Application of Sequential Methods to Testlet Based Educational Testings

Chang, Yuan-chin Ivan
Academia Sinica, Institute of Statistical Science
128, Academia Road Section 2
Taipei 11529, Taiwan
E-mail: ycchang@sinica.edu.tw

The terminology – testlet in educational/psychological testings refers to a group of test items which are administered together to test takers, and often used in, for example, a comprehensive test. Therefore, unlike those in the item response theory, the items in a testlet are usually correlated, since they may refer to a passage of a paragraph or share a common content. There are several models used to model testlest response data. Among them, a modified logistic model below, originally used in the item response theory, is commonly used for modeling the dichotomous responses to testlets:

$$(1) \qquad P(Y_{ij} = 1) = c_j + (1 - c_j)\text{logit}^{-1}(t_{ij}),$$

$$(2) \qquad t_{ij} = a_j(\theta_i - b_j - r_{i,k(j)})$$

where Y_{ij} denotes the response of a test-taker with the ability θ_i to the item j, parameters a_j, b_j and c_j retain their original interpretation as in the item responser theory, and the newly introduced parameter $r_{i,k(j)}$ is the parameter of *testlet effect* of item j with person i that is nested within testlet k. Note that if $r_{i,k(j)} = 0$ for all i, j and k, then (1) becomes the classical 3-parameter logistic model used in the item response theory.

Suppose those item parameters are known in advance, and our goal is to estimate the ability levels, θ's, of test takers. Then for this kind of the correlated binary responses data, the method of the generalized estimating equation (GEE) (Liang and Zeger, 1986) can be used. In addition, as in the variable length computerized adaptive testing, it is of interest to know how many testlets used will suffice to obtain an estimate of θ with a satisfactory (prescribed) accuracy. In this study, we use a fixed width confidence interval to manage the accuracy of estimation of θ, and a sequential method is employed such that the test is stopped as long as the prescribed accuracy for the estimate is reached.

Assume the exchangeable correlations among items within a testlet, and items from different testlets are mutually independent. Thus, GEE with a "diagonal block" working covariance matrix is used. For a given test-taker i, let $\text{corr}(Y_{ik(j)}, Y_{ik(j')}) = \rho_k$, $j \neq j'$, be the correlation between responses $Y_{ik(j)}, Y_{ik(j')}$ to items j, j', respectively, in the k-th testlet. Then correlation matrix for observations within a testlet is equal to

$$\begin{bmatrix} 1 & \rho_k & \cdots & \rho_k \\ \rho_k & 1 & \cdots & \rho_k \\ \vdots & \vdots & \ddots & \vdots \\ \rho_k & \rho_k & \cdots & 1 \end{bmatrix},$$

and set the working correlation matrix $V_i = (A_i)^{1/2}R_i(\rho)(A_i)^{1/2}$, where A_i is a $n_i \times n_i$, ($n_i = \sum_{h=1}^{k_i} J_h, k_i \leq K$), diagonal matrix with $P_{ik(j)} \times (1 - P_{ik(j)})$ as its $k(j)$th diagonal element, and $R_i(\rho)$ is a $n_i \times n_i$ symmetric matrix as mentioned above. Let $\hat{\theta}_i^{(k)}$ be the estimate of θ_i after taking k testlets. Then it follows that

$$(3) \qquad \hat{\theta}_i^{(k)} = \theta_0 + (\sum_k D'_{ik}V_{ik}^{-1}D_{ik})^{-1}\sum_k D'_{ik}V_{ik}^{-1}(Y_{ik} - P_{ik}),$$

Table 1: The average estimates(standard deviation) and mean square error with the fixed width confidence interval $d = 0.3$, and the target coverage frequency 0.95 ($\alpha = 0.05$)

c	θ	$\hat{\theta}$	MSE of θ	Coverage prob.	No.of testlets
0.1	-2	-1.995(.147)	0.022	0.970	30.48(3.05)
	-1.5	-1.493(.149)	0.022	0.946	27.38(2.50)
	-1	-0.995(.150)	0.022	0.964	26.13(2.14)
	-0.5	-0.487(.155)	0.024	0.950	25.29(1.94)
	0	0.002(.150)	0.023	0.952	25.28(1.97)
	0.5	0.500(.149)	0.022	0.948	25.16(1.95)
	1	1.011(.148)	0.022	0.950	25.19(1.93)
	1.5	1.506(.149)	0.022	0.952	25.80(2.07)
	2	2.003(.151)	0.023	0.952	27.06(2.33)

