KURAM VE UYGULAMADA EĞİTİM BİLİMLERİ EDUCATIONAL SCIENCES: THEORY & PRACTICE

Received: April 14, 2016
Revision received: October 2, 2016
Accepted: December 1, 2016
OnlineFirst: December 25, 2016

www.estp.com.tr
DOI 10.12738/estp.2017.1.0270 • February 2017 • 17(1) • 321–335

Research Article

The Effects of Test Length and Sample Size on Item Parameters in Item Response Theory*

Alper Şahin[1]
Middle East Technical University
Northern Cyprus Campus

Duygu Anıl[2]
Hacettepe University

Abstract

This study investigates the effects of sample size and test length on item-parameter estimation in test development utilizing three unidimensional dichotomous models of item response theory (IRT). For this purpose, a real language test comprised of 50 items was administered to 6,288 students. Data from this test was used to obtain data sets of three test lengths (10, 20, and 30 items) and nine different sample sizes (150, 250, 350, 500, 750, 1,000, 2,000, 3,000 and 5,000 examinees). These data sets were then used to create various research conditions in which test length, sample size, and IRT model variables were manipulated to investigate item parameter estimation accuracy under different conditions. The results suggest that rather than sample size or test length, the combination of these two variables is important and samples of 150, 250, 350, 500, and 750 examinees can be used to estimate item parameters accurately in three unidimensional dichotomous IRT models, depending on test length and model employed.

Keywords

Small samples in IRT • Item parameter estimation • Item response theory • Language testing • Short tests

* This paper is a part of the first author's PhD dissertation at the department of Educational Measurement and Evaluation, Institute of Social Sciences, Hacettepe University, Ankara, Turkey.

An earlier version of this paper was presented at the 77th Annual meeting of National Council on Measurement in Education, Chicago, April 15-19, 2015.

1 **Correspondence to:** Alper Şahin (PhD), Modern Languages Program, School of Foreign Languages, Middle East Technical University Northern Cyprus Campus, Kalkanlı, Güzelyurt, Mersin 99738 Turkey. Email: alpersahin2@yahoo.com

2 Department of Educational Measurement and Evaluation, Faculty of Education, Hacettepe University, Ankara Turkey. Email: aduygu@hacettepe.edu.tr

Citation: Şahin, A., & Anıl, D. (2017). The effects of test length and sample size on item parameters in item response theory. *Educational Sciences: Theory & Practice*, *17*, 321–335. http://dx.doi.org/10.12738/estp.2017.1.0270

Developed following controversies over the intelligence testing movement (Baker, 1992), classical test theory has successfully served its practitioners in terms of test development and interpretation of test results for decades (Embretson & Reise, 2000; Hambleton, Swaminathan, & Rogers, 1991). However, the need for a theory that can predict examinees' responses to any item, even if items have never been seen by the examinees before (Lord, 1980), resulted in the development of a new test theory whose basic concepts were developed over seven and a half years (Baker, 1992). After spending a long time searching for an agreed-upon name, item response theory (IRT) became the contemporary theoretical foundation for measurement (Embretson & Reise, 2000). It is now widely used by test publishers, as well as educational, military, and industrial institutions, in their research on test development, test equating, and detecting differentially functioning items (Hambleton et al., 1991). However, despite the advantages IRT offers, a major obstacle to its use in test development is its demanding requirement for calibration sample size. More specifically, while developing a test, "the difficulty lies in the need to assess the properties of items by trying them out on a sample of subjects" (Woods & Baker, 1985, p. 117). IRT requires this sample size to be large (around 1,000) in order to obtain accurate item-parameter estimates (Hambleton, 1989) that results in accurate estimates of ability, upon which some high-stakes decisions are made.

There is a large volume of published studies in the literature on the effects of sample size and test length on item-parameter estimation in IRT-based test development. Lord's (1968) study is the first one of its kind in which he investigated the sample sizes required for estimating item parameters (*a*-item discrimination, *b*-item difficulty, and *c*-pseudo chance) accurately in the three-parameter logistic model (3PLM) using data from the Scholastic Aptitude Test. As a result, Lord concluded that a minimum of 50 items and 1,000 examinees were required to estimate *a* parameters with high accuracy. With the support of subsequent studies (Patsula & Gessaroli, 1995; Tang, Way, & Carey, 1993; Yen, 1987; Yoes, 1995), 1,000 was taken as the minimum sample size required for accurate item-parameter estimation in IRT.

Studies have also been conducted that investigated the effects of using sample sizes of less than a 1,000 on item parameter estimates. However, their findings have tremendous discrepancies. To illustrate, a limited number of studies on one-parameter logistic model (1PLM) have suggested that 250 (Goldman & Raju, 1986), 300 (Guyer & Thompson, 2011), or 500 (Thissen & Wainer, 1982) are the minimum feasible sample sizes for this dichotomous unidimensional model. As the complexity of the model increases, the discrepancy in findings also increases. For example, a sample of 250 with 15 items (Harwell & Janosky, 1991), of 500 (Stone, 1992) or 750 (Lim & Drasgow, 1990) with 20 items, of 200 (Weiss & Minden, 2012) or 250 (Harwell & Janosky, 1991) with 25 items, of 500 with 30 items (Hulin, Lissak, & Drasgow, 1982), or of 300 with 75 items (Yoes, 1995) have been suggested for two-

parameter logistic model (2PLM). The situation gets more confusing for 3PLM. A sample of 1,000 with 20 items (Patsula & Gessaroli, 1995; Swaminathan & Gifford, 1979; Yen, 1987); of 200 (Weiss & Minden, 2012) with 25 items; of 500 (Akour & Al-Omari, 2013) with 30 items; of 300 (Chuah, Drasgow, & Luecht, 2006) with 50 items; of 1,000 with 40 (Patsula & Gessaroli, 1995; Tang et al., 1993; Yen, 1987), 50 (Lord, 1968), 60 (Hulin et al., 1982) and 75 (Yoes, 1995) items; and of 500 (Ree & Jensen, 1980) with 80 items have been suggested in 3PLM.

