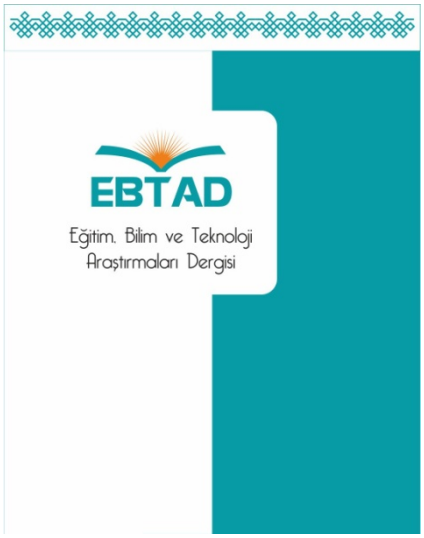


Eğitim, Bilim ve Teknoloji Araştırmaları Dergisi

Madde Tepki Teorisinde Ölçeklerin Ortak Değişkenler ile İlişkilendirilmesi

Valentina Sansivieri[1], Marie Wiberg[2]

[1]Bologna Üniversitesi
[2]Umeå Üniversitesi

Bu makaleye atıf için:
Sansivieri, V., & Wiberg, M. (2018). Madde tepki teorisinde ölçeklerin ortak değişkenler ile ilişkilendirilmesi. *Eğitim, Bilim ve Teknoloji Araştırmaları Dergisi*, *3*(2), 12-32.

Dergi web sayfası için lütfen tıklayınız…

Journal of Research in Education, Science and Technology

Linking Scales in Item Response Theory with Covariates

Valentina Sansivieri[1], Marie Wiberg[2]

[1]University of Bologna
[2]Umeå University

To cite this article:
Sansivieri, V., & Wiberg, M. (2018). Linking scales in item response theory with covariates. *Journal of Research in Education, Science and Technology*, *3*(2), 12-32.

Please click here to access the journal web site...

Eğitim, Bilim ve Teknoloji Araştırmaları Dergisi

Volume 3, Issue 2, Autumn 2018, Page 12-32 ISSN: 2548-0286

Madde Tepki Teorisinde Ölçeklerin Ortak Değişkenler ile İlişkilendirilmesi

Valentina Sansivieri[1*], Marie Wiberg[2],
[1]Bologna Üniversitesi
[2]Umeå Üniversitesi

Makale Bilgisi

Makale Tarihi

Gönderim Tarihi:
09 Kasım 2018

Kabul Tarihi:
27 Aralık 2018

Anahtar Kelimeler

MTT,
Ortak değişkenler,
Birleştirme,
DMF

Özet

Bir biriyle eşit özelliklere sahip olmayan farklı gruplara test formları uygulandığı ve sonuçlar madde tepki kuramına (MTT) göre puanlandırıldığı zaman, iki grup için ayrı ayrı tahmin edilen madde parametrelerinin aynı ölçeğe yerleştirilmesi gerekmektedir. Test edilenlerle ilgili değişkenleri içeren MTT modellerinde, aynı skalaya koyulması gereken, düzgün olan ve düzgün olmayan değişken madde fonksiyonunu (DMF) modelleyen iki farklı parametre vardır. Bu çalışma düzgün olan ve düzgün olmayan DMF parametrelerini aynı skalaya yerleştiren dönüşüm denklemlerini önermeyi amaçlamaktadır. Dönüşüm denklemlerinin katsayılarını tahmin etmek amacıyla bu çalışmada şu dört yöntem kullanılmıştır: ortalama/ortalama, ortalama/sigma, Haebara ve Stocking-Lord. Araştırmamızda bir simülasyon çalışması ve deneysel bir örnek vermekteyiz. Bu simülasyon çalışmasının sonuçları bizlere eşitlik denklemlerinin katsayılarının büyük ölçüde Haebara ve Stocking-Lord yöntemleri için aynı olduğunu göstermiş olsa da, diğer yöntemler için farklı olduğunu göstermiştir. Deneysel örneğimizin sonuçları ise yüksek beceri değerleri için değişkenlerle birlikte olan MTT'nin değişkenler olmaksızın uygulanan MTT'den daha bilgilendirici bir sonuç ürettiğini göstermiştir. Bunun yanında ortalama/ortalama ve ortalama/sigma yöntemleri kullanıldığında eş zamanlı kalibrasyon yöntemine göre daha aydınlatıcı sonuçlar elde edilmiştir.

Linking Scales in Item Response Theory with Covariates

Valentina Sansivieri[1†], Marie Wiberg[2],
[1]University of Bologna
[2]Umeå University

Article Info

Article History

Received:
November 09, 2018

Accepted:
December 27, 2018

Keywords

IRT,
Covariates,
Linking,
DIF

Abstract

When test forms are administered to different non-equivalent groups of examinees and are scored by item response theory (IRT), it is necessary to put item parameters estimated separately on two groups on the same scale. In the IRT models which include covariates about the examinees, we have two parameters which model uniform and non-uniform differential item functioning (DIF) and that have to be put on the same scale. The aim of this study is to propose conversion equations, which are used to put the uniform and non-uniform DIF parameters on the same scale. To estimate the coefficients of the conversion equations we will use four methods: mean/mean, mean/sigma, Haebara and Stocking-Lord. We give a simulation study and an empirical example. The results of the simulation study show that the coefficients of the conversion equations are substantially equal for the Haebara and Stocking-Lord methods, while they are different for the other methods. The results of the empirical example is that IRT with covariates produces a more informative test than using IRT without covariates for high abilities' values and, when the mean-mean and the mean-sigma methods are used, we obtain more informative tests than when using concurrent calibration.

[*]İletişim: *Valentina Sansivieri, Bologna Üniversitesi, Yönetim Bölümü, valentina.sansivier2@unibo.it*
[†]Corresponding Author: *Valentina Sansivieri, Department of Management, University of Bologna, valentina.sansivier2@unibo.it*

INTRODUCTION

Item response theory (IRT; Lord, 1980) is widely used in educational and psychological testing to construct items and to analyze test results as well as making inferences about the examinees. If we wish to compare two groups from different populations who have taken different test forms, the item parameters and examinees for each of the groups are not by default on the same scale and thus cannot be directly comparable. In order to compare the item parameters from two test forms one must place them on the same scale (Kolen & Brennan, 2014; González & Wiberg, 2017). To estimate the item parameters from two test forms and place those on the same scale one can use common items, i.e. an anchor test. One way to place the item parameters on the same scale is to use concurrent calibration, which is a multi-group estimation procedure where one estimates the item parameters from two different test forms at the same time. In the concurrent calibration, one can take into account the differences between the ability distributions of the groups the test forms have been given and the item parameters are estimated with one of the groups used as a reference group.

