frontiers in Education

METHODS
published: 07 December 2017
doi: 10.3389/feduc.2017.00063

Robust Automated Test Assembly for Testlet-Based Tests: An Illustration with Analytical Reasoning Items

Bernard P. Veldkamp[1] and Muirne C. S. Paap[2]*

[1] University of Twente, Enschede, Netherlands, [2] University of Groningen, Groningen, Netherlands

In many high-stakes testing programs, testlets are used to increase efficiency. Since responses to items belonging to the same testlet not only depend on the latent ability but also on correct reading, understanding, and interpretation of the stimulus, the assumption of local independence does not hold. Testlet response theory (TRT) models have been developed to deal with this dependency. For both logit and probit testlet models, a random testlet effect is added to the standard logit and probit item response theory (IRT) models. Even though this testlet effect might make the IRT models more realistic, application of these models in practice leads to new questions, for example, in automated test assembly (ATA). In many test assembly models, goals have been formulated for the amount of information the test should provide about the candidates. The amount of Fisher Information is often maximized or it has to meet a prespecified target. Since TRT models have a random testlet effect, Fisher Information contains a random effect as well. The question arises as to how this random effect in ATA should be dealt with. A method based on robust optimization techniques for dealing with uncertainty in test assembly due to random testlet effects is presented. The method is applied in the context of a high-stakes testing program, and the impact of this robust test assembly method is studied. Results are discussed, advantages of the use of robust test assembly are mentioned, and recommendations about the use of the new method are given.

Keywords: automated test assembly, high-stakes testing, robust optimization, robust test assembly, testlet response theory

OPEN ACCESS

Edited by:
Mustafa Asil,
University of Otago, New Zealand

Reviewed by:
Caterina Primi,
University of Florence, Italy
Haci Bayram Yilmaz,
Ondokuz Mayis University, Turkey

****Correspondence:***
Bernard P. Veldkamp
b.p.veldkamp@utwente.nl

Specialty section:
This article was submitted to Assessment, Testing and Applied Measurement, a section of the journal Frontiers in Education

Received: *03 August 2017*
Accepted: *16 November 2017*
Published: *07 December 2017*

Citation:
Veldkamp BP and Paap MCS (2017) Robust Automated Test Assembly for Testlet-Based Tests: An Illustration with Analytical Reasoning Items. Front. Educ. 2:63. doi: 10.3389/feduc.2017.00063

INTRODUCTION

In many tests, a reading passage, graph, video fragment, or simulation is presented to a test taker; and after reading the passage, studying the graph, watching the video fragment, or participating in the simulation, the test taker is presented with a number of items pertaining to the stimulus. Such a group of items can be referred to as a testlet (Wainer and Kiely, 1987). The responses of the test takers to items in the testlet depend on the correct reading, interpretation, and understanding of the stimulus. This causes a dependency among the responses given to the items pertaining to the same stimulus. Even after controlling for latent ability, responses to items within a testlet tend to be correlated. This violates the assumption of local independence (LI) (Koziol, 2016). For an item pair that shows positive local dependence, the information provided by the items consists of a redundant (overlapping) part and a unique part. If the local dependency is ignored, the overlapping

part is counted twice, and thus the measurement precision is overestimated[1] (e.g., Junker, 1991; Braeken, 2011; Baghaei and Ravand, 2016). To deal with these kinds of issues, testlet response theory (TRT; Wainer et al., 2007) models were proposed. In these models, the dependency between responses to items in the same testlet is modeled by adding a testlet effect to the item response theory (IRT) models that accounts for the excess within-testlet variation.

Applying TRT models to practical testing problems was found to reduce overestimation of the precision of the ability estimates (e.g., Wainer and Wang, 2000; Chang and Wang, 2010). It was demonstrated that for some ability levels, information in the test was overestimated by 15%, when the testlet structure was not taken into account. Application of these new TRT models also led to new questions. For example, in many large-scale testing programs, automated test assembly (ATA) methods are applied to select items from an item bank to build new test forms. Depending on the amount of information the items provide, they are generally selected either consecutively (e.g., Lord, 1977) or simultaneously (e.g., van der Linden, 2005). For some test assembly problems, the amount of information in the test has to be maximized, whereas for other test assembly problems, the amount of information has to meet a prespecified target. van der Linden (2005) (Chap. 1) describes how targets might vary depending on the goal of testing. For making pass/fail decisions, the target information function (TIF) has to be peaked around the cutoff score, while for broad ability testing, the TIF might be uniform for all relevant ability values. ATA methods typically model the test assembly problem as a mathematical programming problem that maximizes an objective function, for example, related to the amount of information in the test, while a number of constraints, for example, related to the test specifications, have to be met. Mathematical programming solvers like CPLEX (IBM, 2016) or Gurobi (Gurobi Optimization, 2016) can be applied to solve the problems and to assemble the tests.

One of the main assumptions of ATA is that the coefficients of the test assembly models are fixed and known. In TRT, this might be a problem, because testlet effects are typically modeled as random effects (Wainer et al., 2007), and the random testlet effects cause uncertainty in the information functions. The question arises: how can we assemble test forms when Fisher Information varies from person to person?