While numerous studies have been conducted on the effects of sample size and test length on item-parameter accuracy in IRT-based test development, the present study differs from them in two important aspects. Firstly, the data used in previous research has been mostly simulated data, and it is not known whether simulated data reflects the qualities of a real test (Sireci, 1991). In this manner, the results of the present study are expected to be closer to real test development than most previous research. Moreover, limited numbers of sample sizes, mostly 4 or 5, were tested in the previous research. However, a large number (9) of small and large sample sizes are tested at the same time in the present study, thus making it possible to compare and contrast results from both sides. That makes this study unique in the literature. With the help of these two aspects, this study sets out to provide all IRT practitioners with a blueprint of how large or small a sample size can be to estimate item parameters accurately in IRT-based test development by answering the question "How do sample size and test length affect item-parameter estimation in IRT-based test development?"

Method

Participants and Data Collection Instrument

The participants in this foundational study are 6,288 freshman students enrolled in an English language course at a large university. The data collection instrument was a 50-item English language test. In order to fulfill the content validity concerns, items in the test were written in parallel with the course objectives by the English language instructors who teach the course and are deemed as subject matter experts. Moreover, before administering the test, expert opinions were received from a doctoral candidate and language testing specialists working for the testing office at the university where the data was collected. In this way, the items were revised multiple times based upon expert opinions until reaching their final form. Afterwards, the instrument was administered simultaneously to 6,288 test takers in one session.

Item Selection for Shorter Tests

Items from the full data set were selected to form sub-tests of different lengths. While deciding the number of items for shorter tests, previous studies (Baker, 1998;

Gao & Chen, 2005; Patsula & Gessaroli, 1995; Yen, 1987) in the literature were considered. So, from the 50-item *full data* set (6,288 x 50), sub-tests of n = 30, n = 20, and n = 10 items were selected. While forming sub-tests, unidimensionality assumption of IRT was satisfied. Unidimensionality refers to having a single common factor underlying examinees' correct responses to items. In order to select the item that collectively satisfy unidimensionality, exploratory factor analysis on the tetrachoric inter-item correlation matrix of the full data was used (Edelen & Reeve, 2007). Items that had factor loadings less than 0.30 for the first factor were removed from the data as they failed to load on the first factor adequately (Fidelman, 2012). Out of the 50 items, 30 items that had the highest factor loadings on the first factor were chosen as the items for the 30-item test. The tetrachoric correlation matrix of these 30 items was computed and another exploratory factor analysis was done on this matrix. Then, 20 items with highest factor loadings on the first factor were selected for the 20-item test. The same procedure was followed to select the items for 10-item test. A summary of the results related to the unidimensionality assumption and the factor loadings can be found in Table 1.

Table 1

Summary of Exploratory Factor Analysis in Item Selection

Test Length	Variance	λ_1/λ_2	Range of Factor Loadings
30-Item Test	37.50 %	8.50	0.45 to 0.75
20-Item Test	43.78 %	8.15	0.59 to 0.75
10-Item Test	51.29 %	6.10	0.68 to 0.77

Note. Variance = Variance accounted for by the first factor, λ_1/λ_2 = ratio of the first to the second eigenvalue.

Lord (1980) suggests that if the first eigenvalue is large compared to the second one, and if the second one is not much larger than any of the other eigenvalues, this can be taken as a proof of unidimensionality. These were observed in the data sets of the current study. Apart from these, moderate to high factor loadings for the first factor and moderate to high variance accounted for by the first factor were also taken as proof of unidimensionality. Upon completing the item selection for the sub-tests, the KR-20 internal consistency coefficient was calculated for each test, which were found to be 0.76 (10-item test), 0.85 (20-item test), and 0.88 (30-item test).

Sampling of Examinees

After selecting the items that would be used for each sub-test (10-item, 20-item, and 30-item), nine different sample sizes (N = 150, 250, 350, 500, 750, 1,000, 2,000, 3,000, and 5,000) were drawn from the full dataset for each test by stratified random sampling using the complex samples module of SPSS 20 (International Business Machines Corporation, 2011). This was done in order to simulate different research conditions (e.g., 250 examinees' responses to 20 items). While deciding sample sizes, the sample sizes most frequently used in previous IRT research on sample size were

reviewed. As a result, sample sizes of 150, 250, 500, 1,000, 2,000, 3,000, and 5,000 were found to be used the most in similar studies (Akour & Al Omari, 2013; Baker, 1998; Goldman & Raju, 1986; Hulin et al., 1982; Lord, 1968; Tang et al., 1993; Thissen & Wainer, 1982; Yen, 1987). Samples of 350, which had never been tested, and of 750, which had only been previously tested once (Lim & Drasgow, 1990), were also added to the study as alternatives to 500 and 1,000. Moreover, while sampling the examinees, their colleges were used as strata to reflect examinee diversity in the full data across samples. This process resulted in 30 (See Table 2) different data sets, including the full data set (marked with an *).