Another way to put the item parameters on the same scale is to estimate them separately in the groups. If we have common items, their item parameters should be identical in the two groups. Thus, one can use information from the common items to estimate linking parameters, which are used to place all the item parameters from the two test forms on the same scale. This paper focuses on the linking parameters. To estimate the linking parameters there are two estimator types which have commonly been used; methods based on response functions and methods based on moments. The response function methods minimize a distance measure between the test or item characteristics functions for estimated common item parameters from two different groups (Haebara, 1980; Stocking & Lord, 1983). The methods based on moments utilize the means and the variances of the estimated common item parameters to estimate the linking parameters (Loyd & Hoover, 1980; Marco, 1977). Previous researches have shown that the response function methods have more favorable properties than the methods based on moments (Hanson & Béguin, 2002; Kim & Cohen, 1998; Kim & Kolen, 2007). Much recent research has aimed to find expressions for the variance of the linking coefficient estimators for different models (see e.g. Ogasawara, 2001; 2011; Andersson, 2018).

When constructing a test it is important to have good items. However, even in well-designed tests we can have items with differential item functioning (DIF). DIF is present in an item if examinees from different subgroups (for example male and female examinees) display different probabilities to solve the item even if they have the same ability (Embretson & Reise, 2000). Ideally, when we construct a test or an anchor test they should be free from DIF in the items. Anchor items are ideally free from DIF to avoid artificially augmented false alarm rates (Kopf, Zeileis, & Strobl, 2015). If the anchor contains DIF items the construction of a common scale for the item parameters may fail and the result can be an increase of false alarm rates (see e.g. Finch, 2005; Stark, Chernyshenko, & Drasgow, 2006; Wang & Yeh, 2003; Wang, 2004). In other words, true DIF free items may appear to have DIF, which may jeopardize the results of a DIF analysis (Jodoin & Gierl, 2001). Also, when anchor item DIF varies across test forms in a differential manner across subpopulations, population invariance of equating can be compromised (Huggins-Manley, 2014). Although a possible way would be to exclude DIF items from an anchor using a number of iterative step, it may not solve the problem if the test contains many DIF items (Wang, Shih, & Sun, 2012). Another route to follow if one has DIF items is to incorporate DIF when modeling the items. Tay, Newman and Vermunt (2011; 2016) proposed to incorporate parameters to model DIF in IRT models. If we incorporate DIF parameters in an IRT model we need tools to link the DIF coefficients between test forms. The overall aim of this paper is to propose linking coefficients when using IRT models with covariates. The proposed linking coefficients are illustrated in an empirical study with real test data and a simulation study. The results of this paper are important because linking coefficients are used in many different educational measurements, especially when equating or linking achievement tests. Although it is true that one wants to avoid DIF in achievement tests, there are examples where DIF are present in several items in achievement tests (Wedman, 2018). The proposed linking coefficients will be useful so we can still perform a good equating or linking of achievement tests, which have DIF items present and where it is not suitable to remove those items.

The rest of this paper is structured as follows. First, traditional methods for linking scales in IRT is described, followed by a description of IRT with covariates and how one can link scales when an IRT model with covariates is used. Next, a simulation study and an empirical study are described in which the proposed linking coefficients are used. The result section is given next and the paper ends with a discussion with some concluding remarks is given.

METHODS

When traditional IRT models are used, there are well-known procedures for placing the item parameters on the same scale. A common IRT model is the traditional three-parameter logistic IRT (3PL) model (Birnbaum, 1968). Let a_i be the item discrimination parameter, b_i the item difficulty parameter, and c_i the pseudo-guessing parameter. Let p_{ji} be the probability of a correct response of examinee j on item i, for the 3PL be defined as follows

$$p_{ji} = c_i + (1-c_i)\frac{\exp[Da_i(\theta_j - b_i)]}{1+\exp[Da_i(\theta_j - b_i)]}, \quad (1)$$

where D is a constant equal to 1.7. Let J be the scale of test form 1 and let I be the scale of test form 2, then we can use the conversion equations given in Lord (1980) to put item parameters and examinees abilities estimated on test form 2 on the scale J of test form 1:

$$\theta_{Jj} = A\theta_{Ij} + B \quad (2)$$

$$a_{Ji} = \frac{a_{Ii}}{A} \quad (3)$$

$$b_{Ji} = Ab_{Ii} + B \quad (4)$$

$$c_{Ji} = c_{Ii}, \quad (5)$$

where θ_{Jj} is the ability of the examinee j, a_{Ji} is the discrimination of item i, b_{Ji} is the difficulty of item i, and c_{Ji} is the pseudo-guessing of item i on the scale J of test form 1. Θ_{Ij}, a_{Ii}, b_{Ii}, and c_{Ii} are the corresponding values on the scale I of test form 2. The coefficients A and B can be estimated by using different methods. The most commonly used methods to estimate A and B are the mean/mean method (Loyd & Hoover, 1980), the mean/sigma method (Marco, 1977), the Haebara method (Haebara, 1980) and the Stocking-Lord method (Stocking & Lord, 1983). These four methods can be used when test form 1 and test form 2 have a set of common items, i.e. an anchor test. The mean/mean method estimates A and B as follows:

$$A = \frac{\mu(a_I)}{\mu(a_J)} \quad (6)$$

$$B = \mu(b_J) - A\mu(b_I), \quad (7)$$

where $\mu(a_J)$ and $\mu(a_I)$ are the means of the discriminations parameters of common items on the scale J of test form 1 and on the scale I of test form 2, respectively. Further, $\mu(b_J)$ is the means of the difficulty parameters of the common items on the scale J of test form 1 and on the scale I of test form 2, respectively. The mean/sigma method estimates B by using Equation 7 and A as follows:

$$A = \frac{\sigma(b_J)}{\sigma(b_I)}, \quad (8)$$

where $\sigma(b_J)$ and $\sigma(b_I)$ respectively, are the standard deviations of the common items difficulties on the scale *J* of test form 1 and on the scale *I* of test form 2.

