regularly used and could be expected to increase the difficulty of a text (both with respect to logical reasoning and sentence complexity).

[a]Based on work by Drum, Calfee, and Cook (1981); Embretson and Wetzel (1987); and Gorin and Embretson (2006).

that can predict the magnitude of the testlet effect (i.e., variance of the testlet parameter), data from real tests, where testlet effects are present, have to be analyzed. Text mining is especially suitable for analyzing verbal stimuli, such as text passages in a standardized test. The aim of text mining is to extract information from a piece of text using an automated procedure, which is usually followed by applying a statistical method to select the text features that are most discriminating when predicting the dependent variable.

In text mining, the raw ("unstructured") text is first structured. A common first step is to reduce words in the stimulus (text passage) to their stems (e.g., the words "sleepy" and "sleeping" would both be reduced to "sleep"). A second step is to filter out words that are thought not to be relevant to the analysis, such as "and," "to," and so forth. After this text structuring is performed, statistical methods are used to identify patterns in the structured data. The methods used depend on the purpose of the analysis.

The surface linguistic variables that were generated can be found in Table 1. Note that to calculate the Brown News Popularity variable, the Porter Stemmer algorithm was used to standardize each word; this algorithm was not used during the construction of the other variables.

Statistical Analysis

First, the testlet response model was estimated, to obtain the input variables for the TBR. Both the 2PNO testlet model and the 3PNO testlet model were estimated, so we could evaluate whether different results are obtained when 2PNO testlet model estimates are used instead of the 3PNO testlet model estimates. A fully Bayesian approach using a Markov chain Monte Carlo (MCMC) computation method was applied. The software package MIRT (Glas, 2010) was used.

Tree-based regression (TBR). Before describing our model building strategy, we will first provide the reader with a short introduction to TBR.

Introduction to TBR. There are two categories of trees: classification or decision trees (dependent variable is categorical) and regression trees (dependent variable is continuous). In this study, we focus on regression trees, which we refer to as TBR. TBR could be seen as a special case of regression modeling, where an underlying regression function is approximated by splitting the "predictor space" recursively into disjoint regions and subsequently fitting constant models to each region; this results in a piecewise-constant approximation to the underlying regression function for the dependent variable (Su et al., 2011). Tree-based methods have gained popularity after Breiman, Friedman, Olshen, and Stone (1984) introduced the Classification and Regression Trees (CART) methodology. CART provides a powerful framework that provides highly interpretable presentations of (nonlinear) relationships between variables and adequately addresses practical issues, such as tree size selection. This CART methodology is still considered the current standard for tree modeling. TBR is a good alternative for least squares (linear) regression if the studied relationships are nonlinear because TBR has been shown to have a higher prediction precision in those situations (Breiman et al., 1984; Finch et al., 2011). Furthermore, it has been shown to give clues to data structure that may not be apparent in a linear regression analysis.

Tree-building: Splits and nodes. The C&RT module in SPSS (SPSS, 2007b) closely follows the algorithm called CARTs described by Breiman et al. (1984). Using TBR, clusters are formed by successively splitting the data set based on the value of a predictor variable into increasingly homogeneous subsets ("nodes"). The predictor variable that maximizes the homogeneity with respect to the dependent variable in the two nodes is identified and selected at each stage of the algorithm. The split that maximizes the difference in deviance between the parent node (original set of items) and the sum of the child nodes (subsets of items created by the independent variable) results in a low value for the impurity measure. The impurity measure *R(t)* is measured by the prediction error in node *t*:

$$R(t) = \frac{1}{N}\sum_{i\in t}\left(y_i - \bar{y}(t)\right)^2, \tag{5}$$

where y_i are the observed values of the dependent variable and $\bar{y}(t)$ is the mean value of the dependent variable in the node *t*. The impurity of the tree is given by the sum of the impurity measures of all terminal nodes (nodes that are not split further). In SPSS, this value is reported as the "risk estimate" (within node variance divided by total variance). To indicate the fit of a tree, the proportion of explained variance can be calculated by subtracting the risk estimate from one.

Initially, a large tree that overfits the data is grown, to avoid missing important structures. In this large initial tree, the true patterns are mixed with numerous spurious splits that are then removed via pruning. Through pruning the large tree, a nested sequence of subtrees are obtained; subsequently, a subtree of optimal size is selected from the sequence. Pruning entails collapsing pairs of child nodes with common parents by removing a split at the bottom of the tree.