Table 2
Data Sets of the Study

	Sample Size										
Test Length	150	250	350	500	750	1,000	2,000	3,000	5,000	6,288*	Total
10	1	1	1	1	1	1	1	1	1	1	10
20	1	1	1	1	1	1	1	1	1	1	10
30	1	1	1	1	1	1	1	1	1	1	10
Total	3	3	3	3	3	3	3	3	3	3	30

Item Parameter Estimation

After sampling, item parameters (item difficulty (*b*) in 1PLM; *b* and discrimination (*a*) in the 2PLM; and *a*, *b*, and pseudo-chance (*c*) parameters in the 3PLM) were estimated under 90 different research conditions (3 IRT models x 30 data sets) using Xcalibre 4.1 (Guyer & Thompson, 2011) with marginal maximum likelihood estimation (MMLE) and the default options readily available in the program. The item parameters estimated from the full data set ($N = 6{,}288$, n = 50) were taken as the "true" (baseline) parameters against which the accuracy of the item parameters estimated from the other samples in the study could be compared and contrasted. As the data had a considerably large number of examinees, it was possible to estimate item parameters with a reasonable amount of accuracy and thus could be considered as baseline parameter estimates (Swaminathan, Hambleton, Sireci, Xing, & Rizavi, 2003).

Estimation Evaluation Criteria

Accuracy of the parameter estimates under different research conditions was evaluated through product-moment correlations (Gao & Chen, 2005; Harwell & Janosky, 1991; Hulin et al., 1982; Swaminathan et al., 2003; Tang et al., 1993; Yen, 1987) and the root-mean-squared difference (RMSD; Gao & Chen, 2005; Harwell & Janosky, 1991; Swaminathan et al., 2003; Yen, 1987) between the baseline and estimated parameters, as has been done in similar studies. RMSD is defined as in Equation 1 below.

$$\text{RMSD} = \sqrt{\frac{\sum_{i=1}^{n}(\bar{\pi}_i - \pi_i)^2}{n}} \quad (1)$$

$\hat{\pi}_i$ represents one of the estimated item parameters (*a*, *b*, or *c*) for item *i*, and π_i represents the corresponding baseline parameter, taken as the true item parameter for item *i*. Under all research conditions, correlations of $r \geq 0.70$, which is considered acceptable (Yoes, 1995) and sometimes as high (Field, 2013), and RMSD ≤ 0.33, which corresponds to the classical reliability value of 0.90 (Rudner, 1998), were taken as the criteria for minimum feasible sample size for that particular test length and IRT model.

Model Data Fit

An item-by-item fit analysis was not conducted; rather, an overall model-data fit analysis was used on all data sets because it might be possible to have items that fit one item response model in a data set while failing to fit the model in other data sets. This would make it difficult to compare the same items' performances in three IRT models with nine sample sizes and three test lengths using real test data from a limited number of items. As the indicator of overall model-data fit, the model chi-square reported by Xcalibre 4.1 (Guyer & Thompson, 2011) was used. As chi-square is sensitive to large sample sizes, chi-square divided by the degrees of freedom (x^2/df) was used as it is not sensitive to large sample sizes (Kline, 2005). A summary of the model-data fit analysis based on x^2/df is shown in Table 3.

Table 3
Summary of the Model Data Fit Analysis Based on x^2/df

	n = 10			*n* = 20			*n* = 30		
	1PLM	2PLM	3PLM	1PLM	2PLM	3PLM	1PLM	2PLM	3PLM
N	x^2/df	x^2/df	x^2/df	x^2/df	x^2/df	x^2/df	x^2/df	x^2/df	x^2/df
150	0.4	0.4	0.6	0.5	0.5	0.5	0.5	0.4	0.4
250	0.5	0.4	0.5	0.5	0.5	0.5	0.5	0.4	0.4
350	0.6	0.8	0.9	0.5	0.4	0.5	0.6	0.4	0.4
500	0.6	1.0	1.4	0.5	0.4	0.4	0.5	0.4	0.4
750	1.2	0.9	1.5	0.7	0.5	0.5	0.8	0.4	0.4
1,000	0.9	0.9	2.1	0.6	0.5	0.6	0.9	0.5	0.5
2,000	2.1	1.7	2.7	0.9	0.7	0.8	1.3	0.5	0.6
3,000	2.7	2.5	4.7	1.0	0.7	0.9	1.9	0.6	0.7
5,000	4.0	3.3	8.0	1.4	1.4	1.3	2.9	0.8	0.8
6,288	5.0	3.9	9.4	1.7	1.6	1.5	3.4	0.9	0.9

Values of 3 or less are suggested as an indicator of model fit (Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001). Table 3 indicates misfit of few large data to the model. These were ignored, as they were thought to result from the very large size of these data sets.

Findings

Figures 1, 2, and 3 show correlations and RMSD values obtained from the parameter estimates. As Figure 1 shows, correlations pertaining to the *b* parameters estimated in 1PLM ranged between 0.939 (N = 150, n = 10) and 1.00 (N = 5,000; n = 10, 20, and 30). In addition, RMSD values pertaining to the *b* parameter estimates ranged between 0.33 (N = 150, n = 10) and 0.01 (N = 5,000; n = 10). As can be seen from the figures, a sample of 150 met both criteria for being deemed acceptable in 1PLM for all test lengths.