The Haebara and Stocking-Lord methods start by the consideration that the probability p_{ji} of a correct response of the examinee *j* on the item *i* must be the same regardless of the scale on which the items parameters and the examinees abilities are. In mathematical terms, using the quantities defined previously, this can be written as:

$$p_{ji}(\theta_{Jj}, a_{Ji}, b_{Ji}, c_{Ji}) = p_{ji}(A\theta_{Ij} + B, \frac{a_{Ii}}{A}, Ab_{Ii} + B, c_{Ii}). \quad (9)$$

Haebara (1980) suggests calculating *A* and *B* by minimizing the following quantity:

$$\sum_j \sum_{i \in V} [p_{ji}(\theta_{Jj}, \hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji}) - p_{ji}(A\theta_{Ij} + B, \frac{\hat{a}_{Ii}}{A}, A\hat{b}_{Ii} + B, \hat{c}_{Ii})]^2 \quad (10)$$

where *V* denotes the set of anchor items, $\hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji}$ are the estimates of the corresponding quantities on scale *J* and $\hat{a}_{Ii}, \hat{b}_{Ii}, \hat{c}_{Ii}$ are the estimates of the corresponding quantities on scale *I* .

Stocking and Lord (1983) suggest calculating *A* and *B* by minimizing the following quantity:

$$\sum_j \left[\sum_{i \in V} p_{ji}(\theta_{Jj}, \hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji}) - \sum_{i \in V} p_{ji}(A\theta_{Ij} + B, \frac{\hat{a}_{Ii}}{A}, A\hat{b}_{Ii} + B, \hat{c}_{Ii}) \right]^2 \quad (11)$$

There are two different types of DIF; uniform DIF and non-uniform DIF. Uniform DIF means that an item can vary only in difficulty among the subgroups. Non-uniform DIF implies that an item can vary also in discrimination among the subgroups. The 3PL with covariates (3PL-C) model (Tay, Newman & Vermunt, 2011; 2016) is defined as

$$p_{ji} = c_i + (1 - c_i) \frac{\exp[Da_i(\theta_j - b_i - d_i z_j - e_i z_j \theta_j)]}{1 + \exp[Da_i(\theta_j - b_i - d_i z_j - e_i z_j \theta_j)]}, \quad (12)$$

where p_{ji}, *D*, a_i, b_i, c_i and θ_j are defined as in Equation 1, d_i is a parameter which models uniform DIF and e_i, is a parameter modeling non-uniform DIF and z_j is a covariate for examinee *j*.

When we use the 3PL-C model in Equation 12 on two non-equivalent groups of examinees, we need conversion equations to put the ability and item parameters on the same scale. By using the previous notation, θ_{Ij}, a_{Ii}, b_{Ii}, and c_{Ii} can be put on the scale *J* by using Equations 13-18 as described in Propositions 1-6.

Proposition 1

The ability θ_{Ii} of the examinee *j* from scale I can be placed on scale J through the conversion

$$\theta_{Jj} = \theta_{Ij} + B + A. \quad (13)$$

Proposition 2

The discrimination a_{Ii} of the item i from scale I can be placed on scale J through the conversion

$$a_{Ji} = a_{Ii}. \quad (14)$$

Proposition 3

The difficulty b_{Ii} of the item i from scale I can be placed on scale J through the conversion

$$b_{Ji} = b_{Ii} + B + A. \quad (15)$$

Proposition 4

The pseudo-guessing c_{Ii} of the item i from scale I can be placed on scale J through the conversion

$$c_{Ji} = c_{Ii}. \quad (16)$$

Proposition 5

The uniform DIF parameter d_{Ii} of item *i* from scale I can be placed on scale J through the conversion

$$d_{Ji} = d_{Ii} - Be_{Ii}. \quad (17)$$

Proposition 6

The non-uniform DIF parameter e_{Ii} of item *i* from scale I can be placed on the scale J through the conversion

$$e_{Ji} = e_{Ii}. \quad (18)$$

Proof of the appropriateness of proposed conversion equations in propositions 1-6

By substituting Equations 13-18 in Equation 12 we obtain

$$p_{ji} = c_{Ii} + (1 - c_{Ii}) \frac{\exp\left[Da_{Ii}(\theta_{Ij}+B+A-b_{Ii}-B-A-d_{Ii}z_j+Be_{Ii}z_j-e_{Ii}z_j\theta_{Ij}-Be_{Ii}z_j)\right]}{1+\exp\left[Da_{Ii}(\theta_{Ij}+B+A-b_{Ii}-B-A-d_{Ii}z_j+Be_{Ii}z_j-e_{Ii}z_j\theta_{Ij}-Be_{Ii}z_j)\right]}, \quad (19)$$

which, after simplifications becomes:

$$p_{ji} = c_{Ii} + (1-c_{Ii}) \frac{\exp[Da_{Ii}(\theta_{Ij} - b_{Ii} - d_{Ii}z_j - e_{Ii}z_j\theta_{Ij})]}{1+\exp[Da_{Ii}(\theta_{Ij} - b_{Ii} - d_{Ii}z_j - e_{Ii}z_j\theta_{Ij})]}. \quad (20)$$

Equation 20 shows that, by using the proposed conversion equations, the probability of a correct response is the same regardless of which scale is used.

The coefficients *A* and *B* can be estimated by using the mean/mean and the mean/sigma method as shown in Equations 6-8. The Haebara and Stocking-Lord methods, however have to be slightly modified to consider the new parameters. The updated Haebara method will minimize the following quantity:

$$\sum_j \sum_{i \in V} [p_{ji}(\theta_{Jj}, \hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji}, \hat{d}_{Ji}, \hat{e}_{Ji}) - p_{ji}(\theta_{Jj}, \hat{a}_{Ii}, \hat{b}_{Ii} + B + A, \hat{c}_{Ii}, \hat{d}_{Ii} - B\hat{e}_{Ii}, \hat{e}_{Ii})]^2, \quad (21)$$

where $\hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji}, \hat{d}_{Ji}, \hat{e}_{Ji}$ are the estimates of the corresponding quantities on scale *J* and $\hat{a}_{Ii}, \hat{b}_{Ii}, \hat{c}_{Ii}, \hat{d}_{Ii}, \hat{e}_{Ii}$ are the estimates of the corresponding quantities on scale *I*. The partial derivatives of the Haebara minimization criterion for the 3PL IRT-C model are given in Appendix A.