To answer this question, TRT models will be presented in more detail first. Existing methods for robust ATA will be described and evaluated theoretically in case of ATA with testlets. We will focus on the specific problems in ATA caused by the uncertainty in Fisher Information at the individual level. The implications and shortcomings of the existing methods will be discussed. After that, a method for robust ATA will be presented. It will then be applied in the context of a high-stakes testing program. The resulting test forms will be compared for various settings of the method. Finally, implications of this new method for automated assembly of tests with testlets will be discussed, and recommendations will be given.

[1] Conversely, for negative local dependence measurement, precision may be underestimated (Braeken, 2011).

TESTLET RESPONSE THEORY

One of the assumptions of IRT states that the observed responses to items are independent of each other given a test taker's score on the latent ability. For items within testlets, the assumption of LI does not hold. Besides the latent ability, the responses also depend on the common stimulus. To account for this dependency a testlet effect can be added to a response model. For example, let the response behavior of a test taker be described by the 3-parameter logistic model (3PLM). Define,

$$\tau_{ij} = a_i(\theta_j - b_i), \tag{1}$$

where a_i denotes the discrimination parameter of item i, b_i denotes the difficulty parameter, and θ_j denotes the latent ability of person j. When c_i denotes the guessing parameter for item i, the 3PLM can be formulated as follows:

$$P_i(\theta_j) = c_i + (1-c_i)\frac{\exp(\tau_{ij})}{1+\exp(\tau_{ij})}. \tag{2}$$

To extend the 3PLM to a 3-parameter testlet model (3PL-T), a random testlet effect $\gamma_{jt(i)} \sim N\left(0, \sigma^2_{t(i)}\right)$ for person j on testlet $t(i)$, where $\sigma^2_{t(i)}$ indicates the strength of the testlet effect, can be added to the exponent:

$$\tau_{ij} = a_i(\theta_j - b_i + \gamma_{jt(i)}). \tag{3}$$

Several procedures for estimating testlet response models have been developed and applications of TRT have been studied (Glas et al., 2000; Wainer et al., 2007). Recently, Paap et al. (2013) proposed to reduce the variance of the testlet effect by adding a fixed effect to the model in Eq. 3, which depended on features that described the stimulus (e.g., word diversity, topic, or structure of the stimulus):

$$\tau_{ij} = a_i\left(\theta_j - b_i + \sum_q x_{t(i)q}\beta_q + \gamma_{jt(i)}\right), \tag{4}$$

where β_q is the regression coefficient associated with feature q (see also Glas, 2012a). Paap et al. (2015) showed how regression trees can be used to select testlet features with predictive value for the testlet effect.

Testlet response theory can be used to estimate the latent abilities more realistically, by taking the dependency between the items into account. Glas et al. (2000) showed that ignoring the testlet effect can lead to biased item parameter estimates. Wang et al. (2002) illustrated that ignoring testlet effects provides SEs that will be potentially too small when the testlet effect is neglected.

Fisher Information

Fisher Information is defined as the negative inverse of the asymptotic variance. For the 3PL-T model, Fisher Information for item i at ability level θ_j can be formulated by the following equation:

$$I_i(\theta_j) = a_i^2\left(\frac{\exp\tau_{ij}}{1+\exp\tau_{ij}}\right)^2\frac{1-c_i}{c_i+\exp\tau_{ij}}, \tag{5}$$

where τ_{ij} is the same latent linear predictor as in Eq. 4. An interesting feature of this information function is that it has some uncertainty in it due to the probabilistic nature of the testlet effect. At an individual level, the location of the information function varies based on the testlet effect.

Maximum Fisher Information is obtained for

$$\theta_j = b_i + \gamma_{jt(i)} + \frac{1}{a_i}\left\{\ln\frac{1+(1+8c_i)^{1/2}}{2}\right\}. \quad (6)$$

In other words, the testlet effect does have an impact on the ability level θ_j for which the maximum amount of information is obtained. Besides, given an ability level, it can be deducted that the larger the testlet effect, the larger the deviation in Fisher Information between TRT models that take the effect into account ($\gamma_{jt(i)} \neq 0$), and IRT models that assume $\gamma_{jt(i)} = 0$. In the next section, a method is introduced to deal with this uncertainty in the test assembly process.

ROBUST ATA

In ATA, items are selected from an item bank based on their properties. In this selection process, 0–1 linear programming (0–1 LP) techniques are generally applied (e.g., van der Linden, 2005). The first step in ATA is to formulate the test assembly problem as a linear programming model. These models are characterized by decision variable $x_i \in \{0,1\}$ for $i = 1, \ldots, I$, which denotes whether an item i is selected for the test ($x_i = 1$) or not ($x_i = 0$). An objective function, for example, to maximize the amount of information in a test or to minimize the deviation from a TIF, is defined, and restrictions related to the test specifications are imposed.

Let

c be the vector of coefficients of the objective function;
A be a matrix of coefficients of the various constraints;
b be a vector of bounds;
N be the number of items in the bank;
x be a vector of decision variables;

a general model for ATA can now be formulated as follows:

$$\max c^T x, \quad (7)$$

subject to

$$Ax \leq b, \quad (8)$$

$$x \in \{0,1\}^N. \quad (9)$$

For an extensive introduction to the problem of model building in 0–1 LP, see Williams (1999) or van der Linden (2005) (Chap. 2–3). These optimization problems are solved either by applying Branch-and-Bound-based solvers that search for optimal solutions (van der Linden, 2005) (Chap. 4) or by using heuristical approaches (e.g., Swanson and Stocking, 1993; Armstrong et al., 1995; Luecht, 1998; Veldkamp, 2002; Verschoor, 2007).