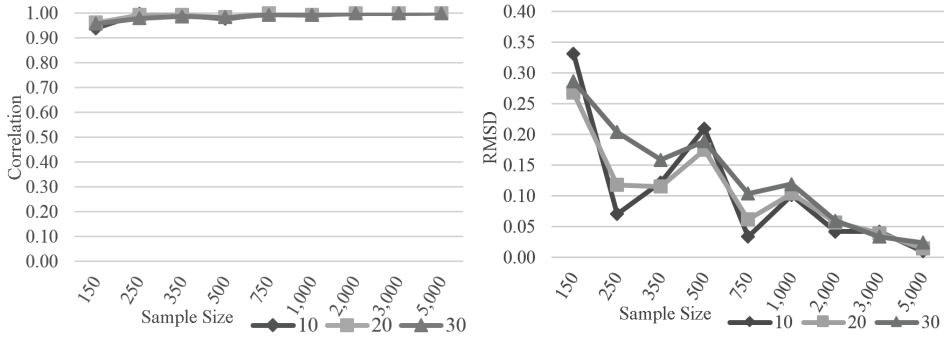
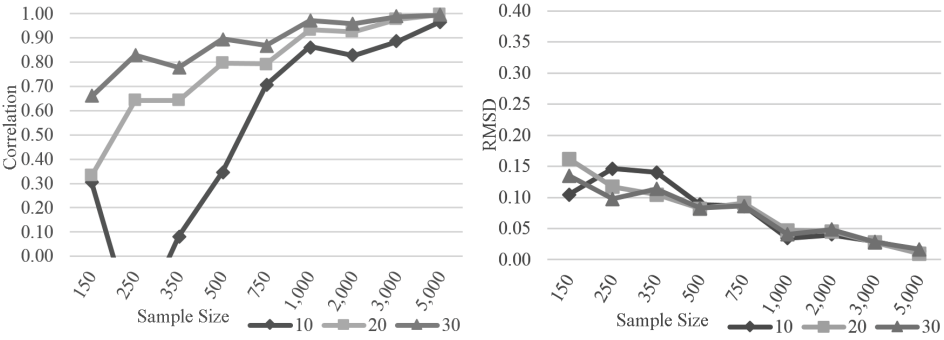
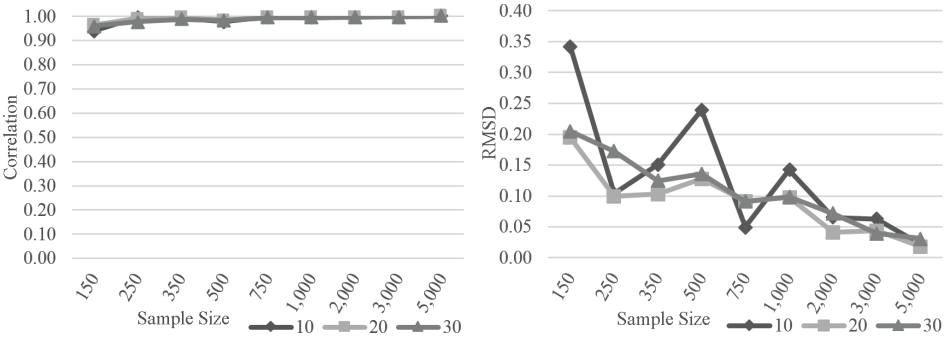


Figure 1. Correlations (left) and RMSD (right) values obtained for the *b* parameter in 1PLM.

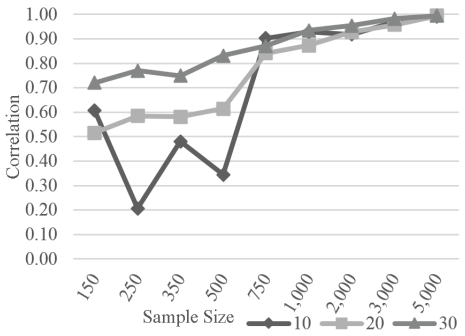


2a. Correlations for the *a* parameter in 2PLM 2b. RMSD for the *a* parameter in 2PLM

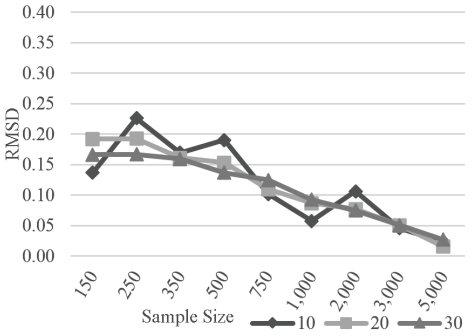


2c. Correlations for the *b* parameter in 2PLM 2d. RMSD for the *b* parameter in 2PLM

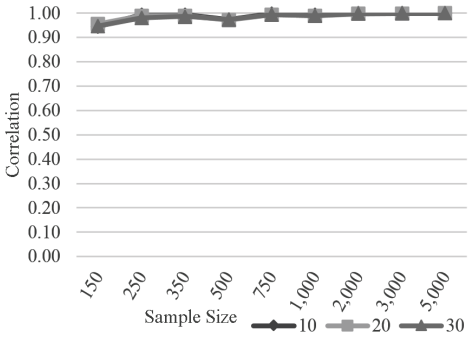
Figure 2. Correlations and RMSD values obtained for the *a* and *b* parameters in 2PLM.



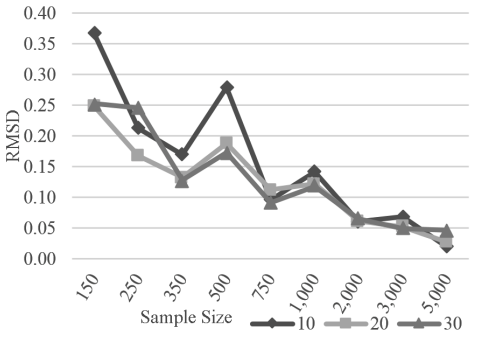
3a. Correlations for the *a* parameter in 3PLM



