Chapter 6

Minimizing the Testlet Effect: Identifying Critical Testlet Features by Means of Tree-Based Regression

Muirne C.S. Paap and Bernard P. Veldkamp

Abstract Standardized tests often group items around a common stimulus. Such groupings of items are called testlets. The potential dependency among items within a testlet is generally ignored in practice, even though a basic assumption of item response theory (IRT) is that individual items are independent of one another. 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. Knowledge about these features might help to develop item sets with small testlet effects. This study illustrates the merits of TBR in the analysis of test data.

Keywords: testlets, testlet effects, testlet response theory, tree-based regression

Introduction

Standardized educational tests (which are often high-stakes tests) commonly contain sets of items grouped around a common stimulus, for example, a text passage, graph, table, or multimedia fragment, creating a dependence structure among items belonging to the same stimulus. Such groups of items are generally referred to as item sets or testlets (Wainer & Kiely, 1987), and this kind of dependence has been referred to as *passage dependence* (Yen, 1993). Testlets are popular for several reasons, including time efficiency and cost constraints, reducing the effects of context in adaptive testing, and circumventing concerns that a single independent test might be too atomistic in nature (measuring a concept that is very specific or narrow) (Wainer, Bradlow, & Du, 2000). In the Netherlands, testlets are, for example, used in the final examinations at the end of secondary education and in the "Cito test" (van Boxtel, Engelen, & de Wijs, 2011).

In most high-stakes tests, item response theory (IRT) models (Lord, 1980) are applied to relate the probability of a correct item response to the ability level of the candidate. A basic assumption underlying these models is that the observed responses to any pair of items are independent of each other given an individual's score on the latent variable (local independence, or LID).

However, for pairs of items grouped around the same testlet, responses might also depend on the common stimulus. Examinees might misread or misinterpret the stimulus, not like the topic, have particular expertise on the subject matter addressed by the stimulus, and so on.

In certain situations, the testlet structure could be accounted for by applying a polytomous IRT model, like the partial credit model, at testlet level, where the sumscore of the items in the testlets would function as the score on this polytomous item (e.g., Thissen, Steinberg, & Mooney, 1989; Verhelst & Verstralen, 2008). This polytomous approach to testlets would not result in any violations of local independence, and standardized software could be applied to estimate the models. However, there are some drawbacks. Until now, this approach has only been proposed for situations where the items within a testlet adhere to the very strict Rasch model. Furthermore, in calculating sumscores, the exchangeability of items is assumed, which may not be realistic in practice. Moreover, a guessing parameter at the item level cannot be taken into account. Alternatively, an approach can be used that accounts for the multilevel structure (items within testlets). Bradlow, Wainer, and Wang (1999) proposed to model the testlet effect by introducing a new parameter to the IRT models that accounts for the random effect of a person on items that belong to the same testlet, in order to adjust for the nested structure. This parameter, γ_{nt}, is referred to as the testlet effect for person *n* on testlet *t*. It represents a random effect that exerts its influence through its variance: the larger the variance σ_{1t}^2, the larger the amount of local dependence (LD) between the items *j* within the testlet *d* (Wainer & Wang, 2000).

Although several procedures for estimating testlet response models have been developed and applications of testlet response theory (TRT) have been studied (Glas, Wainer, & Bradlow, 2000; Wainer, Bradlow, & Wang, 2007), the dependency is often ignored in practice, and standard IRT models are used instead. The reason is obvious: assuming that LID holds, allows the use of simpler and well-known IRT analyses using easily accessible software. However, ignoring LD may lead to underestimation of the standard error of the ability estimates, as well as bias in the estimated item difficulty and discrimination parameter if the testlet effect is of a medium to large size (Wainer & Wang, 2000; Yen, 1993).

One way to approach this issue is to design testlets that show a small testlet effect. In a simulation study, Glas et al. (2000) investigated what the effect would be on the accuracy of item calibration if the testlet structure were to be ignored.

Their data-set was generated using the 3PL model and the following structure: $a_i \sim U(0.8,1.2), b_i \sim U(-1,1), c_i = 0.25,$ and $\theta \sim N(0,1)$. They compared the outcomes for the two values of σ_{1t}^2: 0.25 and 1.00. It should be noted that values of 1.00 or larger are often found in real data-sets. Their findings showed that the σ_{1t}^2 value of 0.25 resulted in negligible bias in item parameter estimates, whereas moderate effects were found for the σ_{1t}^2 value of 1.00 (Glas et al., 2000). Thus, if the testlet effects are small, the LD violation would be in an acceptable range, and models such as the 2PL or 3PL could be used without sacrificing the quality of the parameter estimation. A requirement for designing such testlets, however, is knowing which testlet characteristics are related to the testlet effect size.