The updated Stocking-Lord method calculates *A* and *B* by minimizing the following quantity:

$$\sum_j [\sum_{i \in V} p_{ji}(\theta_{Jj}, \hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji}, \hat{d}_{Ji}, \hat{e}_{Ji}) - \sum_{i \in V} p_{ji}(\theta_{Jj}, \hat{a}_{Ii}, \hat{b}_{Ii} + B + A, \hat{c}_{Ii}, \hat{d}_{Ii} - B\hat{e}_{Ii}, \hat{e}_{Ii})]^2, \quad (22)$$

where $\hat{a}_{Ji}, \hat{b}_{Ji}, \hat{c}_{Ji}, \hat{d}_{Ji}, \hat{e}_{Ji}$ and $\hat{a}_{Ii}, \hat{b}_{Ii}, \hat{c}_{Ii}, \hat{d}_{Ii}, \hat{e}_{Ii}$ are defined as in the previous equation. The partial derivatives of the Stocking-Lord's minimization criterion for the 3PL-C model in Equation 12 are given in Appendix A.

To illustrate the proposed scale linking samples from the Swedish Scholastic Assessment test (SweSAT) was used. The SweSAT is a paper and pencil test used in Sweden to select students for admissions to college or university. The test is composed of 160 multiple-choice items divided into two sections. A verbal section articulated in four subsections, each of which contains 20 items (Vocabulary, Swedish reading comprehension, English reading comprehension and Sentence completion). A quantitative section divided in four subsections with a different number of items (Data sufficiency containing 12 items, Diagrams, tables and maps containing 24 items, Mathematical problem solving containing 24 items and Quantitative comparisons containing 20 items). The two sections are scored and equated separately. In this study we only used the quantitative section from two administrations; labeled test form 1 test form 2 for which an external anchor test of 40 items was available. Thus, the examinees were administered 120 items at each administration (the 80 unique items of the quantitative section and the 40 items anchor test). The covariate gender was used in the DIF analysis and it was coded 0 for males and 1 for females. We chose the covariate gender because DIF on a demographic variable can result in DIF across scales. A sample of N = 1997 examinees who took the first administration (males comprised 43.9% of the sample, and females comprised the remaining 56.1%) was used. Another sample of N = 2014 examinees who took the second administration (males comprised 42.8% of the sample and females comprised the remaining 57.2%) was also used. Table 1 displays that the male examinees obtained higher mean score than the female examinees on the overall test and on the anchor test in both administrations.

Table 1. Means and standard deviations of the scores (SweSAT)

Administration	SweSAT		Anchor test	
	Male	Female	Male	Female
Test form 1	39.65 (12.38)	34.83(10.63)	18.66 (6.45)	15.49 (5.78)
Test form 2	38.89 (12.46)	32.66 (11.02)	19.45 (6.65)	16.81 (5.59)

The simulation study and the empirical study were carried out in R (R Core Development Team, 2019). The items parameters a_i, b_i, c_i were estimated using the package ltm (Rizopoulos, 2006) while the uniform and non-uniform DIF coefficients (d_i and e_i, respectively) were estimated using the package difR (Magis, Beland, Tuerlinckx, & De Boeck, 2010), which accepts as input the estimates obtained by using the package ltm. All codes can be obtained from the corresponding author upon request.

To estimate the uniform and non-uniform DIF coefficients (d_i and e_i, respectively) we used logistic regression (Swaminathan & Rogers, 1990), which was implemented using the package difR. In our example, males are the reference subgroup (g = 0), and females are the focal subgroup (g = 1). The method fits the following three models:

$$M_0 : \text{logit}(\pi_g) = \alpha + \beta X + \gamma_g + \delta_g X \quad (23)$$

$$M_1 : \text{logit}(\pi_g) = \alpha + \beta X + \gamma_g \quad (24)$$

$$M_2 : \text{logit}(\pi_g) = \alpha + \beta X \quad (25)$$

where π_g is the probability of answering correctly the item in group g and X is the matching variable (we used the raw score). Parameters α and β are the intercept and the slope of the logistic curves. For identification reasons, the parameters γ_0 and δ_0 for reference subgroup (g = 0) are set to zero, while the

parameters γ_1 and δ_1 of the focal subgroup (g = 1) represent, respectively, the uniform DIF effect and the non-uniform DIF effect. For the uniform DIF we test the hypothesis $\gamma_1=0$ by comparing models M_1 in Equation 24 and M_2 in Equation 25; for the non-uniform DIF we test the hypothesis $\delta_1=0$, by comparing the models M_0 in Equation 23 and M_1 in Equation 24. To compare the nested models we used the likelihood ratio test statistic. The results are displayed in Table B.1. To minimize Equations 21 and 22 we used the package pracma (Borchers, 2017).

RESULTS

In Tables 2, the coefficients A and B estimated by using the mean/mean, mean/sigma, Haebara and Stocking-Lord methods are shown. The difference between the different rows is that different initial values have been used to implement the Haebara and Stocking-Lord methods. In the first two rows initial values of the coefficients were estimated by using the mean/mean method. In the next two rows, the mean/sigma method was used and in the last two rows random values (sampled between 0 and 2 for A and between -1 and 0 for B) were used. The coefficients estimated by using the Haebara and Stocking-Lord methods are always identical with each other and smaller than those estimated by using the mean/mean and the mean/sigma methods.