Pruning and choosing the best tree size. Following Matteucci, Mignani, and Veldkamp (2012), the rule of one standard error (*SE*) was adopted to choose the best tree size. According to this rule, the residual variance was evaluated for all levels of pruning, and the tree with a difference in residual variance less than one *SE* between the pruned tree and the subtree with the smallest residual variance was considered the best tree. SPSS (SPSS, 2007b) was used to conduct the TBR analyses. A number of stopping rules can be used to end a TBR analysis. A minimum change of improvement smaller than 0.000001 was used as a stopping rule. The change of improvement equals the decrease in impurity required to split a node; for continuous dependent variables, the impurity is computed as the within-node variance, adjusted for any frequency weights or influence values (SPSS, 2007a). Larger values for the change of improvement tend to result in smaller trees. Also, the maximum tree depth was set to 10 levels, and the minimum number of cases was set to five for parent nodes and three for child nodes.

Typically, *v*-fold cross-validation is applied to further assess the quality of the final model (i.e., tree), but as we had a small data set (100 cases), *v*-fold cross validation resulted in trees with little explained variance and little stability (large effect of the random splitting of the data set). Hence, we decided not to use cross-validation in this study.

Applying TBR. The standard deviation (*SD*) of the testlet parameter, which we will denote as σ_{1t}, was used as a dependent variable. We chose to use σ_{1t} and not σ_{1t}^2 in our

model because σ_{1t} capitalizes on the difference between testlets and is thus more informative in this setting. The *SD* of the testlet parameter was estimated using the testlet response model with no covariates.

The independent variables used in model building can be found in Table 1. Model building was done as follows. First, separate models were evaluated for each cluster of items (structure, theme, linguistic). The variables that were selected by the algorithm were then retained per cluster, and subsequently, one of the other clusters was added to the selected variables to see whether any of those were selected in the regression tree. For example, say we would enter all structure variables into the model, and Variables 1 and 2 in the cluster "structure" would be selected by the algorithm to be part of the regression tree; our next step would be to remove the other (not selected) structure variables from the analysis and add the linguistic variables to see if any of those variables would end up in the tree. We would then remove the variables that were not selected from the independent variable list and enter the theme variable to see whether that one would be selected. In case of competing models, the final model would be the model with the greatest number of splits resulting in a large difference in the mean testlet *SD* for the resulting nodes. To see whether any difference would emerge, the analysis was performed for σ_{1t} as estimated by the 3PNO testlet model and then by the 2PNO testlet model (with guessing parameter set to zero).

Results

When the σ_{1t} parameter was estimated using the 3PNO testlet model, it had a mean of 0.71 (*SD* = 0.16), and when it was estimated using the 2PNO testlet model, it had a mean of 0.50 (*SD* = 0.12). Inspection of the *SE*s of σ_{1t} for each testlet revealed that these were generally larger for the 3PNO (median = 0.060, interquartile range = 0.036-0.081) model than for the 2PNO (median = 0.027, interquartile range = 0.020-0.033) model. This is not entirely surprising because a greater number of parameters are estimated using the 3PNO testlet model, which is typically accompanied by a decrease in measurement precision.

The Two Regression Trees

Figures 1 and 2 show the final models resulting from the TBR analyses. Figure 1 depicts the model using the testlet effect estimated by the 2PNO testlet model as a dependent variable, and for Figure 2, the dependent variable was estimated using the 3PNO testlet model. The figures show which predictors are associated with the testlet effect. Furthermore, they show which cut-offs (in case of continuous variables) and categories (in case of nominal variables) resulted in the highest homogeneity that could be achieved in terms of the dependent variable. For example, by inspecting the means presented in the different nodes in Figure 1, it can be seen that the lowest mean can be found in Node 6 (0.381). In other words, the testlet effect was smallest for stimuli that contained more than 31% "if" clauses but did not have an educational theme. To aid interpretation, the nodes are printed in blue for $\sigma_{1t} < 0.5$, in yellow if $0.5 < \sigma_{1t} < 0.75$, and in red if $\sigma_{1t} > 0.75$.

The two final trees both contained the independent variables Percentage of "If" Clauses, Number of Entities, and Theme/Topic. However, the 2PNO-based model also contained the variable Ordered Positions, whereas the 3PNO-based model contained the variable Predicate Propositional Density. Note that the distribution of Percentage of "If" Clauses was highly skewed, with 63% of the stimuli containing zero "if" clauses. The Number of Entities was slightly positively skewed with a mean of 7.33 (*SD* = 3.28). Of all themes, Business (29%), Education (19%), and Recreation (21%) were most common, whereas Health (2%), Media (3%), and Nature (3%) were least common. Sixty percent of the stimuli contained Ordered Positions, 35% did not, and the remaining 5% contained only Partially Ordered Positions. Predicate Propositional Density was normally distributed with a mean of 7.6% (*SD* = 2.1%). The 2PNO-based model contained 12 nodes, and for this model, the explained variance equaled 33.11%; the 3PNO-based model contained 16 nodes, and the explained variance equaled 37.5%.