Predicting Testlet Effects

In a recent study (Paap, He, & Veldkamp, submitted), which will be referred to here as "study 1," we used tree-based regression (TBR) to identify the key features of the stimuli that can predict the testlet effect in a standardized test measuring analytical reasoning. TBR is a popular method in the field of data mining, but it is becoming more popular in other fields as well, including educational measurement (e.g., Gao & Rogers, 2011). Like in other forms of regression analysis, TBR involves a set of independent variables and one or more dependent variables. Independent variables can be nominal, ordinal, or interval variables. A dependent variable is a continuous variable; if it is categorical in nature, a classification tree is generated. Independent variables can enter the tree more than once. Among TBR's advantages are its nonparametric nature, ease of interpretation, and flexibility in dealing with high-order interaction terms. An example of such a high-order interaction can be found in Figure 1: nodes 11 and 12, which are positioned in the right branch. These two nodes are the result of an interaction between four independent variables!

TBR can be used to divide the set of testlets iteratively in increasingly homogeneous subsets (so-called "nodes"). At each stage of the analysis, the testlet feature with the largest influence on the dependent variable is identified by using a recursive partitioning algorithm called the "classification and regression tree" (CART) (Breiman, Friedman, Olshen, & Stone, 1984). The CART algorithm starts by growing a large initial tree which overfits the data so as to not miss any important information.

In the next step, the tree is "pruned": a nested sequence of subtrees is obtained and, subsequently, one of them is selected based on pre-defined criteria. Typically, the final step consists of cross-validating the tree to determine the quality of the final model further.

Since we had a relatively small data-set (100 testlets)[1] in our study, the cross-validation resulted in trees with little explained variance, and there was a substantial effect of the random splitting of the data-set on the findings. Therefore, we chose not to use cross-validation in our study.

The dependent variable in our TBR is the standard deviation of the testlet parameter, denoted as σ_{1t}. Note that we deliberately chose to use σ_{1t} as opposed to σ_{1t}^2 in our model, since σ_{1t} capitalizes on the difference between testlets and is thus more informative in this setting. We estimated the testlet effect using a three-parameter normal ogive (3PNO) model, which is highly similar to the well-known 3PL model. The responses were coded as Y_{ni} = 1 for a correct response and Y_{ni} = 0 for an incorrect response. The probability of a correct response is given by

$$P(Y_{ni} = 1) = c_i + (1 - c_i)\Phi\left(a_i\theta_n + b_i + \gamma_{nt(i)}\right), \quad (1)$$

where $\Phi(.)$ is the probability mass under the standard normal density, and c_i is the guessing parameter of item *i*. $\gamma_{nt\,(i)}$ has a normal distribution; that is,

$$\gamma_{nt} \sim N(0, \sigma_{1t}^2). \quad (2)$$

The parameters were estimated in a fully Bayesian approach using an MCMC computation method. For details, see Glas (2012). Note that the model fit of (1) will be investigated in a future study. The average testlet effect estimated with the 3PNO equaled 0.71 (SD = 0.16). It should be noted that a value of σ_{1t} smaller than 0.50 has been shown to have a negligible effect on parameter estimates, whereas an effect near the size of 1.00 has a more substantial influence (Glas et al., 2000).

[1] Each respondent was presented with four out of 100 testlets; the four testlets were comprised of around 26 items each.

Identifying Testlet Characteristics

The features used as independent variables in our study can be divided into four categories: (1) variables describing the logical structure of the stimuli, (2) variables describing the themes contained in the stimuli, (3) surface linguistic variables, and (4) aggregated item characteristics. Two raters independently coded the variables in categories 1 and 2. In the case of discordant scorings, a consensus was reached through discussion; a discussion log was kept for these stimuli. The surface linguistic features were generated by using the specialized text-mining software Python (Python Software Foundation, 2009). The aggregated item characteristics were computed by averaging the attributes over all of the items in a testlet. In total, 22 independent variables were generated.

Study 1: Prediction Based on Testlet Features only

In our first study, we did not include information about the items in our prediction model. A two-step procedure was applied to build the prediction model. First, separate models were evaluated for each variable category (structure, theme, linguistic). The variables that were selected by the algorithm were then retained for each category, and subsequently all of the variables belonging to one of the other categories were added to the selected variables to see if any of them would be selected in the regression tree. In the next step, all of the variables that were not selected by the CART algorithm were removed from the list of independent variables and the variables of the remaining category were added. We then removed the variables that were not selected from the independent variable list again. In the case of competing models, the final model was selected based on the amount of the explained variance and the greatest number of splits resulting in a large difference in the mean testlet *SD* for the resulting nodes.

