10-2019

An Applied Example of a Two-Tier Multiple-Group Testlet Model

Paulius Satkus
James Madison University, satkuspx@jmu.edu

Christine E. DeMars
demarsce@jmu.edu

An Applied Example of a Two-Tier Multiple-Group Testlet Model

Paulius Satkus

Christine E. DeMars

James Madison University

Satkus, P., & DeMars, C. E. (2019, October). *An Applied Example of a Two-Tier Multiple-Group Testlet Model.* Paper presented at the annual meeting of the Northeastern Educational Research Association. Rocky Hill, CT.

An Applied Example of a Two-Tier Multiple-Group Testlet Model

Appropriate application of many traditional item response theory (IRT) models are conditional on satisfying the assumptions these models make. The three most common assumptions are: appropriate model form, unidimensionality, and local independence. Appropriate model form refers to the hypothesized relationship between probability of getting an item correct and the ability level. In other words, in IRT probability of getting an item correct depends on several parameters, which determine the name of the IRT model. That is, 1PL model only has one parameter, which tells us how difficult an item is to the group of examinees. In 2Pl models, a second parameter – discrimination is estimated. Discrimination refers to how well an item can discriminate or differentiate among different levels of ability. In 3PL models, the third parameter- pseudo guessing is estimated. Pseudo guessing refers to the probability that even examinees of extremely low ability levels would have to answer any given item correctly. Stated simply, when making the appropriate model form assumption, researchers believe that the empirical data or the administered items can be modeled using one of the three models. An example when this assumption might be violated is when the administered items use a "fill in the blank" format, thus eliminating majority of correct guessing, but a 3PL IRT model is used to score them.

Second, unidimensionality may be discussed with respect to the items or with respect to the test. Starting at the item level, a unidimensional item is said to measure only one construct (e.g., addition or 9th century Viking history). In other words, only one construct (or underlying ability) is necessary to correctly answer the item. A test is a collection of related items, however some tests may measure more than one constructs, thus violating the unidimensionality

assumption. In psychological research, many instruments are often comprised of multiple subscales that may measure related, yet distinct constructs. For example, the Student Opinion Survey (SOS, Thelk, Sundre, Horst, & Finney, 2009) measures two constructs test-taking importance and test-taking effort. Scoring students' responses to this instrument using a unidimensional IRT model would not be appropriate. However, multidimensional IRT models have been developed and are available for researchers' use (Reckase, 2009).

Third, in unidimensional IRT models, local independence of items is often assumed. Local independence means that any given two items are not related, after controlling for the primary construct. That is, two items do not share any similarities, after the influence of the primary ability is "taken out". In multidimensional tests, this assumption could be extended by stating that items are not related after controlling for the two or more primary abilities.

In practice, there are several situations in which the local independence assumption may be violated. For example, consider a speeded test. As the name suggests, these tests often have strict time constrains. Consider an example. An examinee may spend too much time on the items at beginning of the test and there is not enough time to devote to the remaining items on the test. Thus, when completing the items at the end of the test, an examinee may have to rush to complete the remaining items on the test. In this situation, the responses to the remaining items would be driven by the primary ability the test was supposed to measure but also by the pressure to complete the items in time.

Differential item functioning (DIF) is another situation in which local independence could be violated. DIF is often described as a situation in which two examinees of the same ability but from different groups (e.g., ethnic or social) have a different probability of getting an item correct. Thus, group membership is related to the probability of correctly answer an item. If

several items on the test favor one group, then this violates the local independence as "getting" any two DIF items correct depends on the primary construct but also group membership. Literature on DIF is vast, as there are many legal and ethical issues that DIF brings about.