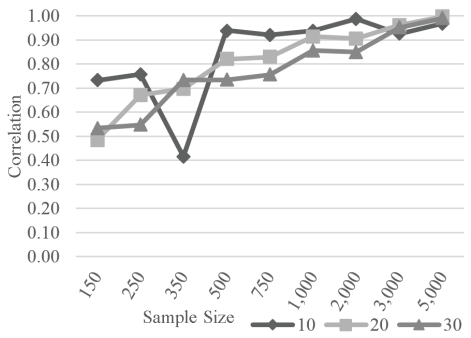
3b. RMSD for the *a* parameter in 3PLM



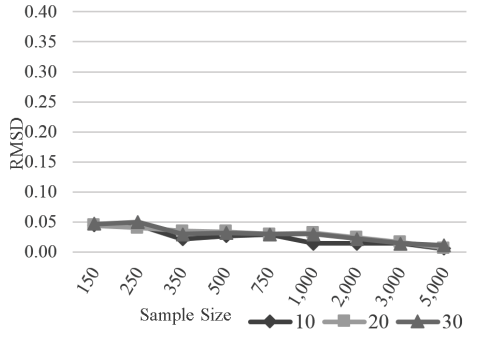
3c. Correlations for the *b* parameter in 3PLM



3d.RMSD for the *b* parameter in 3PLM



3e. Correlations for the *c* parameter in 3PLM



3f. RMSD for the *c* parameter in 3PLM

Figure 3. Correlations and RMSD values obtained for the *a*, *b*, and *c* parameters in 3PLM.

The correlations pertaining to the *a* parameter (Figure 2a) in 2PLM ranged between -0.311 ($N = 250$, $n = 10$) and 1.00 ($N = 5{,}000$; $n = 20, 30$). The RMSD values (Figure 2b) obtained from the *a* parameters in 2PLM varied from 0.16 ($N = 150$, $n = 20$) to 0.00 ($N = 5{,}000$; $n = 10, 20$). As for the *b* parameters, the correlations (Figure 2c) ranged between 0.940 ($N = 150$, $n = 10$) and 1.00 ($N = 2{,}000, 3{,}000, 5{,}000$; $n = 10, 20, 30$) while the RMSD values for the *b* parameter (Figure 2d) ranged between 0.34

(N = 150, n = 10) and 0.02 (N = 5,000; n = 10, 20). In line with these, the minimum sample size that met both evaluation criteria for both parameters (a and b) estimated in 2PLM was 750 ($r_{a\hat{a}}$ = 0.706, RMSD = 0.09; $r_{b\hat{b}}$ = 0.999, RMSD = 0.05) for the 10-item test, 500 ($r_{a\hat{a}}$ = 0.795, RMSD = 0.08; $r_{b\hat{b}}$ = 0.985, RMSD = 0.13) for the 20-item test, and 250 ($r_{a\hat{a}}$ = 0.830, RMSD = 0.10; $r_{b\hat{b}}$ = 0.978, RMSD = 0.17) for the 30-item test in 2PLM.

The correlations pertaining to the a parameter estimates in 3PLM (Figure 3a) ranged between 0.207 (N = 250, n = 10) and 0.995 (N = 5,000; n = 20, 30), whereas the RMSD values obtained from the a parameter estimates in 3PLM (Figure 3b) ranged between 0.23 (N = 250, n = 10) and 0.02 (N = 5,000; n = 10, 20). The correlations pertaining to the b parameter estimates (Figure 3c) ranged between 0.944 (N = 150, n = 10) and 1.00 (N = 2,000, 3,000, 5,000; n = 10, 20, 30) while the RMSD values obtained from the b parameter estimates (Figure 3d) ranged between 0.37 (N = 150, n = 10) and 0.02 (N = 5,000; n = 10). As for the c parameter, the correlations (Figure 3e) ranged between 0.415 (N = 350, n = 10) and 0.996 (N = 5,000, n = 20) whereas the RMSD values pertaining to the c parameter estimates (Figure 3f) ranged between 0.05 (N = 150, n = 30; N = 250, n = 10, 30) and 0.01 (N = 1,000, 2,000, 3,000, 5,000, n = 10; N = 5,000, n = 20; N = 3,000, 5,000, n = 30). As a result, the smallest sample sizes that met the criteria for all parameters (a, b, and c) estimated in 3PLM were 750 ($r_{a\hat{a}}$ = 0.901, RMSD = 0.10; $r_{b\hat{b}}$ = 0.998, RMSD = 0.10; $r_{c\hat{c}}$ = 0.992, RMSD = 0.03) when n = 10; 750 ($r_{a\hat{a}}$ = 0.840, RMSD = 0.11; $r_{b\hat{b}}$ = 0.995, RMSD = 0.11; $r_{c\hat{c}}$ = 0.829, RMSD = 0.03) when n = 20; and 350 ($r_{a\hat{a}}$ = 0.749, RMSD = 0.16; $r_{b\hat{b}}$ = 0.987, RMSD = 0.13; $r_{c\hat{c}}$ = 0.734, RMSD = 0.03) when n = 30.

Discussion

This study aimed to investigate the effects of sample size and test length on item-parameter estimation accuracy in IRT-based test development. The results indicate an association between test length and sample size that amplifies as the number of items in the test and the number of parameters in the model increase. As mentioned earlier, there are two criteria set for estimation evaluation. One is that $r \geq 0.70$; the other is that RMSD $\leq$ 0.33 between the baseline and estimated item parameters. A summary of the minimum sample sizes that met both of these criteria in all IRT models can be found in Table 4.