Summary of Results

Four independent variables were selected for the final prediction model: the percentage of "if" clauses, the predicate propositional density (the number of verbs divided by the total number of words, excluding punctuation), theme/topic, and the number of entities (entities are defined as the units in the stimulus that had to be assigned to positions). The latter two variables entered the tree at several splits. The total explained variance equaled 37.5%. The final tree consisted of 16 nodes. For every node, the mean value of σ_{1t} was larger than 0.50. For 6

nodes, the value of σ_{1t} exceeded 0.75. For all 6 nodes with a medium-large testlet effect, the percentage of "if" clauses was smaller or equal to 31%.

The largest testlet effects were found for the stimuli with a predicate propositional density of 0.098 or larger: 0.898 for stimuli with more than 10 entities and 0.980 for stimuli with 4 entities or fewer. For stimuli with a predicate propositional density smaller than 0.098, the largest testlet effect was found for the stimuli with the theme/topic that was either business, education, transport, or nature related, which had 5 entities or less, and a predicate propositional density between 0.071-0.097.

Study 2: Including Average Item Difficulty

Since a testlet effect is an additional source of variance in an *item* response function, the question arises whether attributes of items belonging to the testlet can be used to predict the testlet effect. In study 1, the focus was only on stimulus attributes. In this second study, aggregated item attributes will be included as well. Several interesting questions have to be answered, including whether there is a relationship between the average item difficulty in a testlet and the size of the testlet effect, whether characteristics of the testlet are related to average item difficulty, and whether there is an influence of item characteristics and the testlet location on the testlet effect. We made a first step towards illuminating these issues by investigating the relationship between average item difficulty within a testlet and the testlet effect size. We did this by adding the average item difficulty per testlet to the TBR model described in the previous study. The same two-step procedure for building the model was applied. The only difference was that besides the structure, theme, and linguistic variables, a fourth category of independent variables was added to the model.

Summary of Results

The resulting tree can be found in Figure 1. The total variance explained for this model was 41.4%, which implies that adding average difficulty as an independent variable improved the model.