A third example of violating local independence is by using testlet items. A testlet is defined as "A series or a cluster of items based on a common stimulus" (Wainer, Bradlow, & Du, 2000). The common stimulus for the testlet could be a short passage depicting a certain situation or a scenario that examinees then have to analyze. It could also be a series of graphics, perhaps a matrix of several graphs. Anywhere from two to five or more items could be associated with that common stimulus, which together would then form a testlet. Testlets violate the assumption of local independence because after controlling for the primary ability, the items on the testlet share some dependencies from the common stimulus that the items are based on. Wainer et al., (2000) lays out a few reasons why testlets are useful in testing situations. First, one of the most attractive features of the using standard multiple-choice item format is the fact that many different items can be administered to examinees in a relatively short amount of time. It is likely one of the reasons why the multiple-choice item format has gained so much popularity. However, some researchers critique this format by pointing out the "atomistic nature of each item" (Wainer et al., 2000). The items are related to each other in a sense that they were developed to measure the same underlying construct or ability. However, each item by intention measures a different aspect of the same construct and it could be that the items are not capturing the intended construct well. Testlets connect several items and give more information about what the items within a testlet are asking. Second, testlets reduce the effects of prior knowledge. For example, if the test is designed to measure a complex construct such as critical thinking, some examinees may benefit from a scenario that they are more familiar with. Using a testlet, a more

abstract or unlikely stimulus could be presented, which would provide just enough information to complete the items. The third reason to include testlet is to increase the efficiency of items. A lot of time during a test is dedicated to reading the items and processing what is being said. By using testlets, examinees have to read the testlet stimulus only once, which serves a number of items that belong to the testlet. Thus, the time spent reading the question is reduced when the time actually answering items is increased.

Modelling Testlets and Purpose of the Study

It was reviewed above that testlets violate the local independence assumption of the traditional IRT models. One early approach to deal with violations of local independence was to model items comprising a testlets as a polytomous item. That is, say, there are five items that comprise a testlet. These five items would then be summed to create one item that could range from 0 (if an examinee failed to correctly respond to all five items) to 5 (if an examinee correctly responded to all five items. Using polytomous IRT models, we could then score this composite item. The major drawback of this approach is that information about the exact scoring pattern would be lost. That is, it would be unclear which items within the testlet the examinees got correct. A score of three could represent many different response patterns, whereas only scores of zero or five would provide pointed feedback.

Another approach that researchers have used is simply ignoring the testlets and modeling the examinee response with traditional IRT methods. Two simulation studies revealed the consequences for ignoring testlets (Bradlow, Wainer, & Wang, 1999; Wainer et al., 2000) were biased ability estimates and biased item parameters. In these studies, difficulty and pseudo-guessing parameter values were affected only slightly, however the biggest impact was seen on the discrimination parameter. Wainer et al., (2000) offered a conceptual explanation of why the

discrimination parameters suffers the most when testlets are ignored. On page 265, they state "… when there is unmodelled local dependence the model looks upon the resulting nonfit as noise and so the regression of the item on the underlying trait is, in the face of this noise, more gradual." Thus, when the local dependencies are accounted for by the testlets, the relationships between each item and the underlying trait becomes more pure and stronger. In structural equation modelling, this occurs when latent relations are considered, meaning that the measurement error is removed.

So how do testlet models account for the dependencies among certain items on the test? The following equation for the 3PL models is provided by Wainer et al., (2000):

$$P(Y_{ij} = 1) = c_j - (1 - c_j)\frac{e^{\left(a_j(\theta_i - b_j - \gamma_{id(j)})\right)}}{1 + e^{\left(a_j(\theta_i - b_j - \gamma_{id(j)})\right)}}$$

where the probability (P) to correctly (1) answer item (j) for each examinee (i) depends on item's (j) pseudo-guessing parameter (c), item's (j) discrimination parameter (a), examinee's ability level (θ_i), and the person's testlet score ($\gamma_{id(j)}$). In this model, each testlet variance is estimated as additional factor. That is, in addition to the primary factor that the test is measuring, testlet factors for each testlet are estimated. The testlet slopes for each item are set equal to slopes for the primary factors. If the test has multiple testlets, the variances of each of these testlets are estimated independent of each other allowing the comparison of testlet variances.

The purpose of the current study is to illustrate how a 3PL testlet model could be estimated using a real data example and what information the testlet model provides over and above the traditional IRT models. The item parameters estimated from a testlet model and from a regular 3PL model will then be compared.

Methods

The data used in the current study was obtained using Natural World Version 9 (NW9) test. NW9 was designed to assess students' scientific and quantitative reasoning abilities. The test was developed by university faculty as part of the general education program institutional accountability assessment. NW9 contains 66 dichotomously scored items, of which 40 items measure students' scientific ability and 26 items measure students' quantitative reasoning ability. There are 12 different testlets in NW9 that range from two to four items per testlet.