Table 4
Summary of the Findings

Test Length	1PLM	2PLM	3PLM
10	150	750	750
20	150	500	750
30	150	250	350

As Table 4 illustrates, the findings of the present study suggest that a sample of as low as 150 examinees could be used in 1PLM with tests of 10, 20, or 30 items to accurately estimate the *b* parameter. Such a result was partially expected as there had already been some signs regarding the viability of using 150 examinees for 1PLM in the current literature. More specifically, some authors have already suggested a sample size of around 200 as appropriate for 1PLM (Demars, 2010; Wright & Stone, 1979) with no specific reference to test length. The present finding is important as it confirms these authors' findings with a specific reference to test length. Moreover, some previous research had similar findings, like $N = 250$ (Goldman & Raju, 1986) or $N = 300$ (Guyer & Thompson, 2011) with 78 and 50 items, respectively. These sample sizes and test lengths were the smallest sample sizes and shortest test lengths tested in these studies. Thus, the present finding, which suggests a sample size of 150 in tests of 10, 20, or 30 items, may constitute a viable update to the current findings in the literature. However, although the correlations and RMSD values similar to smaller sample sizes confirmed that the full data, as well as $N = 5{,}000$ for 10-, 20-, and 30-item tests and $N = 3{,}000$ in 30-item tests, fit the 1PLM as parameter invariance was reached among the samples, the findings pertaining to 1PLM may still be biased (as x^2/df values indicated misfit) and need to be interpreted accordingly.

As seen in Table 4, the minimum sample size suggested for a 10-item test in 2PLM was 750. This was an interesting finding for two reasons. First, a test of 10 items in 2PLM was not popular in previous studies. Only three studies tested 10-item tests in 2PLM (Baker, 1998; Stone, 1992; Yen, 1987). However, they were unable to yield acceptable accuracy with it. The sample sizes tested in these studies were 1,000 (Yen, 1987); 30, 50, 60, 120, and 500 (Baker, 1998); and 250, 500, and 1,000 (Stone, 1992). The present finding differed from Yen's (1987) and Stone's (1992) studies as they did not report 1,000 as a viable sample size for 10 items. On the contrary, in the present study, the item parameters of 10 items with 1,000 examinee responses were estimated with high accuracy ($r_{a\hat{a}} = 0.864$, RMSD = 0.03; $r_{b\hat{b}} = 0.964$, RMSD = 0.14). Second, as stated earlier, a sample size of 750 was not a popular alternative to 1,000 in the previous research. Only one study (Lim & Drasgow, 1990) tested the performance of 750 with 20 items and reported that the *a* and *b* parameter estimates were close to their true values. This is clearly in parallel with the present finding with a shorter test length.

The present findings indicate that accurate estimates of item parameters can be obtained when $N = 500$ and $n = 20$ in 2PLM. There have been studies that suggested a sample of 500 in 2PLM with 30- (Hulin et al., 1982) and 78-item (Goldman & Raju, 1986) tests. However, it should be noted that Hulin et al. (1982) only tested 15-, 30-, and 60-item tests and used joint maximum likelihood estimation (JMLE) as the parameter estimation method. This could accordingly cause less accurate estimation of item parameters in smaller sample sizes and shorter test lengths as JMLE is more

efficient with sample sizes of over 1,000 and tests with more than 60 items (Baker & Kim, 2004). Moreover, it should also be noted that 78 was the only test length tested by Goldman and Raju (1986). For these reasons, the comparability of these two studies with the present study is poor. However, it is encouraging to compare the present finding with that of Stone's (1992), who reported an RMSD value of approximately 0.30 when $N = 500$ and $n = 20$, indicating a parallel finding with the present study.

The results of the present study also suggest that using a sample of 250 with 30 items in 2PLM is viable. This is consistent with the findings of Harwell and Janosky (1991), who obtained accurate *a* and *b* parameter estimates with a 25-item test when $N = 250$. Apart from this, Hulin et al. (1982) suggested a sample size of 500 with 30 items in 2PLM. As mentioned earlier, due to the estimation method (JMLE) they had employed to estimate item parameters, it would not be wrong to assume that they could have obtained a smaller sample size with 30 items if they had used the MMLE estimation method used in the present study. Moreover, Guyer and Thompson (2011) obtained accurate *a* and *b* parameters in 2PLM when $N = 300$ and $n = 50$. One should note that $n = 50$ was the shortest test length and $N = 300$ was the smallest sample size they tested, which means the present study obtained similar results testing a smaller sample size and shorter test length.

As shown in Table 4, acceptable accuracy is reached when $N = 750$ and $n = 10$ in 3PLM. Although tests with 10 items had been previously tested in two studies (Gao & Chen, 2005; Yen, 1987) in 3PLM, there was no suggestion for this test length in the current literature for 3PLM as previous research did not yield accurate parameter estimates. To begin with, when Gao and Chen's (2005) study was analyzed in terms of test length and sample size ($N = 100, 500, 2{,}000$; $n = 10, 30, 60$), one can see that the closest sample size to 750 was 500. They could not yield accurate estimates of the *a* parameter with a sample of 500. The present study confirmed that 500 examinees with 10 items in 3PLM yields very poor estimates for the *a* parameter ($r_{a\hat{a}} = 0.345$, RMSD = 0.19). Gao and Chen (2005) did not test 750 or even 1,000 in their study as an alternative to 2,000. However, 750 was tested in the present study, and accurate estimates of the *a* parameter have been obtained. Secondly, Yen (1987) obtained poor accuracy with a sample of 1,000 and 10 items in 3PLM. On the contrary, the present findings indicated that when $N = 750$ and $n = 10$ ($r_{a\hat{a}} = 0.901$, RMSD = 0.10; $r_{b\hat{b}} = 1.00$, RMSD = 0.10; $r_{c\hat{c}} = 0.922$, RMSD = 0.03), the item parameters were estimated as accurately as they were when $N = 1{,}000$ ($r_{a\hat{a}} = 0.929$, RMSD = 0.06; $r_{b\hat{b}} = 0.999$, RMSD = 0.14; $r_{c\hat{c}} = 0.938$, RMSD = 0.01) with the same test length.