where $D_{ik} = (\partial P_{ik(1)}/\partial\theta, \ldots, \partial P_{ik(j)}/\partial\theta, \ldots, \partial P_{ik(n)}/\partial\theta)'$. Hence, the estimate of θ is obtained through an iterative algorithm. It is shown that the estimate of θ is asymptotically normally distributed. Based on the asymptotic normality of estimate of θ, the stopping rule for building a two-sided confidence interval with its width no greater than $2d$ and coverage probability no less than $1-\alpha$ is discussed below. (Note that $R_i(\rho) = I$ is equivalent to assuming no correlation within a testlet. For other correlation structures used for other applications in literature, please see, for example, Fitzmaurice, et al. (1993) and Zorn (2001).)

Stopping rule

Suppose we require the width of a $1-\alpha$ confidence interval of θ is no greater than $2d$; that is, we require that $Pr(\theta_i \in [\hat{\theta}_i^{k_i} - d, \hat{\theta}_i^{k_i} + d]) \geq 1-\alpha$, Therefore, the stopping rule for constructing such a confidence interval of θ_i is $T(\theta_i) = \inf\{k_i \geq 1 : \hat{\Sigma}_i^{(k_i)} \leq (d/Z_{\alpha/2})^2\}$, where k_i is the number of testlets. Note that instead assigning a new item each time in the variable length computerized adaptive testing, here we administer a new testlet at a time.

Numerical Study

In following simulation study, the discrimination parameter a is chosen randomly from $U(0.5, 2.5)$, the difficult parameter b is chosen randomly from $U(-3, 3)$ and the guessing parameter $c = 0.1$. The target coverage probability is 0.95 ($\alpha = 0.05$) with $d = 0.3$. We consider the case with $r_{i,k(j)} = 0$ first. Table 1 shows the averages of estimates and their corresponding standard deviations for different θ's based on 500 runs. The number of testlets used to estimate θ is about 26 with 10 items in each testlet. Table 2 shows results for model with $r_{i,k(j)} = -0.1$. It can be seen from this table that due to the testlet effect, the number of testlets used in this case is much larger than that of the case with $r_{i,k(j)} = 0$, while the coverage probabilities are still lower than their counterparts.

Mastery Testing with Group Sequential Method

In some testing situation, the qualifications of test takers, instead of their exact ability levels, are of interest. This kind of a test is usually referred as a mastery/criterion reference test in educational/psychological testing. Thus, by treating each testlet as a group, we employ a group sequential hypotheses testing method to it. Parapallona and Tsiasts (1994), as an extension of Emerson and Fleming (1989), proposed a asymmetric test with unequal type I and II errors (see also Jennison and Turnbulls (2000)). Furthermore, Lee, et al. (1996) proved that the method of group sequential can be

Table 2: The average estimates(standard deviation) and mean square error with $\gamma = -0.1$, the fixed width confidence interval $d = 0.3$, and the target coverage frequency 0.95 ($\alpha = 0.05$)

c	θ	$\hat{\theta}$	MSE of θ	Coverage prob.	No.of testlets
0.1	-2	-1.892(.134)	0.030	0.926	106.10(49.71)
	-1.5	-1.415(.171)	0.036	0.904	56.86(32.28)
	-1	-0.958(.198)	0.041	0.866	36.81(21.91)
	-0.5	-0.462(.192)	0.038	0.896	31.32(15.87)
	0	-0.002(.201)	0.040	0.884	30.51(13.76)
	0.5	0.440(.211)	0.048	0.834	30.87(15.41)
	1	0.934(.207)	0.047	0.840	30.50(14.07)
	1.5	1.433(.193)	0.042	0.878	30.60(14.95)
	2	1.910(.179)	0.040	0.860	30.24(13.80)

applied to longitudinal data under some conditions on their correlation structure (see also Gange and DeMets (1999)). We assume items in different testlets are mutually independent as before. Thus the structure of correlation is assumed to be known in advance. The assumptions in Lee, et al. (1996) are proved to be satisfied, asymptotically, and the group sequential method is therefore applicable here.

Suppose that the threshold of a mastery testing is $\theta = 0$; that is our goal is to test whether θ_i's is greater than 0; that is, to test $H_0 : \theta_i \leq 0$ *vs.* $H_1 : \theta_i > 0$.