Data were collected over four assessment occasions in the spring of 2016, 2017, 2018, and 2019. The samples were combined to create an effective sample size of 2198. The subjects in the current study were university students that have completed 45-70 credits at a large southeastern public university. The 3PL IRT model and the testlet IRT model were estimated using FlexMIRT (Cai, 2012) software.

Results

Before estimating the IRT models, local independence was examined using Stout's DIMTEST 2.0 (Stout, 2005). To perform Stout's test, the sample was split using a 50/50 ratio (Socha & DeMars, 2013). According to the results, the local independence was violated: Stout's $T = 2.6778$, $p = .0037$. Note, since the NW9 contain two primary factors, a secondary dimension could be formed by one of the primary factors. However, according to Stout's test, the items that were identified as forming the secondary dimension were dispersed among the primary two factors. In other words, some other construct is influencing items, after controlling for the two primary factors. Our hypothesis is that the testlet factors are at play. Despite violating the local independence assumption, a two-factor 3PL model was estimated and parameter estimates are

displayed in table 1. Note that item 60 was eliminated from the analyses due to extreme values (i.e., estimated difficulty parameter < -20).

Next, a 3PL model with 12 testlets was estimated. The parameter estimates are presented in table 2. It is noteworthy to point out that replicating the results from the two simulation studies reviewed above, the discrimination parameter estimates are higher when testlets were modeled. Additionally, the variance estimates for each testlet are presented in table 3. The variances of the Quantitative Reasoning and Scientific Reasoning ability was set to one for scaling purposes. This allowed meaningful comparison between variances of the two primary factors and the testlets. Testlet 3 had the largest variance suggesting that for this particular testlet there was more variance after controlling for one of the two primary factors. Some testlets had relatively small variance (e.g., testlet 4 or 5) suggesting that even though the items were ed to represent a testlet, empirically strong dependencies were not found.

Next, a 3PL testlet model was estimated separately for males and females. Doing so allowed to comparison of the two groups mean scores for each primary factors (Scientific and Quantitative Reasoning), the covariance between the two factors, and whether the testlets were more salient for one group than another. Table 4 displays the summaries of the results. It appears that males scored .30 standard deviations higher than females on the Quantitative Reasoning ability factor and .38 standard deviations higher on the Scientific Reasoning ability factor. Interestingly, differences between the two groups emerged on testlet nine, which is comprised of items 47, 48, 49, and 50 (Table 4). Lastly, the relationship between Quantitative Reasoning factor and Scientific Reasoning factor was somewhat different for males (r = .90) than for females (.94), yet remained high.

Discussion

The current project served as an example of a testlet IRT application to a real dataset. Replicating the findings from two simulation studies, it was found that ignoring testlets results in biased parameter estimates. The a parameter – item discrimination suffered the most, which has direct consequences for item development, as items could be deemed poor and eliminated from the test because their discrimination was underestimated. The 3PL testlet model was fit to two groups simultaneously providing an example of multi-group application of testlet models. This allowed comparison of latent means of each primary and testlet factor between the groups. Obvious extension of this application is to add a third group or model data longitudinally, and compare whether how examinees performance on the primary and testlet factors change over time. Additionally, the multi-group application illustrated here could be conceptualized as having a categorical covariate. An extension of this example would be by modeling a continuous predictor; however, a different estimation method would need to be utilized.

References

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

Cai, L. (2012). flexMIRT: Flexible multilevel item factor analysis and test scoring [Computer software]. Seattle, WA: Vector Psychometric Group, LLC.

Stout, W. (2005). Dimtest (Version 2.0)[Computer software]. *Champaign, IL: William Stout Institute for Measurement*.

Socha, A., & DeMars, C. E. (2013). A note on specifying the guessing parameter in ATFIND and DIMTEST. *Applied Psychological Measurement*, *37*(1), 87-92.

Reckase, M. D. (2009). Multidimensional item response theory models. In *Multidimensional item response theory* (pp. 79-112). Springer, New York, NY.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for the 3PL model useful in testlet-based adaptive testing. In *Computerized adaptive testing: Theory and practice* (pp. 245-269). Springer, Dordrecht.