When the test length increased from 10 to 20 items in 3PLM, the sample size with which acceptable accuracy was reached was also 750. Patsula and Gessaroli (1995) and Yen (1987) had previously studied the 20-item test condition and both suggested samples of 1,000 with 20 items. Yen (1987) studied only the sample of 1,000 and

Patsula and Gessaroli (1995) did not have $N = 750$ in their research design. They only had samples of 1,000 and 500 as the closest alternatives to 750. They found that 500 did not yield accurate enough estimates, which was also the case in the present study under the same research condition ($r_{a\hat{a}} = 0.614$, RMSD = 0.15; $r_{b\hat{b}} = 0.972$, RMSD = 0.19; $r_{c\hat{c}} = 0.820$, RMSD = 0.03). As such, it should not be surprising that they suggested 1,000 as the minimum sample size for they did not have the opportunity to test 750 as an alternative. The present findings also indicate that 750 is a highly viable alternative to 1,000 as item-parameter estimates in $N = 750$ ($r_{a\hat{a}} = 0.840$, RMSD = 0.11; $r_{b\hat{b}} = 0.995$, RMSD = 0.11; $r_{c\hat{c}} = 0.829$, RMSD = 0.03) are as accurate as when $N = 1{,}000$ ($r_{a\hat{a}} = 0.874$, RMSD = 0.09; $r_{b\hat{b}} = 0.989$, RMSD = 0.12; $r_{c\hat{c}} = 0.915$, RMSD = 0.03) and $n = 20$ in 3PLM.

As seen in Table 4, the sample size suggested for estimating item parameters accurately in 3PLM with 30 items was 350. Previously, two studies had tested a sample of 500 in 3PLM with 30 items (Akour & Al-Omari, 2013; Gao & Chen, 2005). When these studies were further analyzed, it can be seen that the sample sizes of 100 and 200 were also tested in these studies respectively. Thus, $N = 350$ naturally isn't found in the current literature as it was never tested against 500. Many researchers (Goldman & Raju, 1986; Harwell & Janosky, 1991; Patsula & Gessaroli, 1995; Ree & Jensen, 1980; Yoes, 1995) instead preferred the sample size of 250 as a smaller alternative to 500. That the present finding suggests the use of a sample size of 350 as viable with 30-item tests in 3PLM thus should not be intriguing.

As one can infer from the discussion in the previous paragraphs, the findings of the present study not only support the findings of previous research in the field, but also further its findings by providing researchers with more data in terms of viable sample sizes for IRT-based test development. More importantly, Table 4 provides IRT practitioners with a detailed blueprint on what sample size to use with which model and test length. In addition, the findings presented in Table 4 suggest a tremendous decrease in the minimum sample size required for IRT-based test development. This means that IRT practitioners will need to reach fewer examinees to pretest their items, and this will make the data collection process easier and more budget-friendly. From this point of view, the findings of this study will not only be beneficial for IRT practitioners who can only reach a limited number of examinees, but it will also help to decrease the data collection, item analysis, test production, and personnel costs for companies and institutions that need to recruit more staff to meet the requirements of IRT-based test development.

In order to develop similar tests with similar qualities, the key reliability and validity aspects taken under control in this study may need to be followed. First of all, the KR-20, which is a variant of the alpha internal consistency coefficient for binary items, was high for the tests that were drawn from the full test in the present

study. Moreover, the sub-tests formed in the study were highly unidimensional. In subsequent studies, tests with high internal consistency and high unidimensionality may be needed in order to obtain similar results. As part of content validity concerns, the items in the developed test were written by subject matter experts in accordance with course curriculum, and multiple expert opinions were taken. Afterwards, the test was administered as a real test to real examinees. A similar approach may also be necessary in order to obtain similar results.

Present findings should not be deemed as definitive. However, they should be deemed as a call for further research on minimum sample size necessary to estimate item parameters accurately. One should note that the present findings may only be limited to language-test development and short test lengths up to 30 items with qualities similar to those used in the present study. Moreover, the parameter estimates were obtained using MMLE, so the present findings may be limited to item parameter estimation with MMLE.

This study has thrown up some questions in need of further investigation. A natural progression of this work would be to conduct studies on the following topics by using real and simulated test data. A language test data was used in this study. Replicating this study with data from tests other than language tests would be of value to the practitioners in the field. It would also be a valuable contribution to the field to investigate feasibility of using small sample sizes in item-parameter estimation for tests with more than 30 items.

References

Akour, M., & Al Omari, H. (2013). Empirical investigation of the stability of IRT item-parameters estimation. *International Online Journal of Educational Sciences*, *5*(2), 291–301. Retrieved from http://www.iojes.net//userfiles/Article/IOJES_980.pdf

Baker, F. B. (1992). *Item response theory: Parameter estimation techniques*. New York, NY: Marcel Dekker.

Baker, F. B. (1998). An investigation of the item parameter recovery of a Gibbs sampling procedure. *Applied Psychological Measurement*, *22*(2), 153–169. http://dx.doi.org/10.1177/01466216980222005

Baker, F. B., & Kim, S. H. (2004). *Item response theory: Parameter estimation techniques* (2nd ed.). New York, NY: Marcel Dekker.

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, *36*(4), 523–562. http://dx.doi.org/10.1207/S15327906MBR3604_03

Chuah, S. C., Drasgow F., & Luecht, R. (2006). How big is big enough? Sample size requirements for cast item parameter estimation. *Applied Measurement in Education, 19*(3), 241–255. http://dx.doi.org/10.1207/s15324818ame1903_5

DeMars, C. (2010). *Item response theory: Understanding statistics measurement*. New York, NY: Oxford University Press.