ATA with Testlets

The model in Eqs 7–9 has been formulated to select individual items from an item bank. However, for some testing programs, the item bank may consist of subsets of items, for example, when items belong to a testlet. To deal with the testlet structure during test assembly, additional constraints might have to be added to the test assembly model: (1) the number of testlets to be selected for a test is bounded by a minimum or a maximum and (2) for every testlet that is selected for the test, a minimum and/or maximum number of items has to be selected from the corresponding set.

To model these limitations, an additional set of decision variables $z_{t(i)} \in \{0,1\}$ for $t(i) = 1, \ldots, T$, has to be defined that denotes whether testlet $t(i)$ is selected for the test [$z_{t(i)} = 1$] or not [$z_{t(i)} = 0$]. Imposing the additional constraints on the general model for test assembly in Eqs 7–9 comes down to adding the following constraints:

$$b_{Zl} \leq 1^T z \leq b_{Zu}, \quad (10)$$

$$n_{t(i)l} z_{t(i)} \leq \sum_{i \in V_{t(i)}} x_I \leq n_{t(i)u} z_{t(i)} \; \forall t(i), \quad (11)$$

$$z \in \{0,1\}^T. \quad (12)$$

where 1 denotes the unity vector; $t(i)$ is an indicator for the testlets; z is a vector of decision variables $z_{t(i)}$; b_{Zl} is a lower bound on the number of testlets in a test; b_{Zu} is an upper bound on the number of testlets in a test; $V_{t(i)}$ set of items belonging to testlet s; $n_{t(i)l}$ is the minimum number of items to be selected for testlet s once the testlet is selected; and $n_{t(i)u}$ is the maximum number of items to be selected for testlet s once the testlet is selected.

Please note that testlets can be seen as a special type of item sets. For an overview of how to model ATA problems with item sets, see van der Linden (2005) (Chap. 5).

Robust ATA with Testlets

When TRT is used to model the responses, the coefficients of Fisher Information have uncertainty in them as illustrated in Eq. 5. So either the coefficients of the objective function $c^T x$ become uncertain, or the coefficients of some of the constraints $Ax \leq b$ are affected. Several methods for dealing with uncertainty in 0–1 LP models have been proposed in the literature. First of all, Soyster (1973) proposed to take the maximum level of uncertainty into account in the 0–1 LP model. For large problems with uncertainty in many parameters, this method turned out to be very conservative. In case of ATA with testlets, it would imply that three times the SD of the testlet effect would be subtracted in Eq. 5, and the resulting value for the information function would be close to 0. A less conservative alternative was proposed in De Jong et al. (2009), where only 1 SD was subtracted. Veldkamp et al. (2013) studied the De Jong et al. (2009) method more into detail and found that subtracting 1 SD for all items might not be realistic and for long tests it might be too conservative.

Soyster (1973) based methods assume all the uncertain coefficients parameters have maximum impact on the solution of a 0–1 LP problem, which is usually not the case in practice. Bertsimas and Sim (2003) observed that it hardly ever occurs that uncertainties in all coefficients impact the solution. They developed a method for solving 0–1 LP optimization problems with uncertainty in some of the parameters. They proved that when uncertainty in some of the coefficients affects

the solution, 0–1 LP problems with uncertainty in the coefficients can be solved as a series of 0–1 LP problems *without* uncertainty in the coefficients. Veldkamp (2013) applied their method for ATA problems.

Let

Γ denotes the protection level, which is the number of items for which uncertainty impacts the solution (this number has to be specified by the user);

d_i represents the uncertainty in the coefficients of the objective function c_i;

$\tilde{a}_{ik}$ represents the uncertainty in the coefficients a_{ik} of constraint k.

The first step in modeling ATA problems with testlets is to reorder the items according to their maximum amount of uncertainty $d_1 \geq d_2 \geq \ldots \geq d_n$, and define $d_{n+1} = 0$. Note that for every item belonging to the same testlet the deviations d_i are identical. Once the items have been reordered, the following sets can be defined. For any item m, let

S_m be the subset of items with $d_i > d_m$;

$\tilde{S}_{mk}$ be the subset of items with $\tilde{a}_{ik} > \tilde{a}_{mk}$.

Following Veldkamp (2013), a generic model for robust testlet assembly problems with protection level Γ can be formulated as follows:

$$\max_{m=1,\ldots,n+1}\left\{\max c^T x - d_m\Gamma + \sum_{i\in S_m}\left(d_i - d_m\right)x_i\right\}, \tag{13}$$

subject to

$$Ax + \Gamma h_1 + \sum_{i\in\tilde{S}_{mk}} H_{2i} \leq b, \tag{14}$$

$$h_{1j} + H_{2ik} \geq \tilde{\alpha}_{ik} y_i \;\forall i \in \tilde{S}_{mk}, k, \tag{15}$$

$$x \leq y, \tag{16}$$

$$b_{Zl} \leq 1^T z \leq b_{Zu}, \tag{17}$$

$$n_{t(i)l} z_{t(i)} \leq \sum_{i\in V_{t(i)}} x_I \leq n_{t(i)u} z_{t(i)} \;\forall t\left(i\right), \tag{18}$$

$$h_1, H_2, y \geq 0, \tag{19}$$

$$z \in \left\{0,1\right\}^T, \tag{20}$$

$$z \in \left\{0,1\right\}^I. \tag{21}$$

where h_1 is an auxiliary vector; H_2 is an auxiliary matrix; and y is a vector of auxiliary decision variables.