Lemma 1. *Suppose* $\{Z_1, \cdots, Z_K\}$ *is a sequence of test statistics observed at* K *analyses for a group sequential study, which has a canonical joint distribution with information levels* $\{I_1, \cdots, I_K\}$ *for the parameter* θ*. If (i)* $\{Z_1, \cdots, Z_K\}$ *is multivariate normal, (ii)* $E(Z_k) = \theta\sqrt{I_\theta}, k = 1, \cdots, K$*, (iii)* $Cov(Z_{k_1}, Z_{k_2}) = \sqrt{I_{k_1}/I_{k_2}}, 1 \leq k_1 \leq k_2 \leq K$*; that is,* $Cov(\hat{\theta}_{k_1}, \hat{\theta}_{k_2}) = I_{k_2}^{-1}$*, (iv)*$I_k = (k/K)I_K, k = 1, \cdots, K$*. Then, for testing* $H_0 : \theta = 0$ *vs.* $H_1 : \theta > 0$ *with Type I error probability* α *and power* $1 - \beta$ *at* $\theta = \delta$*, a general one-sided group sequential test is defined by pairs of constants* (s_k, t_k) *with* $s_k < t_k$ *for* $k = 1, \cdots, K - 1$ *and* $s_K = t_K$*, and the group sequential procedure is described below:*

(i) For $k = 1, \cdots, K - 1$*, if* $Z_k \geq t_k$*, then stop samping and reject* H_0*; if* $Z_k \leq s_k$ *stop sampling and accept* H_0*; otherwise administer testlet k+1.*

(ii) When $k = K$*, if* $Z_K \geq t_K$ *then stop the test and reject* H_0*; if* $Z_K < s_K$ *then stop the test and accept* H_0*.*

(Note that we set $s_K = t_K$ to ensure that the test will be terminated at analysis K.) For $k = 1, \cdots, K$, the critical value with parameter Δ are $t_k = \widetilde{C}_1(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2}$ and $s_k = \delta\sqrt{I_k} - \widetilde{C}_2(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2}$. The constants $\widetilde{C}_1(K, \alpha, \beta, \Delta)$ and $\widetilde{C}_2(K, \alpha, \beta, \Delta)$, which do not depend on δ, are chosen to ensure the Type I error and power conditions. Wang and Tsiatis (1987) proposed boundaries of different shapes, via different Δ's, for a family of two-sided and one-sided tests. Figure 1 is the boundaries of power family tests with four testlet of observations at $\alpha = \beta = 0.05$ and $\Delta = -0.5, -0.25, 0, 0.25$. As Δ goes large, the range of boundary becomes narrow. Wang & Tsiatis tests includes Pocock and O'Brien & Fleming tests as special cases (When $\Delta = 0.5$, it is Pocock's test, while $\Delta = 0$, it becomes O'Brien & Fleming's test). In order to have $s_K = t_K$, the final information level must be

$$I_K = \left\{ \left(\widetilde{C}_1(K, \alpha, \beta, \Delta) + \widetilde{C}_2(K, \alpha, \beta, \Delta) \right)^2 \right\} / \delta^2.$$

Then, s_k and t_k, $k = 1, \ldots, K$, are $t_k = \widetilde{C}_1(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2}$, and

$$s_k = \left(\widetilde{C}_1(K, \alpha, \beta, \Delta) + \widetilde{C}_2(K, \alpha, \beta, \Delta) \right) (k/K)^{1/2} - \widetilde{C}_2(K, \alpha, \beta, \Delta)(k/K)^{\Delta-1/2}.$$