Table 2. Estimated coefficients A and B

	MEAN/MEAN	MEAN/SIGMA	HAEBARA	STOCKING-LORD
M/M				
A	1.1011	1.2284	0.7076	0.7076
B	-0.1833	-0.3245	-0.5767	-0.5767
M/S				
A	1.1011	1.2284	0.8419	0.8419
B	-0.1833	-0.3245	-0.7110	-0.7110
RV				
A	1.1011	1.2284	0.6989	0.4869
B	-0.1833	-0.3245	-0.5718	-0.3562

In Table B.1, in the Appendix, estimated common items parameters and uniform and non-uniform DIF coefficients before linking with their means and standard deviations for both test forms 1 and 2 are shown. From the obtained means of the estimated item parameter values, we notice that all the item parameters are higher on test form 2. Thus it appears that test form 2 is composed by items more discriminating and more difficult than items in test form 1, and also the probabilities of guessing an answer and the differences in answering between groups are higher for test form 2 than for test form 1. It is also important to underline that the estimated item parameters (except the item difficulties) and the DIF coefficients have lower standard deviations on test form 1. However, the parameters are not on the same scale, thus to compare them correctly, it is necessary to link them. Regarding the DIF parameters we are displaying those that are significant on level 0.05 in bold.

It is extremely important to underline that, where the uniform DIF and/or the non-uniform DIF parameter was negligible, we put the corresponding coefficient d_i and/or e_i equal to 0 in Equation 12, before estimating *A* and *B*. Indeed, according to Zumbo and Thomas (1997), DIF can be negligible, moderate or large. We used their proposal of using the difference between Nagelkerke's R^2 (Nagelkerke, 1991) of two logistic models, denoted ΔR^2, as the effect size of DIF. If $\Delta R^2 \leq 0.13$, then DIF is negligible; if $0.13 < \Delta R^2 \leq 0.26$, then DIF is moderate; if $\Delta R^2 > 0.26$, then DIF is large.

Next, we transformed the items parameters of test form 2 which are on Scale I to the scale J of test form 1 and the results are shown in Tables B.2-B.3 in the Appendix. Remembering that A and B are substantially identical for the Haebara and Stocking-Lord methods; we obviously obtain the same transformed item parameters and DIF coefficients for these two methods. Concerning the means and

the standard deviations, we found that there are only small differences between the different methods (for example, b_2 and d_2 are slightly lower when we use the Haebara and Stocking-Lord methods).

Now when the items parameters and DIF coefficients are on the same scale, we can compare our results with the means and the standard deviations of the item parameters of test form 1. Table 3 shows that there are no large differences between the mean and the standard deviations of the transformed values and the corresponding values for the test form 1. Figures B.1-B.3 in the Appendix show the test information functions (TIF) and the test characteristic curves (TCC) for test form 2 when we transformed the item parameters by using the coefficients A and B obtained from the mean-mean, mean-sigma and Haebara/Stocking-Lord methods. The TIF and the TCC are very similar for the mean-mean and the mean-sigma methods, while we underline that the TIF for the Haebara/Stocking-Lord methods show that the test is more informative for ability which goes from 1 to 2 (for the other methods the test is more informative for slightly higher values of ability) and the TCC show that the test discriminates better than the other tests, because the curve is more rapid.

Table 3. Means and standard deviations of the items parameters

	a1	b1	c1	d1	e1
Mean	1.195	1.095	0.174	0.006	-0.000
Sd	0.554	1.360	0.125	0.189	0.010
Scale I converted to Scale J using MM					
	a2	b2	c2	d2	e2
Mean	1.263	1.828	0.173	0.039	-0.002
Sd	0.613	1.448	0.128	0.166	0.010
Scale I converted to Scale J using MS					
	a2	b2	c2	d2	e2
Mean	1.263	1.814	0.173	0.039	-0.002
Sd	0.613	1.448	0.128	0.166	0.010
Scale I converted to Scale J using H and SL					
	a2	b2	c2	d2	e2
Mean	1.263	1.041	0.173	0.038	-0.002
Sd	0.613	1.448	0.128	0.166	0.010

Table 4 shows that the parameters estimated by using concurrent calibration have lower standard errors than those estimated by using IRT with (our proposal) and without (traditional) covariates. However, it is also true that IRT with covariates exhibits standard errors lower than IRT without covariates for the difficulty parameters and equal or very similar for the discrimination and guess parameters. Figures B.4-B.10 in the Appendix B show that the concurrent calibration produces a test much more informative (around a middle ability level) and better discriminating than the IRT with and without covariates. However, it is also true that, by using IRT with covariates, we obtain a more informative test than the IRT without covariates for high abilities' values. In two cases, IRT with covariates over performs the concurrent calibration; when using the mean-mean and the mean-sigma methods, we obtain a more informative test for very high abilities around 3 by using IRT with covariates than by using the concurrent calibration. This fact is clearly visible in Figures 1-2.

Table 4. Means and standard deviations of the items parameters

Scale I converted to Scale J						
IRT without covariates						
Scale I converted to Scale J using MM and MS						
	a2MM	b2MM	c2MM	a2MS	b2MS	c2MS
Mean	1.147	0.819	0.173	1.028	0.793	0.173
Sd	0.557	1.595	0.128	0.499	1.779	0.128
Scale I converted to Scale J using H and SL						
	a2H	b2H	c2H	a2SL	b2SL	c2SL
Mean	1.128	0.803	0.173	1.128	0.803	0.173
Sd	0.548	1.622	0.128	0.548	1.622	0.128
Concurrent calibration						
Scale I converted to Scale J using MM and MS						
	a2MM	b2MM	c2MM	a2MS	b2MS	c2MS
Mean	2.177	0.425	0.000	1.974	0.419	0.000
Sd	0.431	0.266	0.000	0.391	0.294	0.000
Scale I converted to Scale J using H and SL						
	a2H	b2H	c2H	a2SL	b2SL	c2SL
Mean	1.788	1.085	0.000	2.177	0.425	0.000
Sd	0.354	0.324	0.000	0.431	0.266	0.000

Figure 1. Comparison between the Test information functions (mean-mean)