Table 1

Parameter estimates obtained using 3PL model

item	a1	s.e.	a2	s.e.	b	s.e.	c	s.e.
1	-	-	0.39	0.05	-0.23	0.13	0.30	0.03
2	-	-	1.11	0.15	-2.69	0.33	0.21	0.02
3	0.65	0.06	-	-	1.14	0.10	0.30	0.03
4	0.26	0.04	-	-	-0.03	0.11	0.29	0.03
5	-	-	0.61	0.05	0.77	0.09	0.29	0.03
6	-	-	0.44	0.05	0.01	0.11	0.28	0.03
7	0.88	0.07	-	-	1.61	0.10	0.28	0.03
8	0.61	0.05	-	-	1.19	0.09	0.28	0.03
9	-	-	0.73	0.07	2.78	0.13	0.28	0.03
10	1.18	0.09	-	-	0.34	0.10	0.23	0.02
11	0.60	0.06	-	-	0.25	0.10	0.27	0.03
12	0.67	0.06	-	-	1.20	0.09	0.28	0.03
13	0.54	0.05	-	-	0.27	0.10	0.29	0.03
14	-	-	1.12	0.10	3.76	0.20	0.28	0.03
15	-	-	0.63	0.07	-0.47	0.15	0.33	0.03
16	-	-	0.53	0.05	0.41	0.11	0.32	0.03
17	-	-	0.44	0.05	1.35	0.09	0.30	0.03
18	-	-	0.40	0.06	-0.71	0.16	0.30	0.03
19	-	-	0.65	0.06	0.06	0.11	0.27	0.03
20	-	-	0.48	0.05	0.69	0.09	0.28	0.03
21	0.55	0.06	-	-	-0.03	0.12	0.30	0.03
22	-	-	0.43	0.06	-0.39	0.13	0.27	0.03
23	-	-	1.00	0.09	3.34	0.17	0.28	0.03
24	-	-	0.20	0.04	0.34	0.10	0.29	0.03
25	-	-	0.26	0.05	-0.42	0.14	0.30	0.03
26	-	-	0.25	0.04	0.07	0.11	0.30	0.03
27	-	-	0.59	0.06	0.21	0.10	0.27	0.03
28	-	-	0.52	0.05	0.93	0.09	0.27	0.03
29	-	-	0.82	0.06	1.51	0.10	0.29	0.03
30	0.64	0.06	-	-	0.67	0.10	0.29	0.03
31	0.82	0.06	-	-	2.21	0.09	0.12	0.02
32	0.67	0.05	-	-	0.16	0.07	0.11	0.02
33	0.70	0.05	-	-	1.10	0.07	0.12	0.02
34	0.54	0.05	-	-	0.85	0.09	0.27	0.03
35	0.75	0.06	-	-	0.19	0.10	0.25	0.03
36	0.69	0.06	-	-	0.30	0.11	0.29	0.03
37	0.87	0.06	-	-	1.51	0.09	0.26	0.03
38	-	-	0.70	0.06	1.63	0.10	0.30	0.03

39	-	-	0.33	0.05	-0.34	0.14	0.31	0.03
40	-	-	0.23	0.04	0.67	0.09	0.31	0.03
41	-	-	0.66	0.06	0.77	0.11	0.31	0.04
42	-	-	0.79	0.07	0.39	0.12	0.31	0.04
43	-	-	0.73	0.08	-0.65	0.17	0.27	0.03
44	-	-	1.30	0.10	1.77	0.11	0.27	0.03
45	-	-	1.02	0.08	1.08	0.10	0.27	0.03
46	-	-	0.81	0.08	-0.45	0.13	0.27	0.03
47	-	-	1.25	0.09	2.91	0.14	0.26	0.03
48	-	-	0.90	0.07	1.75	0.10	0.27	0.03
49	-	-	0.87	0.07	2.23	0.11	0.26	0.03
50	-	-	1.13	0.08	2.30	0.12	0.27	0.03
51	0.71	0.05	-	-	1.01	0.07	0.11	0.02
52	0.76	0.05	-	-	1.68	0.08	0.12	0.02
53	0.55	0.07	-	-	-1.44	0.16	0.12	0.02
54	-	-	1.04	0.08	1.83	0.10	0.27	0.03
55	-	-	0.50	0.05	0.80	0.10	0.31	0.03
56	-	-	0.87	0.07	2.26	0.11	0.30	0.03
57	-	-	1.20	0.10	3.47	0.18	0.29	0.03
58	0.69	0.06	-	-	0.78	0.09	0.27	0.03
59	0.47	0.05	-	-	1.45	0.09	0.30	0.03
61	0.81	0.10	-	-	-1.56	0.22	0.25	0.02
62	0.84	0.07	-	-	0.62	0.10	0.27	0.03
63	0.54	0.10	-	-	-1.48	0.25	0.28	0.03
64	-	-	0.95	0.07	0.77	0.10	0.28	0.03
65	-	-	0.73	0.07	-0.22	0.12	0.27	0.03
66	-	-	0.58	0.06	0.04	0.11	0.25	0.03