In this model, the original objective function $\max c^T x$ is corrected for uncertainty. For each of the subsequent optimization problems $m = 1, \ldots n + 1$ the correction term is equal to Γ times the maximum deviation of the lth item plus an additional correction when some of the items with a larger maximum deviation than item l are selected. For example, let the protection level $\Gamma = 5$. This implies that the uncertainty in at most five of the items is assumed to impact the test assembly problem. To solve the second optimization problem, the set $S_2 = \{1\}$, since only item 1 has a larger maximum deviation than item 2 in the reordered item bank. Therefore, the correction term for this problem is equal to $(5^*d_2) + (d_1 - d_2)x_1$. To deal with uncertainties in the constraints, the same logic is applied. But since the items cannot be reordered for every constraint, the auxiliary matrix and vectors are needed in the model formulation.

Uncertainty due to the testlet effect either affects the objective function when Fisher Information is maximized, or it affects the constraints when Fisher Information has to meet specific bounds for the TIF. Since Fisher Information is a function of ability and not a scalar, it is generally discretized and the optimization problem is solved as a maximin problem over a number of ability values (Boekkooi-Timminga and van der Linden, 1989). Instead of deviations d_i, deviations d_{ik} have to be defined, which denote the deviation from the objective function for θ_k, where $k = 1, \ldots, K$, and θ_k denote the evaluation points of the information function at the ability scale. Veldkamp (2013) showed how to model a robust maximin problem.

A Different Approach for Defining Deviations d_{ik}

In both Bertsimas and Sim (2003) and Veldkamp (2013), the deviations are related to the maximum uncertainty for item *i*. For the problem at hand, this might be far too drastic. In ATA with testlets, the uncertainty is caused by normally distributed testlet effects $\gamma_{jt(i)} \sim N\left(0, \sigma^2_{t(i)}\right)$. This implies that a testlet effect might be equal to three times the SD. However, setting the deviation to its maximum uncertainty decimates the contribution of the items belonging to this testlet to the objective function. This is not realistic, since such deviations are only expected to occur for 0.27% of the test takers.

Besides, most tests consist of a limited number of testlets. The analytical reasoning (AR) section of the LSAT, for example, consists of four stimuli. Veldkamp (2013) already suggested to replace the maximum uncertainty by the expected maximum uncertainty. For testlets, this would imply that the deviations d_i are based on the expected maximum absolute value of a number of draws from normally distributed testlet effects with mean equal to 0 and known SDs, where the number of draws equals the number of testlets in the test. Tippett (1925) demonstrated that the extreme value of a number of draws from a normal distribution does not have a normal distribution, and that it is far from straightforward to calculate them analytically. For a table of the maximum of a number of draws from a normal distribution, see, e.g., Harter (1960). For example, in case of four testlets, the expected maximum equals 1.027 SDs. This is much smaller than the maximum of 3 SDs. The impact of various settings of the deviations d_i is illustrated in Section "Numerical Examples."

NUMERICAL EXAMPLES

Testlet Pool

The empirical item bank used in this study consists of 594 items nested within 100 testlets. The bank came from the AR section of the Law School Admission Test. AR items test the ability to reason

within a given set of conditions. An example of an AR testlet was provided in the official LSAT Handbook (Law School Admission Council, 2010, pp. 6–8):

> Each of five students – Hubert, Lori, Paul, Roberta, and Sharon – will visit exactly one of three cities – Montreal, Toronto, or Vancouver – for the month of March, according to the following conditions: Sharon visits a different city than Paul. Hubert visits the same city as Roberta. Lori visits Montreal or else Toronto. If Paul visits Vancouver, Hubert visits Vancouver with him. Each student visits one of the cities with at least one of the other four students.

Which one of the following could be true?

(A) Hubert, Lori, and Paul visit Toronto, and Roberta and Sharon visit Vancouver.
(B) Hubert, Lori, Paul, and Roberta visit Montreal, and Sharon visits Vancouver.
(C) Hubert, Paul, and Roberta visit Toronto, and Lori and Sharon visit Montreal.
(D) Hubert, Roberta, and Sharon visit Montreal, and Lori and Paul visit Vancouver.
(E) Lori, Paul, and Sharon visit Montreal, and Hubert and Roberta visit Toronto.

Other items for this testlet are questions such as: which of the following could be false? If Sharon visits Vancouver, which one of the following must be true? Each of these questions comes with five possible answers as well. For more examples, we refer to the Official LSAT Handbook (Law School Admission Council, 2010).

Pretesting data were gathered in an incomplete design, where 49,256 candidates each responded to four testlets. Bayesian estimates of the parameters were made using MCMC methodology (Glas, 2012a). The number of respondents varied from 1,500 to 2,500 respondents per item. Descriptive statistics on the item parameters are provided in **Table 1**.

The average SE of estimation of the parameters was quite reasonable given the small number of respondents per item. Glas (2012b) demonstrated that the TRT model had an acceptable fit. For the purpose of this study, the parameters were transformed from the 3PNO-T framework (Glas, 2012b) to a 3PL-T framework, applying $D = 1.702$.