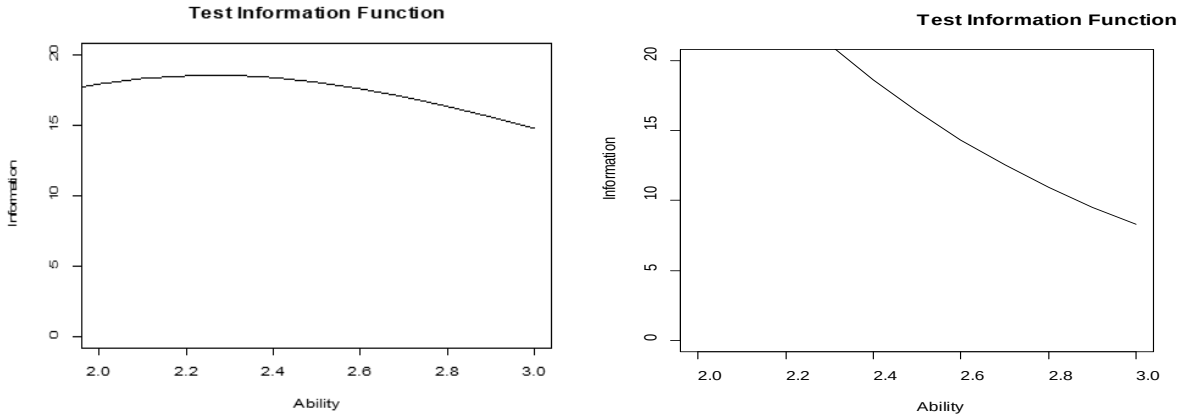
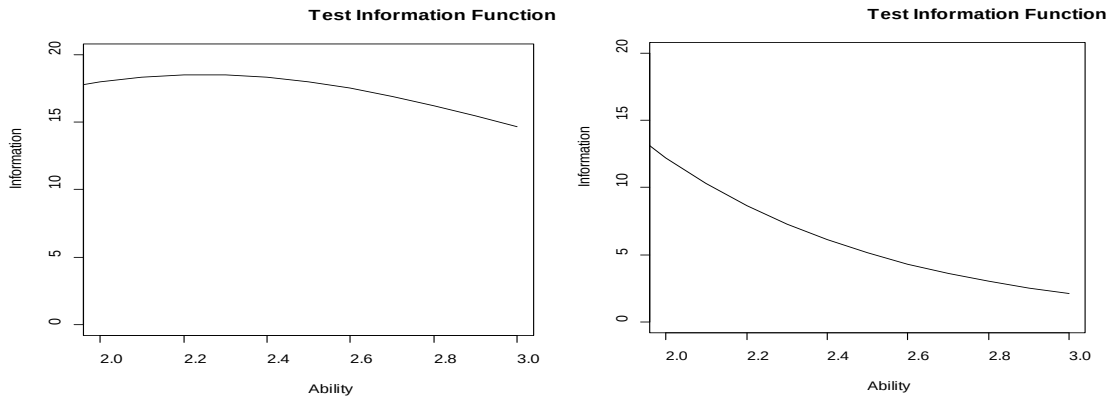


Figure 2. Comparison between the Test information functions (mean-sigma)



DISCUSSION

The main purpose of this study was to propose conversion equations to put the DIF parameters of the IRT-C model (Tay, Newman, & Vermunt, 2011; 2016) on the same scale when non-equivalent groups of examinees are administered test forms, which are scored with the IRT-C model. The research is important as assumption of equivalent groups of examinees is not always fulfilled when different test forms are administered over time (Lyrén & Hambleton, 2011). When we have an anchor test, as in the empirical example, the anchor test can be used to adjust the differences in ability between the groups who have taken different test forms (Kolen & Brennan, 2014). If an anchor test is not available, one could instead use covariates about examinees to correct differences between the examinees groups (Wiberg and Bränberg, 2015). When we do not have an anchor test, the best strategy to solve the problem of linking is to use the concurrent calibration. A problem is however that even in well-designed tests, there might be items with DIF for some subgroups. The proposed research aim to solve the problem to allow for DIF items and still make it possible to model the items with an IRT model which contains parameters for DIF. Thus, if we have an anchor with DIF items, the proposed approach is a good alternative to concurrent calibration. To use separate estimations instead of concurrent calibration has some benefits. For example, one benefit is that it is easier to diagnose potential problems if one estimates the item parameters separately in each group (Hanson & Béguin, 2002). The empirical study showed that the proposed conversion equation can be used with real test data. This is important as we can avoid removing DIF items when comparing test forms, especially if there are a large number of DIF items.

The results show that the IRT with covariates produces a more informative test than using IRT without covariates for high ability values and, when we use the mean-mean and the mean-sigma methods, it also gives more informative tests than concurrent calibration. Summing up, the proposed equations are important as it gives alternatives on how to link scales when we have information from covariates.

REFERENCES

Andersson, B. (2018). Asymptotic variance of linking coefficient estimators for polytomous IRT models. *Applied Psychological Measurement*, 42(3), 192-205.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord and M. R. Novick (Eds.), *Statistical theories of mental test scores* (chaps. 17-20). Reading, MA: Addison-Wesley.

Borchers, H. W. (2017). *Pracma: Practical numerical math functions*. R package.

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

Finch, H. (2005). The MIMIC model as a method for detecting DIF: Comparison with Mantel-Haenszel, SIBTEST, and the IRT likelihood ratio. *Applied Psychological Measuremen*t, 29, 278-295.

González, J., & Wiberg, M. (2017). *Applying test equating methods – using R*. Cham, Switzerland: Springer.

Haebara, T. (1980). Equating logistic ability scales by a weighted least squares method. *Japanese Psychological Research*, *22*, 144-149.

Hanson, B. A., & Béguin, A. A. (2002). Obtaining a common scale for item response theory item parameters using separate versus concurrent estimation in the common-item equating design. *Applied Psychological Measurement*, *26*, 3-24.

Huggins-Manley, A. C. (2014). The effect of differential item functioning in anchor items on population invariance of equating, *Educational and Psychological Measurement*, *74*(4), 627-658,

Jodoin, M. G., & Gierl, M. J. (2001). Evaluating type I error and power rates using an effect size measure with the logistic regression procedure for DIF detection. *Applied Measurement in Education*, 14, 329-349.

Kim, S. H., & Cohen, A. S. (1998). A comparison of linking and concurrent calibration under item response theory. *Applied Psychological Measurement*, *22*, 131-143.