Prediction accuracy. As the dependent variables in the TBR models were in fact estimated in another model (vs. fixed observed traits of the stimuli), they are associated with a certain amount of measurement error. This measurement error is not accounted for in the TBR models. As can be seen in Figure 3, larger testlet effects were associated with higher standard errors, and the 3PNO estimates were estimated with less precision than the 2PNO estimates. To investigate whether this impacted the quality of the TBR predictions, we plotted the standard errors of the testlet effects against the absolute prediction errors of the TBR models (Figure 4).

The absolute prediction error of the TBR model, $\left|(y_i - \hat{y}_i)\right|$, had a mean value of 0.07 (*SD* = 0.062) for 2PNO-based σ_{1t} estimates, and 0.10 (*SD* = 0.079) for 3PNO-based σ_{1t} estimates. The 2PNO-based σ_{1t} estimates showed a modest correlation with their *SE*s (.31). The *SE*s and absolute prediction errors showed a very weak correlation (.05). The 3PNO-based σ_{1t} estimates showed a medium correlation with their *SE*s (.47). The *SE*s and absolute prediction errors also showed a weak correlation (.20). In conclusion, we did not find evidence for a clear relationship between the measurement error associated with the testlet effects and the prediction error of the TBR models.

Discussion

In this study, we demonstrated the advantages of TBR to identify features that can predict the testlet effect. Our approach was prompted by what is common practice; many