Various Conditions

In this study, the impact of various settings of robust test assembly models is compared. The resulting test had to meet the following specifications. First of all, a TIF was defined. For five θ-values, both a lower and an upper bound for Fisher Information in the test were imposed. The TIF was formulated based on the average amount of information provided by the items in the bank.

TABLE 1 | Descriptive statistics testlet pool.

	Minimum	Maximum	Average	SE
Item discrimination α	0.077	3.361	1.260	0.079
Item difficulty β	−1.482	3.188	0.605	0.136
Item guessing *c*	0.036	0.865	0.222	0.035
Testlet effect $\sigma^2_{t(i)}$	0.428	1.289	0.707	0.063

Furthermore, several test specifications were imposed. For the items in the bank, several item types were distinguished. The number of testlets per item type was fixed. Besides, the number of testlets per test was set equal to four, and the number of items per test was set equal to 24. Because of this, the total number of constraints was equal to 16. For the test assembly model, this implied that uncertainties just played a role in the constraints related to the TIF. For these constraints, $\tilde{d}_{ik}$ represents the uncertainty in Fisher Information for item i at ability level $\theta_k \in \{-0.5,0,0.5,1,1.5\}$. Due to the effects of uncertainty on Fisher Information, it might be possible that either the lower bounds imposed by the TIF, the upper bounds, or both, can no longer be met. The test assembly model might become infeasible. Following Huitzing et al. (2005) and Veldkamp (1999), we forced a solution in these cases by minimizing the sum of violations of these bounds. The violations were defined as the absolute difference between Fisher Information and its bound.

Several conditions were compared. In condition 1 no uncertainty due to testlet effects was taken into account. This condition was used as a bench mark. In condition 2, the Veldkamp (2013) model with deviations $\tilde{d}_{ik}$ equal to their maximum values was applied. We compared four different settings, where uncertainty due to the testlet effect in one, two, three, or all four testlets was assumed to have an impact on Fisher Information of the test. In the original Bertsimas and Sim (2003) and in the Veldkamp (2013) model, the maximum number of items for which uncertainty was assumed to have an impact on the objective function, that is, on the test information function, had to be specified at item level. But due to the nested nature of items within testlets, and since the uncertainty was caused by a parameter at testlet level, is was decided to assume impact of uncertainty at testlet level as well. As a consequence, Γ was only allowed to take values equal to the total number of items in the affected testlets, and the deviations for the items belonging to the same testlet were identical. In condition 3, the modified version of the Veldkamp (2013) model was implemented, where the expected maximum deviation was used to calculate the deviations $\tilde{d}_{ik}$. The same settings as in the second condition were applied and compared. Impact of uncertainty on one, two, three, or all four testlets was studied.

The resulting tests were compared based on the sums of violations of the upper and lower bounds of the TIF over the five ability values θ_k. In this study, software packages Microsoft Excel and Cran R were used. The simulations were run in R (R Development Core Team, 2012) version 2.13.

RESULTS

In the test assembly process, both a lower and an upper bound for the TIF had to be met. The information functions of the test assembled without taking uncertainties due to testlet effects into account (condition 1) are shown in **Figure 1**. The gray lines in **Figure 1** represent the TIF and both bounds. It has to be mentioned that any test that met the specifications would have been acceptable as a solution to the first test assembly model. The current solution was randomly drawn from the set of feasible solutions,

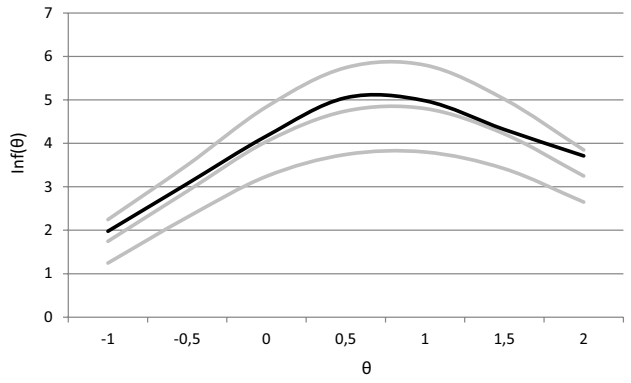

FIGURE 1 | Test information function for condition 1 without taking uncertainty into account.

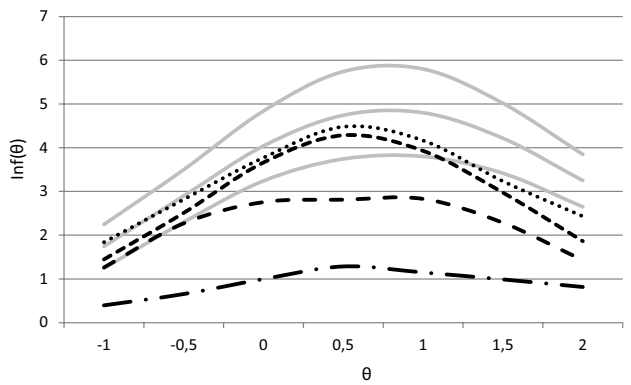

FIGURE 2 | Test information functions for condition 2 (A_t = 1 with small dots, A_t = 2 with small dashes, A_t = 3 with medium dashes, and A_t = 4 with dash/dot).

by the test assembly algorithm. The information function is close to the TIF, and none of the bounds is violated. Since the target was defined based on the average amount of information provided by the items in the bank, neither the very informative testlets nor the uninformative testlets were selected for this test. The testlet effects for this solution varied from $\sigma_S = 0.469$ to $\sigma_{33} = 0.995$.