Note. a1 refers to discrimination parameter for items measuring Quantitative Reasoning ability, whereas a2 refers to discrimination parameter for items measuring Scientific Reasoning ability

Table 2

Parameter estimates obtained using 3PL model

Item	a_1	s.e.	a_2	s.e.	testlet #	$a_{testlet}$	s.e.	b	s.e.	c	s.e.
1	-	-	0.51	0.07		-	-	-0.18	0.12	0.30	0.03
2	-	-	1.45	0.21		-	-	-2.53	0.33	0.20	0.02
3	0.87	0.08	-	-		-	-	1.15	0.09	0.30	0.03
4	0.33	0.06	-	-		-	-	-0.01	0.11	0.29	0.03
5	-	-	0.90	0.07	1	0.90	0.07	0.87	0.09	0.28	0.03
6	-	-	0.61	0.06	1	0.61	0.06	0.04	0.11	0.27	0.03
7	1.16	0.09	-	-		-	-	1.61	0.10	0.28	0.03
8	0.80	0.07	-	-		-	-	1.19	0.09	0.28	0.03
9	-	-	0.98	0.10		-	-	2.77	0.13	0.28	0.03
10	1.78	0.13	-	-	2	1.78	0.13	0.48	0.10	0.22	0.02
11	0.85	0.07	-	-	2	0.85	0.07	0.31	0.10	0.26	0.03
12	0.93	0.07	-	-	2	0.93	0.07	1.32	0.09	0.27	0.03
13	0.70	0.06	-	-	2	0.70	0.06	0.31	0.10	0.28	0.03
14	-	-	1.51	0.14		-	-	3.73	0.20	0.28	0.03
15	-	-	1.89	0.15	3	1.89	0.15	-0.16	0.13	0.24	0.02
16	-	-	2.03	0.17	3	2.03	0.17	1.12	0.11	0.30	0.02
17	-	-	0.59	0.07		-	-	1.36	0.09	0.30	0.03
18	-	-	0.51	0.08		-	-	-0.64	0.15	0.29	0.03
19	-	-	0.90	0.08		-	-	0.09	0.11	0.27	0.03
20	-	-	0.66	0.07		-	-	0.71	0.09	0.28	0.03
21	0.73	0.07	-	-	4	0.73	0.07	0.01	0.12	0.29	0.03
22	-	-	0.58	0.07	4	0.58	0.07	-0.36	0.13	0.27	0.03
23	-	-	1.31	0.12		-	-	3.29	0.16	0.28	0.03
24	-	-	0.26	0.05	5	0.26	0.05	0.35	0.10	0.29	0.03
25	-	-	0.33	0.06	5	0.33	0.06	-0.39	0.13	0.29	0.03