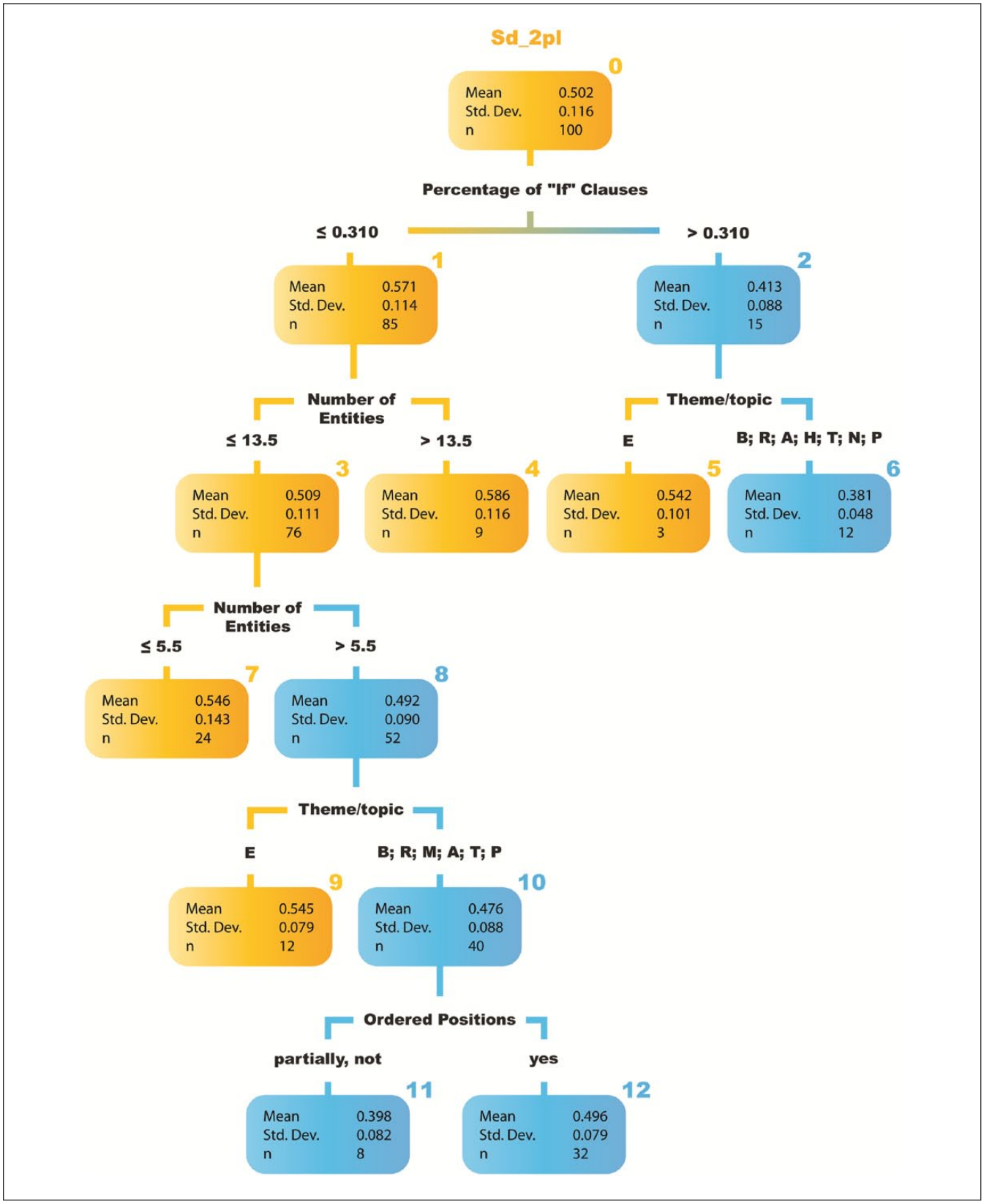

Figure 1. Regression tree for the 2PNO-based testlet effect.
Note. To aid interpretation, the nodes are printed in blue for $\sigma_{1t} < 0.5$, in yellow if $0.5 < \sigma_{1t} < 0.75$, and in red if $\sigma_{1t} > 0.75$. PNO = parameter normal ogive.

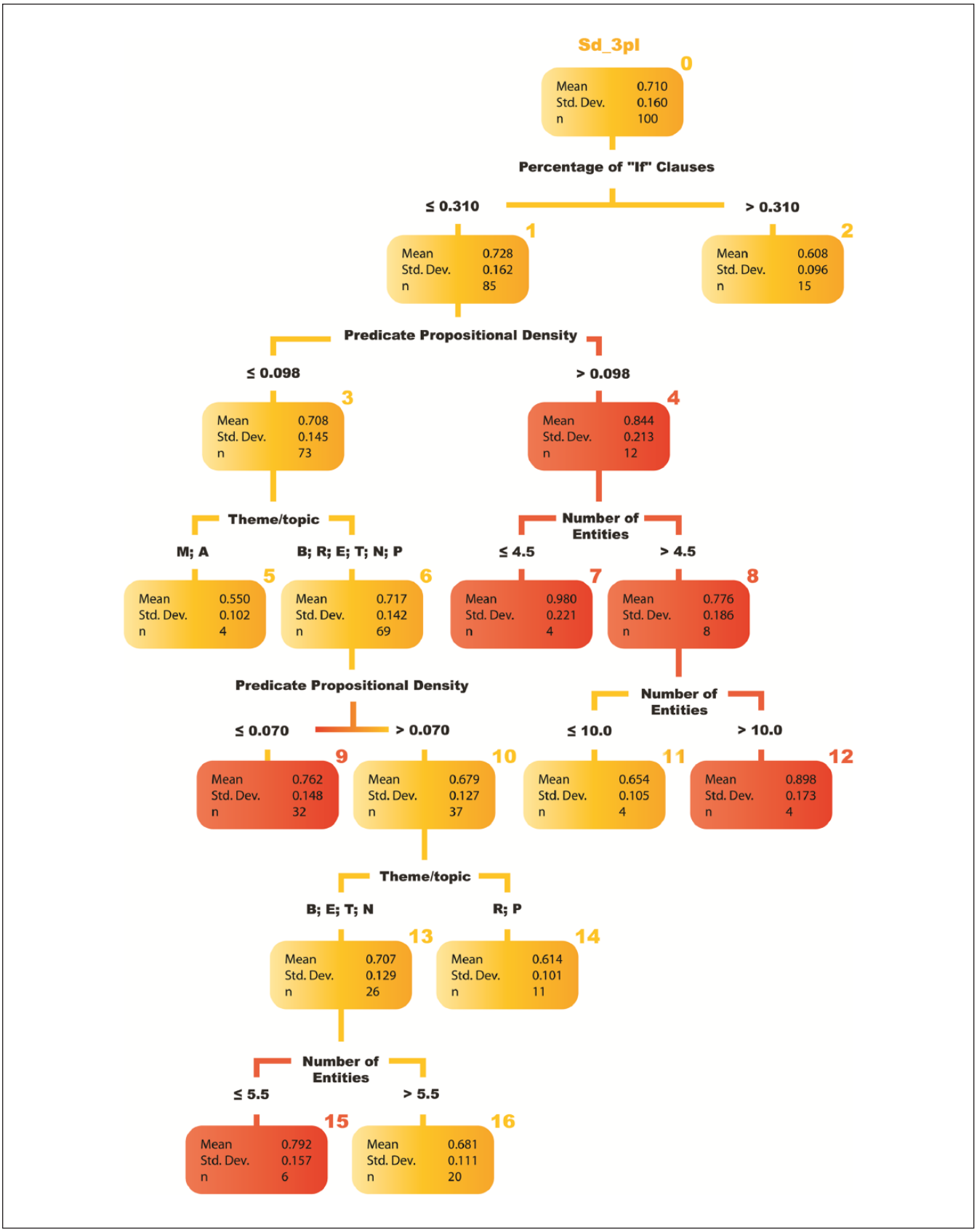

Figure 2. Regression tree for the 3PNO-based testlet effect.
Note. To aid interpretation, the nodes are printed in blue for $\sigma_{1t} < 0.5$, in yellow if $0.5 < \sigma_{1t} < 0.75$, and in red if $\sigma_{1t} > 0.75$. PNO = parameter normal ogive.