In condition 2, the Veldkamp (2013) model was applied, where uncertainty due to testlet effects was assumed to affect the solution for at most one, two, three, or four testlets, respectively. These settings are denoted by $A_t \in \{1,2,3,4\}$ in **Figure 2**. For the problem denoted by $A_t = 1$, uncertainty played a role in one testlet. In the optimization model (Eqs 13–21), Γ was set equal to the number of items in the testlet; for all items in this testlet, the deviations $\tilde{d}_{ik}$ for $\theta_k \in \{-0.5,0,0.5,1,1.5\}$ were calculated by setting the testlet effect equal to $\gamma_{jt(i)} = 3^*\sigma_{t(i)}$ in Eq. 5, and calculating the difference with $\gamma_{jt(i)} = 0$. For the problem denoted by $A_t = 2$, Γ was set equal to the sum of the number of items in both the affected testlets, and so on. Taking the uncertainty into account resulted in a decreasing contribution of the affected testlets to the objective function. Defining $\tilde{\alpha}_{ik}$ based on maximum deviations resulted in an average loss of information by 85%. For some items, the information was reduced by 66%, but especially for those testlets that were informative at a specific range of the ability scale, the amount of information was reduced by almost 95%.

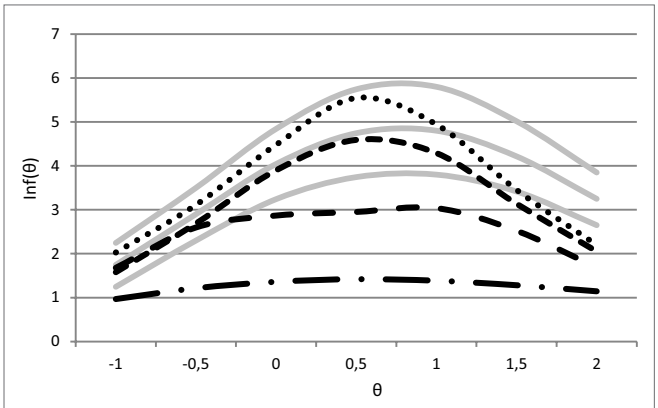

FIGURE 3 | Test information functions for condition 3 (A_t = 1 with small dots, A_t = 2 with small dashes, A_t = 3 with medium dashes, and A_t = 4 with dash/dot).

For $A_t = 1$, the consequence was that one testlet only contributed at most 33% of its information to the objective function. The test assembly algorithm could compensate this, by selecting more informative testlets or testlets with smaller testlet effects. In comparison with the solution of condition 1, one alternative testlet was selected, and the maximum testlet effect reduced to $\sigma_{92} = 0.789$.

When the number of testlets for which uncertainty was assumed to play a role increased, larger violations and more violations of the lower bound for the TIF occurred. For $A_t = 2$, only one (larger) violation occurred. For $A_t = 3$ and $A_t = 4$, the lower bound was violated for all five evaluations points θ_k. In comparison with the solutions in conditions 1 and 2, different testlets were selected, and the maximum testlet effect of the selected testlets reduced to $\sigma_{74} = 0.518$.

In the third condition, the deviations $\tilde{d}_{ik}$ were defined based on the maximum expected effect of four draws from a standard normal distribution. This resulted in an average loss of information of 64%. For some items, the information was reduced by 30%, and for those testlets that were informative at a specific range of the ability scale, the amount of information was still reduced by almost 85%. The information functions of the resulting tests for $A_t \in \{1,2,3,4\}$ are shown in **Figure 3**. By selecting different testlets that were more informative and had smaller testlet effects, a feasible test was assembled in case of $A_t = 1$. For $A_t = 2$, only one violation occurred. For $A_t = 2$, four violations occurred. Finally, for $A_t = 4$, the lower bound was violated for all five evaluation points. The same testlets were selected for $A_t \in \{2,3,4\}$ in conditions 2 and 3. The reason is that even though the size of the deviations $\tilde{d}_{ik}$ differed, the relative order of the testlets did not.

DISCUSSION

Taking the testlet effect into account when estimating the ability level prevents that measurement precision is overestimated, which implies that too much confidence is given to estimated ability levels (e.g., Wainer et al., 2007). In other words, small measurement errors might be a statistical artifact when testlet effects are neglected. In this article, it was illustrated how the presence of testlet effects in IRT models introduces uncertainty in Fisher item information at the individual level and affects ATA. Testlet effects can be seen as an interaction effect between a person and a stimulus, modeling that one test taker perceives the items within one testlet as more difficult or less difficult in comparison with other test takers, depending on characteristics of the stimulus. The testlet parameter, $\gamma_{jt(i)}$, is normally distributed around 0; but for individual persons within a population, it might have an effect, and the amount of Fisher information can decrease as a consequence. In this article, a model was presented to take this uncertainty into account during test assembly.

The Veldkamp (2013) model for robust test assembly was applied. The results showed that straightforward implementation of this model turned out to be very conservative. Using the expected maximum uncertainty as an alternative measure for deviations was more realistic. Tippett (1925) already showed that the expected maximum draw is much smaller than the maximum possible draw. Results illustrate how uncertainty can be taken into account without being over conservative. Especially, in the case where maximum uncertainty in only one testlet is assumed to influence the amount of information in the test, the modified approach resulted in a test that met the requirements.