Kim, S., & Kolen, M. J. (2007). Effects on scale linking of different definitions of criterion functions for the IRT characteristic curve methods. *Journal of Educational and Behavioral Statistics*, *32*, 371-397.

Kolen, M., & Brennan, R. (2014). *Test equating, scaling, and linking: Method and practice*. 3rd edition New York, NY: Springer-Verlag.

Kopf, J., Zeileis, A., & Strobl, C. (2015). Anchor selection strategies for DIF analysis. Review, assessment and new approaches. *Educational and Psychological Measurement*, 75(1), 22-56.

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

Loyd, B. H., & Hoover, H. D. (1980). Vertical equating using the Rasch Model. *Journal of Educational Measurement, 17*, 179-193.

Lyrén, P-E., & Hambleton, R. K. (2011). Consequences of violated the equating assumptions under the equivalent group design. *International Journal of Testing*, *36*(5), 308-323.

Magis, D., Beland, S.,Tuerlinckx, F., & De Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*, 847-862.

Marco, G. L. (1977). Item characteristic curve solutions to three intractable testing problems. *Journal of Educational Measurement, 14*, 139-160.

Nagelkerke, N. J. D. (1991). A note on a general definition of the coefficient of determination. *Biometrika, 78(3)*, 691-692.

Ogasawara, H. (2001). Standard errors of item response theory equating/linking by response function methods. *Applied Psychological Measurement, 25*, 53-67.

Ogasawara, H. (2011). Applications of asymptotic expansion in item response theory linking. In von A. A. Davier, (Ed.), *Statistical models for test equating, scaling, and linking* (pp. 261-280). New York, NY: Springer.

R Core Development Team (2019). *R: A language and Environment for Statistical Computing*. R Foundation for statistical computing. Vienna, Austria: http://www.R-project.org/

Rizopoulos, D. (2006). ltm: An R package for latent variable modelling and item response theory analyses. *Journal of Statistical Software, 17* (5), 1-25.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2006). Detecting differential item functioning with confirmatory factor analysis and item response theory: Toward a unified strategy. *Journal of Applied Psychology*, 91, 1292-1306.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Swaminathan, H., & Rogers, H. J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 27*, 361-370.

Tay, L., Newman, D., & Vermunt, J. K. (2011). Using mixed-measurement item response theory with covariates (MM-IRT-C) to ascertain observed and unobserved measurement equivalence. *Organizational Research Methods, 1*(14), 147-176.

Tay, L., Newman, D., & Vermunt, J. K. (2016). Item response theory with covariates (IRT-C): Assessing item recovery and differential item functioning for the three-parameter logistic model, *Educational and Psychological Measurement, 76*(1), 22-42.

Wang, W.-C. (2004). Effects of anchor item methods on the detection of differential item functioning within the family of Rasch models. *Journal of Experimental Education*, 72, 221-261.

Wang, W.-C., & Yeh, Y.-L. (2003). Effects of anchor item methods on differential item functioning detection with the likelihood ratio test. *Applied Psychological Measurement*, 27, 479-498.

Wang, W.-C., Shih, C.-L., & Sun, G.-W. (2012). The DIF-free-then-DIF strategy for the assessment of differential item functioning. *Educational and Psychological Measurement*, 72, 687-708.

Wedman, J. (2018). Reasons for gender-related differential item functioning in a college admissions test. *Journal of Educational Research, 62*(6), 959-970.

Wiberg, M., & Bränberg, K. (2015). Kernel equating under the non-equivalent groups with covariates design. *Applied Psychological Measurement, 39(5)*, 1-13.

Zumbo, B. D., & Thomas, D. R. (1997). A measure of effect size for a model-based approach for studying DIF. *Working paper of the Edgeworth Laboratory for Quantitative Behavioral Sciences*. Prince George, Canada: University of British Columbia.

Appendixes

The partial derivatives of the Haebara minimization criterion for the 3PL-C. Let H indicate the Haebara minimization criterion (Equation 21), then we obtain:

$$\frac{\partial H}{\partial A} = \sum_j \sum_{i \in V} 2 \{ [c_{Ji} + (1 - c_{Ji}) \frac{\exp[Da_{Ji}(\theta_{Jj} - b_{Ji} - d_{Ji}z_j - e_{Ji}z_j\theta_{Jj})]}{1 + \exp[Da_{Ji}(\theta_{Jj} - b_{Ji} - d_{Ji}z_j - e_{Ji}z_j\theta_{Jj})]}] -$$
$$\left[c_{Ii} + (1 - c_{Ii}) \frac{\exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})]}{1 + \exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})]} \right]$$
$$\{ -c_{Ii} + (1 - c_{Ii}) \frac{\exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})] [D(-a_{Ii})]}{(1 + \exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})])^2} \}$$

and

$$\frac{\partial H}{\partial B} = \sum_j \sum_{i \in V} 2 \{ [c_{Ji} + (1 - c_{Ji}) \frac{\exp[Da_{Ji}(\theta_{Jj} - b_{Ji} - d_{Ji}z_j - e_{Ji}z_j\theta_{Jj})]}{1 + \exp[Da_{Ji}(\theta_{Jj} - b_{Ji} - d_{Ji}z_j - e_{Ji}z_j\theta_{Jj})]}] -$$
$$\left[c_{Ii} + (1 - c_{Ii}) \frac{\exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})]}{1 + \exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})]} \right]$$
$$\{ -c_{Ii} + (1 - c_{Ii}) \frac{\exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})] [D(-a_{Ii} + a_{Ii}e_{Ii}z_j)]}{(1 + \exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})])^2} \}$$