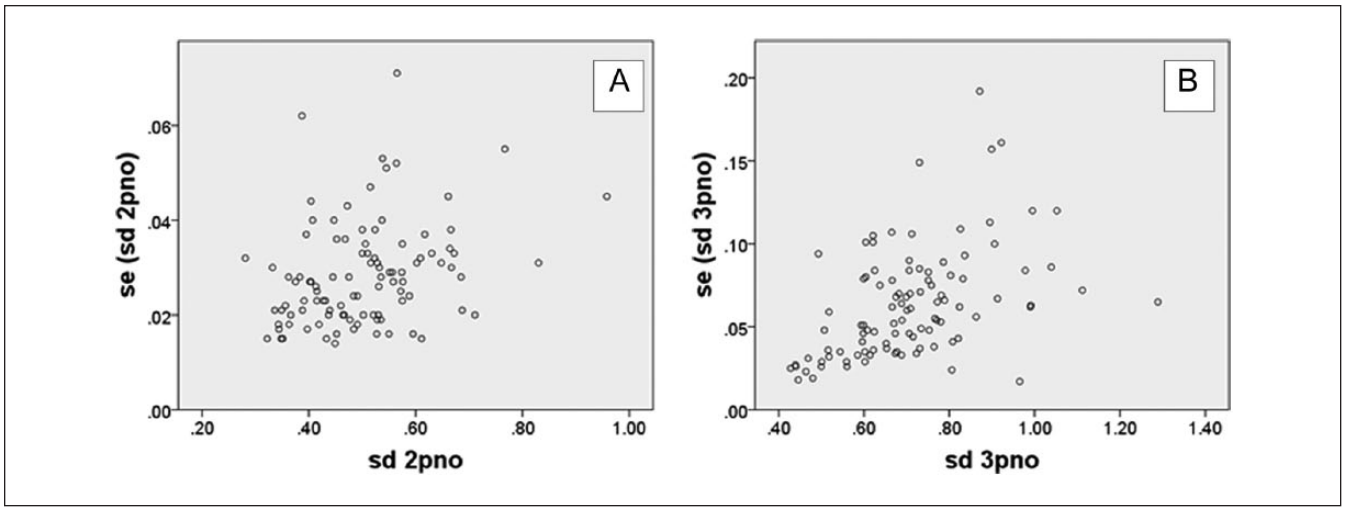

Figure 3. Graphical representation of the relationship between the testlet effects as estimated by the 2PNO model (Panel A) and 3PNO model (Panel B), and their respective standard errors.
Note. PNO = parameter normal ogive.

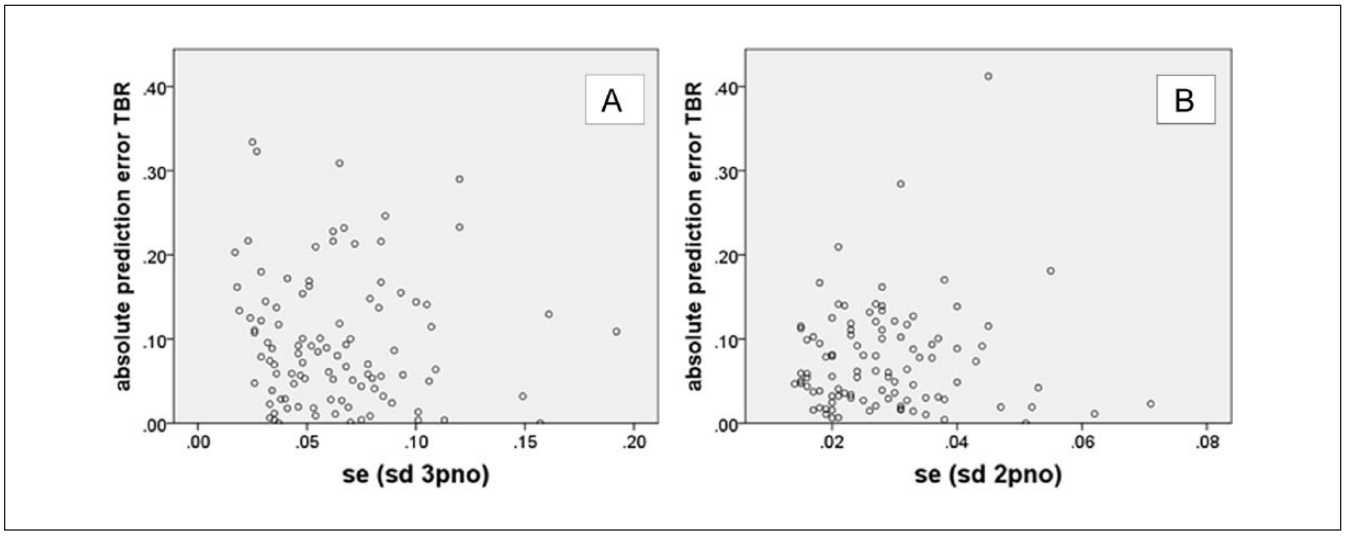

Figure 4. Graphical representation of the relationship between the standard error associated with the testlet effects as estimated by the 2PNO model (Panel A) and 3PNO model (Panel B) and the absolute prediction error of the TBR model.
Note. PNO = parameter normal ogive; TBR = tree-based regression.