The method proposed in this article does depend on choices made during specification of the test assembly model. Choices can be made with respect to definition of the deviations $d_{i,j}$. Besides, a reasonable value has to be chosen for Γ, the number of items for which uncertainty is assumed to play a role. In this article, several values were chosen to illustrate the impact of both kinds of parameters on the resulting tests. A balance has to be found between obtaining a feasible solution and objective value correction, where large values for Γ prevent overestimation of the precision of the ability estimate but might result in infeasible ATA problems. Bertsimas and Sim (2003) call this the price of robustness. For testlet assembly problems where uncertainty is related to a normally distributed testlet effect, the most reasonable value for Γ depends on the number of testlets in the test. For the numerical example at hand, it seems reasonable to assume an effect for uncertainty in only one or two of the testlets, since the probability of three or four draws of at least $d_{i,j}$ from a standard normal distribution, given the total number of four draws, is very small.

It could be argued that the method proposed in the article solves the problem of dealing with uncertainty in the testlet effects by making all kinds of assumptions about the impact of the uncertainty to manage it and to present a solution that is reasonably close to the TIF. Would not it be better to address the issue, to illustrate the consequences, and to advise users to collect more data to reduce uncertainty in the estimates of the testlet effects? The suggestion to collect more data, assuming that this is an option, should always be considered. Although it should be mentioned that budgets and capacity for pretesting the items are often limited. In this article, we attempt to illustrate and to substantiate that subjective choices have to be made in the setting of the algorithms for solving optimization problems with uncertainty in the parameters. Straightforward implementation of the most conservative settings does not always lead to feasible solutions. Especially in the case of automated assembly of tests with testlets, it is our opinion that straightforward implementation of conservative settings of the algorithms does not provide useful results. Many operational tests, like the AR section of the LSAT, consist of a limited number of testlets. Without correction, the resulting test assembly problems will probably be infeasible, which means that the set specifications that have been formulated for a test cannot be met. Infeasibility would imply that testing organizations cannot deliver. Therefore, most organizations implement the approach of condition 1, where uncertainty in the testlet parameters is neglected in ATA. In this article, it is illustrated that this approach might result in an overestimation of the precision of the ability estimates. The modified robust test assembly approach is proposed as an alternative they might consider. Moreover, it should be remarked that after the formal administration of the test to a large group of test takers, both item and testlet parameters can be reestimated and it can be tested whether the assumptions made during test assembly were legitimate.

In previous papers about robust test assembly (De Jong et al., 2009; Veldkamp, 2013; Veldkamp et al., 2013), uncertainty in test assembly was always related to uncertainty in the item parameter estimates. In this article, uncertainty was related to the violation of the assumption of LI, and the presence of testlet effects in TRT models. Even though different kinds of uncertainty were modeled, the same methods for robust ATA were applicable. One could even decide to take the uncertainty in both the item and the testlet parameters into account in ATA, and to model both kinds of uncertainty. The result would be that more uncertainty would be present in the ATA models, and, as a consequence, the resulting tests would be assembled based on a more conservative estimate of the measurement precision. The precise implementation, however, is a topic of further research. Another limitation of the study is that we only varied the number of testlets for which uncertainty plays a role. Besides, all results are based on a single item pool. The AR item pool is well balanced and consists of 100 testlets. Many operational item pools are, for example, much smaller. Because of this, our results can only be generalized to a certain level. In small item pools, for example, the differences between methods will be smaller, since more often the same set of testlets will be selected by the various methods.

Overall, it can be concluded that robust test assembly can be applied to prevent overestimation of the information in the test due to testlet effects. It results in a lower bound for the true information for all candidates in the final test. In this way, robust ATA provides test developers with the tools to handle testlet effects during test assembly, and it gives a greater level of certainty as to the true quality of the resulting test.

AUTHOR CONTRIBUTIONS

Both authors contributed to the article. BV had the lead in developing the method. Both authors codesigned the examples and contributed equally to the writing of the manuscript.

FUNDING

This study received funding from the Law School Admission Council (LSAC). Opinions and conclusions reflect those of the authors and do not necessarily reflect those of LSAC.

REFERENCES

Armstrong, R. D., Jones, D. H., and Wang, Z. (1995). Network optimization in constrained standardized test construction. *Appl. Manag. Sci.* 8, 189–212.

Baghaei, P., and Ravand, H. (2016). Modeling local item dependence in cloze and reading comprehension test items using testlet response theory. *Psicologica* 37, 85–104.

Bertsimas, D., and Sim, M. (2003). Robust discrete optimization and network flows. *Math. Programm.* 98, 49–71. doi:10.1007/s10107-003-0396-4

Boekkooi-Timminga, E., and van der Linden, W. J. (1989). A maximin model for IRT-based test design with practical constraints. *Psychometrika* 54, 237–247. doi:10.1007/BF02294518

Braeken, J. (2011). A boundary mixture approach to violations of conditional independence. *Psychometrika* 76, 57–76. doi:10.1007/s11336-010-9190-4

Chang, Y., and Wang, J. (2010). "Examining testlet effects on the PIRLS 2006 assessment," in *Paper Presented at the 4th IEA International Research Conference* (Gothenburg, Sweden).