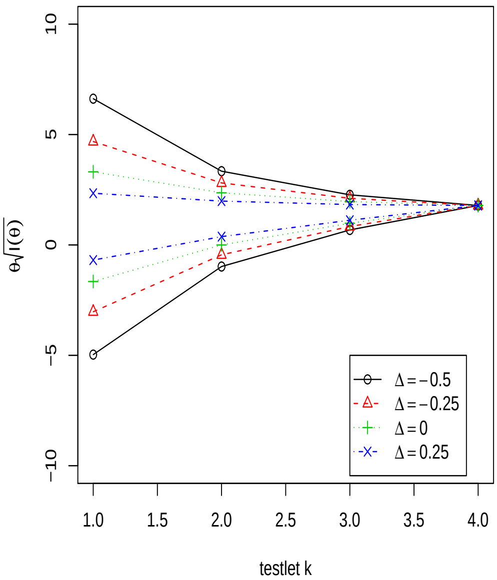

Figure 1: Boundary of one sided test

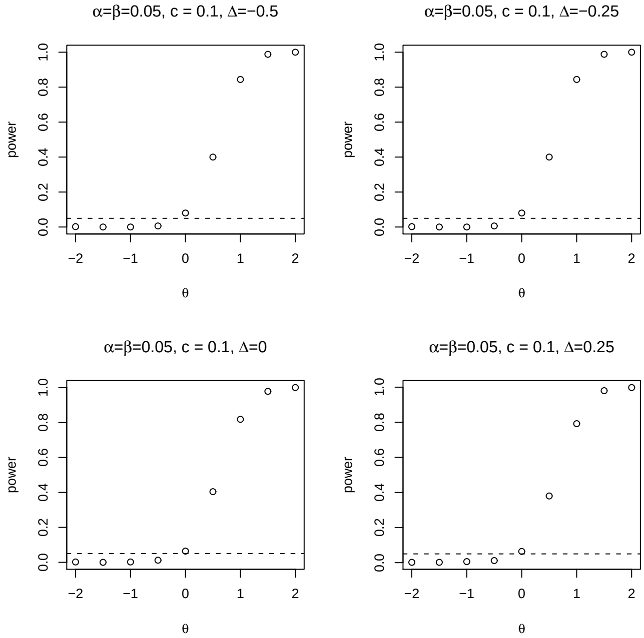

Figure 2: Power functions

Figure 2 shows the power function at different θ's, and numerical study results are summarized in Table 3 below.

Highlights of proofs

Comparing with some ordinary longitudinal studies, where correlation structure are usually unknown, the advantage of our problem is that the correlation structure is known, which makes applications of GEE and the group sequential method to this problem easier. The proof of asymptotic properties of sequential estimation under GEE can be done through the martingale convergence theorems and martingale central limit theorem with respect to a suitable filtration, which can be found in Chow and Teicher (1997) and Pollard (1984). Similar arguments are used in Chang and Ying (2004). The large sample justification for applying the group sequential method to the testlet based mastery testing follows from Lee, et. al (1998), and a similar technique has been employed in Park and Chang (2010).

Summarization and Future Study

The idea of using GEE to deal with clustered/stratified data is classic. In GEE, the correlation is usually assumed to be independent of the unknown regression parameters of interest, while in the current problem the correlation is a function of test-taker's abilities. Current approach takes the advantage of working covariance of GEE under the scenario that the correlation structure is known, and used all test takers to estimate the working covariance matrix as in the ordinary GEE, while in our problem, test takers with different abilities share only the same correlation structure, but not correlation values. Hence, the estimate of correlation is biased. Although, asymptotic properties are obtained as in the ordinary GEE, the current procedures cannot be efficient. To improve the performance of current approach, the traditional approach has to be modified.

As described in (2), the correlations of items within a testlet also depend on the test takers' ability level θ, even for a same testlet. Hence, as in Liang and Zeger (1986), in order to estimate

Table 3: Average Power and Frequency of Terminational testlet in Study 1 at at $c = 0.1$

$\Delta = -0.5$

θ	power	terminal testlet 1	2	3	4	total
-2	0.002	1	485	14	0	500
-1.5	0.000	0	461	39	0	500
-1	0.000	0	389	106	5	500
-0.5	0.006	0	240	235	25	500
0	0.080	0	78	293	129	500
0.5	0.400	0	24	207	269	500
1	0.844	0	50	254	196	500
1.5	0.988	0	154	290	56	500
2	1.000	0	305	189	6	500

$\Delta = -0.25$

θ	power	terminal testlet 1	2	3	4	total
-2	0.002	56	442	2	0	500
-1.5	0.000	27	465	8	0	500
-1	0.000	11	453	34	2	500
-0.5	0.000	2	354	134	10	500
0	0.064	0	171	247	82	500
0.5	0.394	0	62	223	215	500
1	0.858	0	96	251	153	500
1.5	0.988	0	255	207	38	500
2	1.000	1	380	112	7	500

$\Delta = 0$

θ	power	terminal testlet 1	2	3	4	total
-2	0.002	307	193	0	0	500
-1.5	0.000	269	230	1	0	500
-1	0.002	165	318	16	1	500
-0.5	0.012	71	339	84	6	500
0	0.064	27	241	168	64	500
0.5	0.404	5	136	211	148	500
1	0.818	12	173	197	118	500
1.5	0.978	22	302	139	37	500
2	1.000	45	391	61	3	500

$\Delta = 0.25$

θ	power	terminal testlet 1	2	3	4	total
-2	0.002	435	65	0	0	500
-1.5	0.002	408	91	1	0	500
-1	0.006	363	131	5	1	500
-0.5	0.012	257	213	26	4	500
0	0.064	138	224	106	32	500
0.5	0.380	54	167	192	87	500
1	0.792	64	190	171	75	500
1.5	0.980	123	264	85	28	500
2	0.998	183	278	34	5	500

the correlations of items within individual testlet, we need responses from independent test takers with *the same ability level*. However, θ's are the unknown parameters to be estimated. So, based on their ability levels, to group test takers in advance is not possible. Hence, to resolve this obstacle, we propose a multiple stages recursive estimating scheme, which is briefly described below and its results of this method will be reported elsewhere. Similar idea can be applied to mastery testing.