high-stakes tests are equated with "regular" IRT models, even though the test in question might contain testlets. This is also true for the test we used as an example in this study: It is calibrated with a 3-parameter IRT model (e.g., Lord, 1980). Scores reported for the test are based on application of this model as well. This 3-parameter IRT model can be seen as a special case of the testlet model presented in this article, where the testlet effect is equal to zero. By applying the 3-parameter IRT model, it is assumed that no violation of LID occurs. In practice, this assumption is often violated (e.g., Wainer et al., 2000). Passage dependence is a common cause of LD, which also affects the test we used in this study. In several subsections of the test, items are grouped around a common stimulus. To minimize violation of LID and not to be too far off by applying the 3-parameter IRT model, stimuli with a small testlet effect are favored during test assembly. It has been shown that a within-person variance $\left(\sigma^2_{1t}\right)$ of 0.25 or smaller has a negligible effect on the estimation of the discrimination and difficulty parameters (Glas et al., 2000).

In this study, we sought to address the testlet issue from a practical angle, showing how our approach can be used to identify and model features that give an indication of the magnitude of the testlet effect. Our approach has direct practical relevance: It can be used in the process of designing

new testlets, where stimuli with certain features (indicating little LD) would be favored over others (with higher LD) to reduce the risks of mis-specifying the IRT model. The results indicated that the testlet effect was related to both linguistic and structural features of the stimulus as well as to its theme. This might indicate that it could be useful for test developers to carefully consider the features of the stimuli they are designing, if they intend to keep the testlet effect as low as possible. However, our results also indicated that the guidelines resulting from our analyses would depend on the chosen measurement model. Because the 2- and 3-parameter models are both frequently used in educational measurement and assessment settings, it is important that future research addresses this discrepancy. In the application described in this article, we favor the 3PNO solution for several reasons. First, enough data were available to estimate the 3PNO reasonably well; second, the IRT model currently used in calibrating the test we used as an example is also a 3-parameter model; and finally, both the 3PNO-based regression-tree model showed a higher explained variance than its 2PNO counterpart. However, a 2PNO solution might be favored in other settings, especially when estimating the 3PNO would lead to unreasonably high *SE*s (i.e., in smaller data sets). The percentages of explained variance found in our study are relatively low. In a future study, we would like to explore whether we can find a way to increase this number, for example, by using a larger data set.

It should be noted that the standard deviation parameter estimates used as dependent variables in the TBR involve estimation error; in a fully Bayesian approach, like the one we applied, the parameter estimates are shrunken toward the population mean. This implies that our estimates are likely to be on the conservative side. Although this could be considered a limitation, we reason that having conservative estimates is less problematic than overestimating the effects.

As our analyses were based on only 100 cases, the results might not be generalizable to other data sets. In other studies that had a different unit of measurement (item instead of stimulus), *v*-fold cross-validation was used to assess the generalizability of the TBR model (e.g., Sheehan, Kostin, & Futagi, 2007). This was not feasible in our study because using this type of validation would have resulted in our data set becoming too small to fit a TBR model at all. We do not think this is a major drawback, however, because it was not the primary aim of this study to produce a model that would be generalizable to other (sub)tests; we aimed to illustrate how our approach can be used in practice. The quality of the model was established by using pruning to avoid over-fitting the data and inspecting the relationship between prediction accuracy of the TBR model and the *SE*s produced by the testlet models.

Inspection of the prediction accuracy showed a modest relationship between the IRT estimates of the testlet effect and their corresponding *SE*s. More specifically, we found that larger testlet effects were estimated less reliably than smaller testlet effects. However, this did not affect the prediction accuracy of the TBR models; we found weak relationships between the prediction error on one hand, and the size of the testlet effect and the corresponding *SE*s on the other hand. In addition, we ensured satisfactory content validity of our findings by discussing them with a test design expert.