De Jong, M. G., Steenkamp, J. B. E. M., and Veldkamp, B. P. (2009). A model for the construction of country-specific, yet internationally comparable short-form marketing scales. *Market. Sci.* 28, 674–689. doi:10.1287/mksc.1080.0439

Glas, C. A. W. (2012a). "Estimating and testing the extended testlet model," in *LSAC Research Reports Series (LR 12-03)* (Newtown, PA: Law School Admission Council).

Glas, C. A. W. (2012b). "Fit to testlet models and differential testlet functioning," in *LSAC Research Reports Series (LR 12-07)* (Newtown, PA: Law School Admission Council).

Glas, C. A. W., Wainer, H., and Bradlow, E. T. (2000). "MML and EAP estimation in testlet-based adaptive testing," in *Computer Adaptive Testing: Theory and Practice*, eds W. J. van der Linden and C. A. W. Glas (Dordrecht, Netherlands: Kluwer), 271–288.

Gurobi Optimization. (2016). *The Gurobi Optimizer*. Available at: http://www.gurobi.com

Harter, L. H. (1960). Tables of range and studentized range. *Ann. Math. Statist.* 31, 112–1147. doi:10.1214/aoms/1177705684

Huitzing, H. A., Veldkamp, B. P., and Verschoor, A. J. (2005). Infeasibility in automated test assembly models: a comparison study of different models. *J. Educ. Measure.* 42, 223–243. doi:10.1111/j.1745-3984.2005.00012.x

IBM. (2016). *IBM ILOG CPLEX Optimizer*. Available at: https://www-01.ibm.com/software/commerce/optimization/cplex-optimizer/.

Junker, B. W. (1991). Essential independence and likelihood-based ability estimation for polytomous items. *Psychometrika* 56, 255–278. doi:10.1007/BF02294462

Koziol, N. A. (2016). Parameter recovery and classification accuracy under conditions of testlet dependency: a comparison of the traditional 2PL, testlet, and bi-factor models. *Appl. Measure. Educ.* 29, 184–195. doi:10.1080/08957347.2016.1171767

Law School Admission Council. (2010). *The official LSAT Handbook*. Newtown, PA: Law School Admission Council.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *J. Educ. Measure.* 14, 117–138. doi:10.1111/j.1745-3984.1977.tb00032.x

Luecht, R. M. (1998). Computer-assisted test assembly using optimization heuristics. *Appl. Psychol. Measure.* 22, 224–236. doi:10.1177/01466216980223003

Paap, M. C. S., Glas, C. A. W., and Veldkamp, B. P. (2013). "An overview of research on the testlet effect: associated features, implications for test assembly, and the impact of model choice on ability estimates," in *LSAC Research Reports Series (LR 13-03)* (Newtown, PA: Law School Admission Council).

Paap, M. C. S., He, Q., and Veldkamp, B. P. (2015). Selecting testlet features with predictive value for the testlet effect: an empirical study. *Sage Open* 5, 12. doi:10.1177/2158244015581860

R Development Core Team. (2012). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Soyster, A. L. (1973). Convex programming with set-inclusive constraints and applications to inexact linear programming. *Oper. Res.* 21, 1154–1157. doi:10.1287/opre.21.5.1154

Swanson, L., and Stocking, M. L. (1993). A model and heuristic for solving very large item selection problems. *Appl. Psychol. Measure.* 17, 151–166. doi:10.1177/014662169301700205

Tippett, L. H. C. (1925). On the extreme individuals and the range of samples taken from a normal population. *Biometrika* 17, 364–387. doi:10.1093/biomet/17.3-4.364

van der Linden, W. J. (2005). *Linear Models for Optimal Test Design*. New York: Springer Verlag.

Veldkamp, B. P. (1999). Multi-objective test assembly problems. *J. Educ. Measure.* 36, 253–266. doi:10.1111/j.1745-3984.1999.tb00557.x

Veldkamp, B. P. (2002). Multidimensional constrained test assembly. *Appl. Psychol. Measure.* 26, 133–146. doi:10.1177/01421602026002002

Veldkamp, B. P. (2013). Application of robust optimization to automated test assembly. *Ann. Oper. Res.* 206, 595–610. doi:10.1007/s10479-012-1218-y

Veldkamp, B. P., Matteucci, M., and de Jong, M. (2013). Uncertainties in the item parameter estimates and automated test assembly. *Appl. Psychol. Measure.* 37, 123–139. doi:10.1177/0146621612469825

Verschoor, A. J. (2007). *Genetic Algorithms for Automated Test Assembly*. Unpublished doctoral thesis, University of Twente, Enschede, The Netherlands.

Wainer, H., Bradlow, E. T., and Wang, X. (2007). *Testlet Response Theory and Its Applications*. New York: Cambridge University Press.

Wainer, H., and Kiely, G. (1987). Item clusters and computerized adaptive testing: a case for testlets. *J. Educ. Measure.* 24, 185–202. doi:10.1111/j.1745-3984.1987.tb00274.x

Wainer, H., and Wang, X. (2000). Using a new statistical model for testlets to score TOEFL. *J. Educ. Measure.* 37, 203–220. doi:10.1111/j.1745-3984.2000.tb01083.x

Wang, X., Bradlow, E. T., and Wainer, H. (2002). A general Bayesian model for testlets: theory and applications. *Appl. Psychol. Measure.* 26, 109–128. doi:10.1177/0146621602026001007

Williams, H. P. (1999). *Model Building in Mathematical Programming*, 4th Edn. Somerset, Great Britain: John Wiley & Sons Ltd.

Conflict of Interest Statement: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.