26	-	-	0.33	0.06	5	0.33	0.06	0.09	0.11	0.29	0.03
27	-	-	0.81	0.08		-	-	0.23	0.10	0.27	0.03
28	-	-	0.71	0.07		-	-	0.95	0.09	0.27	0.03
29	-	-	1.12	0.09		-	-	1.53	0.10	0.28	0.03
30	0.84	0.07	-	-		-	-	0.69	0.09	0.28	0.03
31	1.38	0.09	-	-	6	1.38	0.09	3.13	0.15	0.12	0.02
32	1.44	0.08	-	-	6	1.44	0.08	0.37	0.08	0.09	0.02
33	1.03	0.06	-	-	6	1.03	0.06	1.37	0.08	0.12	0.02
34	0.73	0.07	-	-		-	-	0.86	0.09	0.27	0.03
35	1.05	0.09	-	-		-	-	0.22	0.10	0.25	0.03
36	0.95	0.08	-	-		-	-	0.33	0.10	0.29	0.03
37	1.17	0.09	-	-		-	-	1.52	0.09	0.26	0.03
38	-	-	1.26	0.09	7	1.26	0.09	2.29	0.13	0.29	0.03
39	-	-	0.56	0.06	7	0.56	0.06	-0.37	0.14	0.31	0.03
40	-	-	0.45	0.04	7	0.45	0.04	0.73	0.10	0.31	0.03
41	-	-	0.90	0.08		-	-	0.79	0.11	0.30	0.04
42	-	-	1.10	0.10		-	-	0.42	0.12	0.30	0.04
43	-	-	0.96	0.11		-	-	-0.57	0.16	0.26	0.03
44	-	-	1.80	0.14		-	-	1.78	0.11	0.26	0.03
45	-	-	1.68	0.12	8	1.68	0.12	1.44	0.10	0.25	0.03
46	-	-	1.19	0.10	8	1.19	0.10	-0.39	0.13	0.25	0.02
47	-	-	1.83	0.14	9	1.83	0.14	3.62	0.20	0.25	0.03
48	-	-	1.33	0.10	9	1.33	0.10	2.09	0.11	0.27	0.03
49	-	-	1.27	0.09	9	1.27	0.09	2.62	0.13	0.26	0.03
50	-	-	1.53	0.11	9	1.53	0.11	2.67	0.14	0.27	0.03
51	1.28	0.07	-	-	10	1.28	0.07	1.40	0.08	0.11	0.02
52	1.34	0.08	-	-	10	1.34	0.08	2.30	0.11	0.12	0.02
53	0.60	0.08	-	-	10	0.60	0.08	-1.61	0.18	0.14	0.02
54	-	-	1.46	0.11		-	-	1.85	0.10	0.27	0.03

55	-	-	0.91	0.07	11	0.91	0.07	1.07	0.10	0.29	0.03
56	-	-	2.08	0.19	11	2.08	0.19	4.24	0.28	0.27	0.03
57	-	-	1.55	0.14	11	1.55	0.14	4.41	0.27	0.28	0.03
58	0.93	0.08	-	-		-	-	0.79	0.09	0.27	0.03
59	0.63	0.07	-	-		-	-	1.46	0.09	0.30	0.03
61	1.06	0.14	-	-		-	-	-1.47	0.21	0.24	0.02
62	1.17	0.09	-	-		-	-	0.64	0.10	0.27	0.03
63	0.66	0.12	-	-		-	-	-1.37	0.23	0.27	0.03
64	-	-	1.48	0.11	12	1.48	0.11	0.98	0.10	0.27	0.03
65	-	-	0.99	0.08	12	0.99	0.08	-0.17	0.12	0.26	0.03
66	-	-	0.85	0.07	12	0.85	0.07	0.07	0.11	0.24	0.03

Note. a_1 *refers to discrimination parameter for items measuring Quantitative Reasoning ability, whereas* a_2 *refers to discrimination parameter for items measuring Scientific Reasoning ability. The* $a_{testlet}$ *refers to the discrimination parameter for the items within one testlet.*

Table 3
Estimated variances for primary and testlet factors

	variance	s.e.
Quantitative Reasoning	1	-
Scientific Reasoning	1	-
testlet 1	0.58	0.02
testlet 2	0.52	0.02
testlet 3	2.25	0.07
testlet 4	0.14	0.00
testlet 5	0.00	0.00
testlet 6	1.79	0.05
testlet 7	1.63	0.05
testlet 8	0.63	0.02
testlet 9	0.81	0.02
testlet 10	1.36	0.04
testlet 11	1.63	0.05
testlet 12	0.64	0.02

Note. The correlation between the two main factors was .90.

Table 4

Latent means for females and males

	females	males	s.e.
Quantitative Reasoning	0	0.30	0.01
Scientific Reasoning	0	0.38	0.03
testlet 1	0	-0.02	0.02
testlet 2	0	0.05	0.03
testlet 3	0	0.01	0.05
testlet 4	0	-0.04	0.01
testlet 5	0	0.00	0.00
testlet 6	0	0.01	0.05
testlet 7	0	0.01	0.05
testlet 8	0	-0.07	0.02
testlet 9	0	0.19	0.03
testlet 10	0	0.00	0.04
testlet 11	0	-0.01	0.04
testlet 12	0	-0.05	0.03

Note. The means of females were set to 0 for identification purposes.