For some of our findings, a theoretical explanation can be given. From a theoretical/conceptual standpoint, one can reason that a large testlet effect (i.e., strong relationship among the items pertaining to the same stimulus after correcting for the latent trait estimate) implies that the stimulus contains a lot or all information necessary to solve the associated items, or that solving the items depends on a special insight (you either "get it" or you don't). If-clauses add information over-and-above the information contained in the main stimulus text. If there are no if-clauses, then all information is contained in the main stimulus text and all items associated with the stimulus pertain to this main text (and thus the same solution set). Typically, an if-statement is key to unlock specific information; the items associated with the same stimulus may differ in the combination of if-clauses that are to be considered. This introduces an independency among the items that is not present in the absence of if-clauses. The greater the number of if-clauses, the greater the variation in solution sets over items, and the smaller the relation among the items (testlet effect). This explanation is in line with our finding that a large number of if-clauses is associated with a smaller testlet effect.

The effect of the number of entities is a bit more complex. We found that both a very small and very large number of entities are associated with an increase in the size of the testlet effect. This may seem contradictory at first, but there may be two different underlying processes at play. When there are only few entities, there is not a lot of room for deduction. Therefore, one usually needs a specific insight to solve items associated with these types of stimuli. However, if there are many entities involved, the test developer may need to include a number of restrictions in the stimulus to ensure that the resulting solution set is still of a manageable size. Therefore, comprehension of the stimulus becomes more important in solving the related items. Both implications were confirmed in TBR. For the other variables selected in TBR, it was less obvious how they were related to the testlet effects.

In conclusion, we showed the merit of TBR in identifying critical testlet features that are associated with the testlet effect, using an empirical data set. We expect that this approach may be helpful for test designers who want to identify the stimulus features that influence testlet effect size in their test and use this information to design testlets with a small testlet effect.

Acknowledgments

We thank Sonja-Vanessa Schmitz for her assistance in preparing Figures 1 and 2, Hanneke Geerlings for useful discussions, and Cees Glas for his feedback on earlier versions of the manuscript.

Authors' Note

The opinions and conclusions contained in this article are those of the author and do not necessarily reflect the policy and position of Law School Admission Council (LSAC).

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) disclosed receipt of the following financial support for the research and/or authorship of this article: This study received funding from the Law School Admission Council (LSAC).

References

Behrend, T., Sharek, D., Meade, A., & Wiebe, E. (2011). The viability of crowdsourcing for survey research. *Behavior Research Methods*, *43*, 800-813.

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

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth International Group.

Chen, W.-H., & Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, *22*, 265-289.

de Jong, M. G., Steenkamp, J.-B. E. M., & Veldkamp, B. P. (2009). A model for the construction of country-specific yet internationally comparable short-form marketing scales. *Marketing Science*, *28*, 674-689.

Douglas, J., Kim, H. R., Habing, B., & Gao, F. (1998). Investigating local dependence with conditional covariance functions. *Journal of Educational and Behavioral Statistics*, *23*, 129-151.

Drum, P. A., Calfee, R. C., & Cook, L. K. (1981). The effects of surface structure variables on performance in reading comprehension tests. *Reading Research Quarterly*, *16*, 486-514.

Egberink, I. J. L., Meijer, R. R., & Veldkamp, B. P. (2010). Conscientiousness in the workplace: Applying mixture IRT to investigate scalability and predictive validity. *Journal of Research in Personality*, *44*, 232-244.

Embretson, S. E., & Wetzel, C. D. (1987). Component latent trait models for paragraph comprehension tests. *Applied Psychological Measurement*, *11*, 175-193. doi:10.1177/014662168701100207

Finch, W. H., Chang, M., Davis, A. S., Holden, J. E., Rothlisberg, B. A., & McIntosh, D. E. (2011). The prediction of intelligence in preschool children using alternative models to regression. *Behavior Research Methods*, *43*, 942-952. doi:10.3758/s13428-011-0102-z

Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing*, *28*, 77-104. doi:10.1177/0265532210364380

Glas, C. A. W. (2010). *Preliminary manual of the software program Multidimensional Item Response Theory (MIRT)*. Enschede, The Netherlands: Department of Research Methodology, Measurement and Data-Analysis, University of Twente.

Glas, C. A. W. (2012). *Estimating and testing the extended testlet model*. Newtown, PA: LawSchool Admission Council.

Glas, C. A. W., Wainer, H., & Bradlow, E. T. (2000). MML and EAP estimation in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computer adaptive testing: Theory and practice* (pp. 271-288). Dordrecht, The Netherlands: Kluwer.

Goetz, C., Ecosse, E., Rat, A.-C., Pouchot, J., Coste, J. l., & Guillemin, F. (2011). Measurement properties of the osteoarthritis of knee and hip quality of life OAKHQOL questionnaire: An item response theory analysis. *Rheumatology*, *50*, 500-505.