Edelen, M. O., & Reeve, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, *16*(1), 5–18. http://dx.doi.org/10.1007/s11136-007-9198-0

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.

Fidelman, C. (2012, January). *A retrospective linking of ECLS-K and ECLS-B reading scores*. Paper presented at Federal Committee on Statistical Methodology Policy Seminar, Washington, DC.

Field, A. P. (2013). *Discovering statistics using IBM SPSS Statistics: And sex and drugs and rock 'n' roll* (4th ed.). London, UK: Sage.

Gao, F., & Chen, L. (2005). Bayesian or non-Bayesian: A comparison study of item parameter estimation in the three-parameter logistic model. *Applied Measurement in Education*, *18*(4), 351–380. http://dx.doi.org/10.1207/s15324818ame1804_2

Goldman, S. H., & Raju, N. S. (1986). Recovery of one- and two-parameter logistic item parameters: An empirical study. *Educational and Psychological Measurement*, *46*(1), 11–21. http://dx.doi.org/10.1177/0013164486461002

Guyer, R., & Thompson, N. A. (2011). *User's manual for Xcalibre 4.1*. St. Paul, MN: Assessment Systems Corporation.

Hambleton, R. K., Swaminathan H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Newbury Park, CA: Sage.

Hambleton, R. K. (1989). Principles and selected applications of item response theory. In R. L. Linn (Ed.), *Educational measurement* (3rd ed., pp. 147–200). New York, NY: Macmillan.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets and varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*(3), 279–291. http://dx.doi.org/10.1177/014662169101500308

Hulin, C. L., Lissak, R. I., & Drasgow, F. (1982). Recovery of two and three-parameter logistic item characteristic curves: A Monte Carlo study. *Applied Psychological Measurement*, *6*(3), 249–260. http://dx.doi.org/ 10.1177/014662168200600301

International Business Machines Corporation. (2011). *IBM SPSS Statistics for Windows, Version 20.0.* Armonk, NY: Author.

Kline, R. B. (2005). *Principles and practice of structural equation modeling* (2nd ed.). New York, NY: Guilford Press.

Lim, R. G., & Drasgow, F. (1990). Evaluation of two methods for estimating item response theory parameters when assessing differential item functioning. *Journal of Applied Psychology*, *75*(2), 164–174. http://dx.doi.org/10.1037/0021-9010.75.2.164

Lord, F. M. (1968). An analysis of the verbal scholastic aptitude test using Birnbaum's three-parameter logistic model. *Educational and Psychological Measurement, 28*(4), 989–1020. http://dx.doi.org/10.1177/001316446802800401

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

Patsula, L. N., & Gessaroli M. E. (1995, April). *A comparison of item parameter estimates and ICCs produced with TESTGRAF and BILOG under different test lengths and sample sizes*. Paper presented at the annual meeting of the National Council on Measurement in Education, San Francisco, CA.

Ree, M. J., & Jensen, H. E. (1980). Effects of sample size on linear equating of item characteristic curve parameters. In D. J. Weiss (Ed.)*, Proceedings of the 1979 Computerized Adaptive Testing Conference*. Minneapolis, MN: University of Minnesota.

Rudner, L. M. (1998). *An on-line, interactive, computer adaptive testing tutorial*. Retrieved from http://echo.edres.org:8080/scripts/cat/catdemo.htm

Sireci, S. G. (1991, June). *Sample independent item parameters? An investigation of the stability of IRT item parameters estimated from small data sets*. Paper presented at the annual Conference of Northeastern Educational Research Association, New York, NY.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimates in the two-parameter logistic response model: An evaluation of Multilog. *Applied Psychological Measurement, 16*(1), 1–16. http://dx.doi.org/10.1177/014662169201600101

Swaminathan, H., & Gifford, J. A. (1979). *Estimation of parameters in the three-parameter latent trait model* (Report No. 90). Amherst, MA: University of Massachusetts, School of Education, Laboratory of Psychometric and Evaluation Research.

Swaminathan, H., Hambleton, R. K., Sireci, S. G., Xing, D., & Rizavi, S. M. (2003). Small sample estimation in dichotomous item response models: Effect of priors based on judgmental information on the accuracy of item parameter estimates. *Applied Psychological Measurement*, *27*(1), 27–51. http://dx.doi.org/10.1177/0146621602239475

Tang, K. L., Way, W. D., & Carey, P. A. (1993). *The effect of small calibration sample sizes on TEOFL IRT-based equating (TOEFL technical report TR-7)*. Princeton, NJ: Educational Testing Service.

Thissen, D., & Wainer, H. (1982). Some standard errors in item response theory. *Psychometrika*, *47*(4), 397–412. http://dx.doi.org/10.1007/BF02293705

Weiss, D. J., & Minden, S. V. (2012). *A comparison of item parameter estimates from Xcalibre 4.1 and Bilog-MG.* St. Paul, MN: Assessment Systems Corporation.

Woods, A., & Baker, R. (1985). Item response theory. *Language Testing*, *2*(2), 117–140. http://dx.doi.org/10.1177/026553228500200202

Wright, B. D., & Stone, M. H. (1979). *Best test design*. Chicago, IL: Mesa Press.

Yen, W. M. (1987). A comparison of the efficiency and accuracy of Bilog and Logist. *Psychometrika*, *52*(2), 275–291. http://dx.doi.org/10.1007/BF02294241

Yoes, M. (1995). *An updated comparison of micro-computer based item parameter estimation procedures used with the 3-parameter IRT model*. Saint Paul, MN: Assessment Systems Corporation.