Multiple-Step Estimation Procedure:

(i) *Initialization:* Estimate the abilities of test-takers with "working covariance matrix" equal to an identity matrix.

(ii) *Clustering:* Group test-takers based on the estimated test-takers' abilities obtained in (i).

(iii) *Covariance Matrix Estimation:* Estimate the covariance matrix within each cluster of test-takers, and re-estimate the abilities of the test-takers within in this group.

(iv) *Re-clustering:* Re-group test-takers using the abilities estimated in (iii).

(v) *Iteration:* Repeat (iii) and (iv) until the estimates of abilities become stable based on some prescribed criterion.

(vi) *Stopping Criterion for Ability Estimation:* Administer next testlet based on the estimated ability levels until the estimated abilities satisfied a prescribed accuracy level.

Some remarks:

1: In (i), we first pretend all items are independent to obtain the initial estimates of abilities. However,

This procedure will also work with other reasonable working covariance matrix.
2: We presume the number of test-takers is large enough such that there is no test-taker's ability is "isolated" and the number of test-takers in each groups is large enough for covariance matrix estimation. Moreover, as the procedure is continuing, if clustering becomes finer, then the test-taker sizes of individual groups become smaller. Hence, the size of group should be controlled such that the covariance matrix can be estimated with certain quality.
3: The stopping criterion for the iteration procedure can be based on the sum of square of the differences between two iterations. The choice of criterion depends on administration and/or the nature of a test, and should become smaller when more testlets are assigned to test-takers as the estimates of abilities become stable.
4: Moreover, no testlet selection method is discussed here. We believe that with a suitable adaptive testlet selection rule, the performance of the proposed methods can be largely improved. The adaptive testlet selection can be integrated here without any technical difficulty. The stopping criterion of individual test-takers will be enforced. That is, the test lengths of different test-takers are different.

REFERENCES (RÉFERENCES)

Chang, Y-c. I. and Ying, Z. (2004). Sequential estimation in variable length computerized adaptive testing, Journal of Statistical Planning and Inference 121, 249 - 264.

Chow, Y. S. and Teicher, (1997). *Probability Theory: Independence, Interchangeability, Martingales*, 3rd Ed., Springer, New York.

Emerson, S. and Fleming, T. (1989). Symmetric group sequential designs. Biometrics, 45:905V923.

Fitzmaurice, G.M. Laird, N. M. and Rotnitzky, A. (1993). Regression models for discrete longitudinal responses. Statistical Science, 8:284V309.

Gange, S and DeMets, D. (1999). Sequential monitoring of clinical trials with correlated responses, Biometrika 83, 157 V 167.

Jennison, C. and Turnbull, B. W. (2000). Group Sequential Methods with Applications to Clinical Trials. Chapman and Hall/CRC Press.

Lee, S. J., Kim, K., and Tsiatis, A. A. (1996). Repeated significance testing in longitudinal clinical trials. Biometrika 83, 779V789.

Liang, K. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. Biometrika 73, 13 - 22.

Pampallona, S. and Tsiatis, A. (1994). Group sequential designs for one-sided and twosided hypothesis testing with provision for early stopping in favor of the null hypothesis. Journal of statistical planning and inference, 42:19V35.

Park, E. and Chang, Y-c. I. (2010). Sequential analysis of longitudinal data in a prospective nested case-control study, Biometrics, 1034 – 1042.

Pollard, D. (1984). Convergence of Stochastic Processes. New York: Springer-Verlag.

Wainer, Bradlow and Wang (2007). Testlet response theory and its applications, Cambridge University Press.

Wang, S. and Tsiatis, A. (1987). Approximately optimal one-parameter boundaries for group sequential trails. Biometrics., 43:193V200.

Zorn, C. (2001). Generalized estimating equation models for correlated data: A review with applications. American Journal of Political Science., 45:470V490.