Gorin, J. S., & Embretson, S. E. (2006). Item diffficulty modeling of paragraph comprehension items. *Applied Psychological Measurement*, *30*, 394-411. doi:10.1177/0146621606288554

Ip, E. (2001). Testing for local dependency in dichotomous and polytomous item response models. *Psychometrika*, *66*, 109-132. doi:10.1007/bf02295736

Jette, A. M., Tulsky, D. S., Ni, P., Kisala, P. A., Slavin, M. D., Dijkers, M. P., . . . Fyffe, D. (2012). Development and Initial Evaluation of the Spinal Cord Injury-Functional Index. *Archives of Physical Medicine and Rehabilitation*, *93*, 1733-1750.

Liu, L. C., Hedeker, D., & Mermelstein, R. J. (2013). Modeling nicotine dependence: An application of a longitudinal IRT model for the analysis of Adolescent Nicotine Dependence Syndrome Scale. *Nicotine & Tobacco Research*, *15*, 326-333.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Erlbaum.

Matteucci, M., Mignani, S., & Veldkamp, B. P. (2012). Prior distributions for item parameters in IRT models. *Communications in Statistics—Theory and Methods*, *41*, 2944-2958.

Mokkink, L. B., Knol, D. L., van Nispen, R. M. A., & Kramer, S. E. (2010). Improving the quality and applicability of the Dutch scales of the communication profile for the hearing impaired using item response theory. *Journal of Speech and Hearing Research*, *53*, 556-571.

Paap, M. C. S., Meijer, R. R., Van Bebber, J., Pedersen, G., Karterud, S., Hellem, F. M., & Haraldsen, I. R. (2011). A study of the dimensionality and measurement precision of the SCL-90-R using item response theory. *International Journal of Methods in Psychiatric Research*, *20*, e39-e55.

Python Software Foundation. (2009). Python Version 2.6.2. Available from https://http://www.python.org

Reise, S. P., & Waller, N. G. (2009). Item response theory and clinical measurement. *Annual Review of Clinical Psychology*, *5*, 27-48.

Rosenbaum, P. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. *Psychometrika*, *49*, 425-435. doi:10.1007/bf02306030

Sheehan, K. M., Kostin, I., & Futagi, Y. (2007). Predicting text difficulty via corpus-based dimensionality reduction and tree-based regression. In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society. Nashville*, TN.

SPSS Inc. (2007a). SPSS Classification Trees™ 16.0. Chicago, IL: SPSS Inc..

SPSS Inc. (2007b). SPSS for Windows, Rel. 16.0.1. Chicago, IL: SPSS Inc.

Stout, W., Habing, B., Douglas, J., Kim, H. R., Roussos, L., & Zhang, J. (1996). Conditional covariance-based nonparametric

multidimensionality assessment. *Applied Psychological Measurement*, *20*, 331-354. doi:10.1177/014662169602000403

Stump, G. S., Husman, J., & Brem, S. K. (2012). The Nursing Student Self-Efficacy Scale: Development using item response theory. *Nursing Research*, *61*, 149-158. doi:10.1097/NNR.0b013e318253a750

Su, X., Azuero, A., Cho, J., Kvale, E., Meneses, K. M., & McNees, M. P. (2011). An introduction to tree-structured modeling with application to quality of life data. *Nursing Research*, *60*, 247-255. doi:10.1097/NNR.0b013e318221f9bc

Tuerlinckx, F., & De Boeck, P. (2004). Models for residual dependencies. In P. De Boeck & M. Wilson (Eds.), *Explanatory item response models: A generalized and nonlinear approach* (pp. 289-316). New York, NY: Springer.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In W. J. van der Linden & C. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practise* (pp. 245-269). Dordrecht, The Netherlands: Kluwer.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York, NY: CambridgeUniversity Press.

Wainer, H., & Kiely, G. (1987). Item clusters and computerized adaptive testing: A case for testlets. *Journal of Educational Measurement*, *24*, 185-202.

Wainer, H., & Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *Journal of Educational Measurement*, *37*, 203-220.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213.

Author Biographies

Muirne Paap is employed as a postdoctoral researcher at the University of Twente, the Netherlands. She has degrees in clinical psychology, psychiatry and psychometrics/statistics. Her research interests include latent variable modeling (e.g., Item Response Theory), classical test theory, standardized testing, educational measurement, and health assessment.

Qiwei He is an associate research scientist at the Educational Testing Service (ETS). Her background and expertise are in psychometrics, data mining, text mining and natural language processing with a focus on building applications for educational and psychological technology. Previously, Dr He conducted research in the areas of text-based online assessments for psychiatric and psychological studies at the University of Twente, the Netherlands.

Bernard Veldkamp is director of the Research Center for Examination and Certification and professor at the University of Twente, the Netherlands. He conducts research in the areas of educational, psychological, and health assessment. His work spans a range of issues in measurement and assessment, from the development of new methods/models for the design and construction of computerized (adaptive) psychological and educational tests, to the development of data mining models for analyzing verbal data and large datasets in fraud detection.