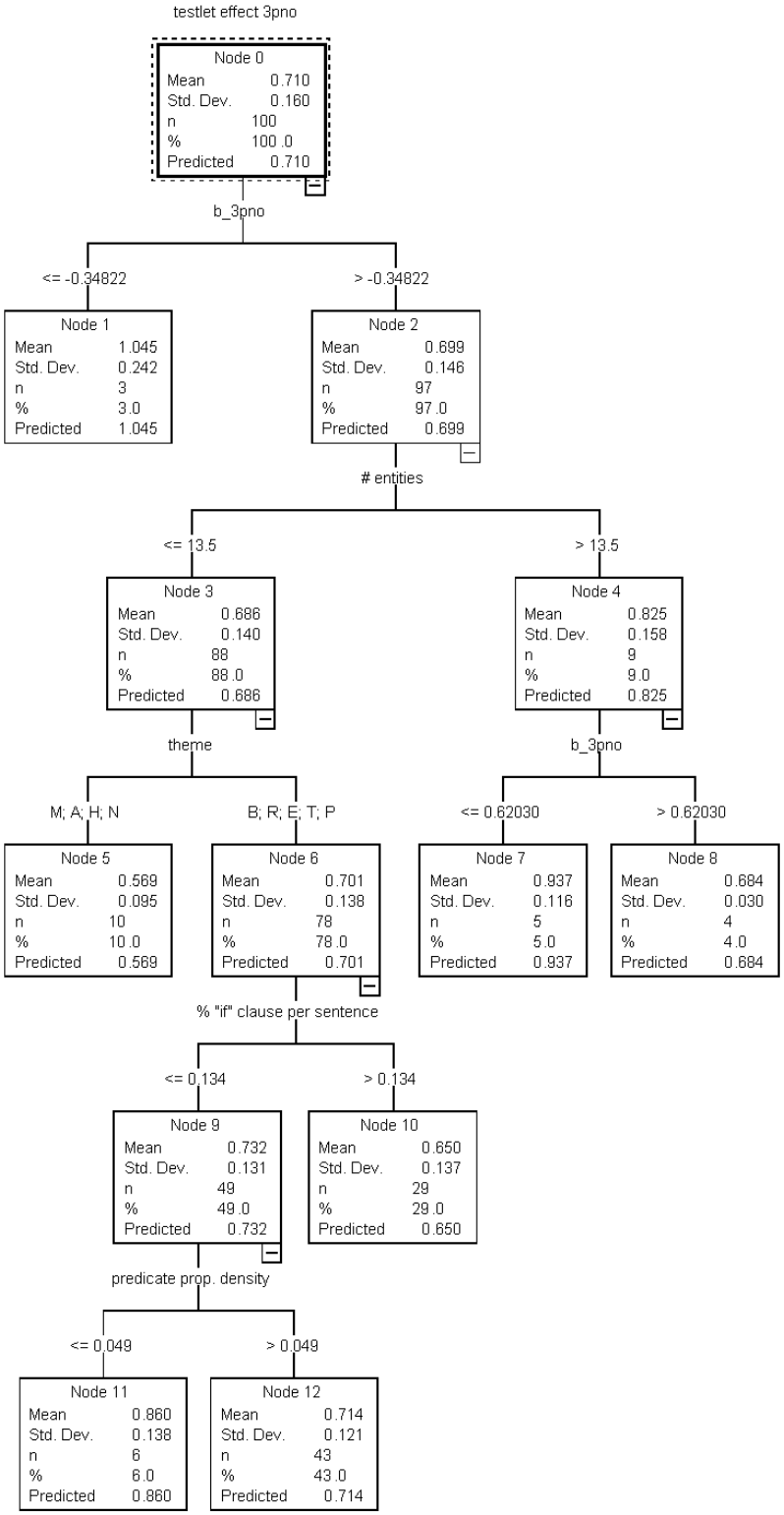

Figure 1 Regression tree based on the final model in study 2 with the 3PNO-based testlet effect as a dependent

When comparing study 2's model depicted in Figure 1 to the 3PNO-based model described in study 1, there are several important similarities. First, all of the variables that were contained in the tree described in study 1 were retained in the new model (Figure 1). Also, both models suggest that a large number of entities is associated with a larger testlet effect, and in a subset of testlets a low predicational propositional density score is associated with a larger testlet effect. However, it is important to note that the average item difficulty is chosen for the first split in the newer model, indicating its relative importance.

It can be seen that testlets containing easy items have a larger testlet effect. Furthermore, testlets with an average item difficulty between -0.35 and 0.62 that also contain 14 entities or more are also associated with a high testlet effect. Finally, testlets with an average item difficulty larger than -0.35; 13 entities or fewer; with a theme related to business, recreation, education, transport, or intrapersonal relationships/family; containing 13.4% or less "if" clauses; and that had a propositional density score of 0.049 or smaller also showed a larger testlet effect.

Conclusion

Our findings indicate that, for most testlets, testlet characteristics are associated with the size of the testlet effect, even when the average item difficulty has been accounted for. Three exceptions were found, all testlets with a relatively low average item difficulty. If these findings can be replicated, they may indicate that if testlets predominantly contain easy items, testlet characteristics are either of less importance to the size of the testlet effect or show considerable overlap with the information provided by the average item difficulty. In order to unravel this issue, we will have to explore the relationship between testlet characteristics (as independent variables) and average item difficulty per testlet (dependent variable) in a future study. In addition, other aggregated item variables might have to be added to the model to explore the relationship between item attributes and testlet effects more extensively.

In summary, we found evidence in our study for stimulus-related variables being associated with the size of the testlet effect. Our findings can be used in item construction, and the analyses we applied can be used as an example for others who construct and analyze similar data-sets to ours. However, a little more research is needed before solid "testlet construction rules" can be formulated.

Acknowledgement

This study received funding from the Law School Admission Council (LSAC). The opinions and conclusions contained in this report are those of the author and do not necessarily reflect the position or policy of LSAC.

References

Bradlow, E. T., Wainer, H., & Wang, X. H. (1999). A Bayesian random effects model for testlets. *Psychometrika, 64*(2), 153–168.

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

Gao, L., & Rogers, W. T. (2011). Use of tree-based regression in the analyses of L2 reading test items. *Language Testing, 28*(1), 77–104. doi: 10.1177/0265532210364380

Glas, C. A. W. (2012). *Estimating and testing the extended testlet model.* LSAC Research Report Series. Newtown, PA: Law School 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, Netherlands: Kluwer.

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

Paap, M. C. S., He, Q., & Veldkamp, B. P. (submitted). Identifying critical testlet features using tree-based regression: An illustration with the analytical reasoning section of the LSAT.

Thissen, D., Steinberg, L., & Mooney, J. A. (1989). Trace lines for testlets: A use of multiple-categorical-response models. *Journal of Educational Measurement, 26*(3), 247–260. doi: 10.1111/j.1745-3984.1989.tb00331.x

van Boxtel, H., Engelen, R., & de Wijs, A. (2011). *Wetenschappelijke verantwoording van de Eindtoets 2010*. Arnhem: Cito.

Verhelst, N. D., & Verstralen, H. H. F. M. (2008). Some considerations on the partial credit model. *Psicologica, 29*, 229–254.

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. 245–270). Dordrecht, Netherlands: Kluwer.

Wainer, H., Bradlow, E. T., & Wang, X. (2007). *Testlet response theory and its applications*. New York: Cambridge University 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*(3), 203–220.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187–213.