The partial derivatives of the Stocking-Lord's minimization criterion for the 3PL-C model in Equation 12 is as follows. Let SL indicate the Stocking-Lord's minimization criterion (Equation 22), then we obtain:

$$\frac{\partial S}{\partial A} = \sum_j 2 \{ \sum_{i \in V} c_{Ji} + (1 - c_{Ji}) \frac{\exp[Da_{Ji}(\theta_{Jj} - b_{Ji} - d_{Ji}z_j - e_{Ji}z_j\theta_{Jj})]}{1 + \exp[Da_{Ji}(\theta_{Jj} - b_{Ji} - d_{Ji}z_j - e_{Ji}z_j\theta_{Jj})]} -$$
$$\sum_{i \in V} c_{Ji} + (1 - c_{Ji}) \frac{\exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})]}{1 + \exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})]}$$
$$\{ \sum_{i \in V} [-c_{Ii} + (1 - c_{Ii}) \frac{\exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})] [D(-a_{Ii})]}{(1 + \exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})])^2}] \}$$

and

$$\frac{\partial S}{\partial B} = \sum_j 2 \{ \sum_{i \in V} c_{Ji} + (1 - c_{Ji}) \frac{\exp[Da_{Ji}(\theta_{Jj} - b_{Ji} - d_{Ji}z_j - e_{Ji}z_j\theta_{Jj})]}{1 + \exp[Da_{Ji}(\theta_{Jj} - b_{Ji} - d_{Ji}z_j - e_{Ji}z_j\theta_{Jj})]} -$$
$$\sum_{i \in V} c_{Ji} + (1 - c_{Ji}) \frac{\exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})]}{1 + \exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})]}$$
$$\{ \sum_{i \in V} [-c_{Ii}$$
$$+ (1 - c_{Ii}) \frac{\exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})] [D(-a_{Ii} + a_{Ii}e_{Ii}z_j)]}{(1 + \exp[Da_{Ii}(\theta_{Jj} - b_{Ii} - B - A - d_{Ii}z_j + Be_{Ii}z_j - e_{Ii}z_j\theta_{Jj})])^2}] \}$$

Table B.1. Estimated common items parameters before linking

	SCALE J					SCALE I				
item	a1	b1	c1	d1	e1	a2	b2	c2	d2	e2
1	0.651	-0.506	0.000	**-0.437**	0.011	0.906	-0.112	0.110	**-0.410**	-0.001
2	1.420	0.328	0.299	-0.060	0.010	1.133	0.610	0.289	-0.100	-0.004
3	1.517	0.843	0.157	**0.301**	-0.002	2.591	0.970	0.188	0.185	0.007
4	1.230	0.700	0.303	**0.589**	0.000	1.654	1.152	0.355	**0.626**	0.010
5	0.677	-1.165	0.001	**0.231**	0.010	0.715	-0.679	0.000	**0.302**	**0.015**
6	1.360	0.857	0.091	0.104	**0.017**	1.578	1.365	0.118	0.092	0.012
7	0.319	0.126	0.001	0.089	0.001	0.434	1.285	0.115	-0.024	0.000
8	1.579	1.204	0.163	0.128	0.005	2.221	1.153	0.175	-0.156	-0.014
9	1.992	1.351	0.074	-0.003	-0.003	2.384	1.387	0.080	0.001	**-0.020**
10	1.782	1.773	0.218	-0.166	0.010	1.867	1.747	0.196	0.104	-0.013
11	1.648	0.604	0.184	0.129	0.010	1.783	0.917	0.253	0.021	-0.009
12	1.649	1.917	0.133	-0.076	-0.004	1.792	1.995	0.134	0.216	**-0.016**
13	1.091	1.257	0.346	-0.044	-0.001	1.011	1.476	0.339	0.182	-0.005
14	2.269	1.304	0.177	0.119	-0.009	2.321	1.392	0.133	**-0.236**	-0.002
15	1.285	1.828	0.374	-0.060	-0.009	0.691	1.725	0.337	-0.001	-0.011
16	1.174	0.326	0.163	-0.011	0.003	1.133	0.383	0.166	0.072	**-0.019**
17	0.907	0.149	0.110	-0.155	-0.002	1.428	0.745	0.259	-0.125	-0.011
18	0.906	-0.170	0.099	0.067	**0.019**	1.130	0.390	0.205	-0.045	0.002
19	1.444	0.522	0.199	-0.031	**-0.018**	1.940	0.787	0.213	**-0.209**	**-0.030**
20	1.324	1.295	0.243	0.032	0.001	1.676	1.639	0.268	0.054	-0.001
21	1.747	1.235	0.098	-0.064	**-0.020**	2.062	1.526	0.122	-0.178	**-0.018**
22	1.091	1.457	0.101	-0.104	0.012	1.392	1.622	0.142	-0.169	-0.010
23	1.760	0.154	0.266	**-0.287**	-0.001	1.461	-0.081	0.153	**-0.327**	-0.009
24	1.136	0.899	0.123	-0.064	0.003	0.977	0.960	0.127	0.061	-0.004
25	0.634	1.647	0.000	**-0.474**	-0.001	0.695	1.264	0.000	**-0.249**	0.011
26	0.718	1.200	0.065	0.003	-0.012	0.808	1.521	0.119	-0.152	-0.006
27	0.994	1.321	0.020	0.019	0.001	1.568	1.227	0.093	0.169	-0.015
28	1.404	1.748	0.183	0.131	-0.003	1.767	1.599	0.190	-0.144	-0.004
29	0.593	-1.543	0.059	-0.147	**0.020**	0.650	-1.362	0.000	**-0.217**	0.010
30	0.552	-0.592	0.000	0.000	0.008	0.797	0.207	0.169	-0.155	-0.011
31	0.294	0.138	0.037	-0.162	0.004	0.294	-0.232	0.000	0.111	0.001
32	0.697	0.523	0.181	-0.140	0.012	0.633	0.181	0.114	-0.189	0.006
33	0.924	1.930	0.318	-0.041	-0.002	1.789	1.907	0.406	0.111	**-0.019**
34	1.445	1.919	0.247	0.029	-0.008	1.383	1.820	0.229	**-0.236**	**-0.020**
35	0.522	1.905	0.185	**0.198**	-0.002	0.339	1.180	0.000	0.016	0.008
36	0.844	1.420	0.131	**-0.343**	0.000	0.729	1.182	0.059	**-0.529**	0.005
37	0.382	4.018	0.011	-0.170	0.002	0.671	2.362	0.044	-0.112	0.005
38	0.461	2.005	0.103	-0.086	0.002	0.356	1.276	0.001	**-0.209**	0.012
39	2.057	2.727	0.280	0.065	-0.013	0.763	3.030	0.255	0.076	-0.001
40	1.741	2.860	0.295	**0.286**	-0.007	1.371	2.842	0.289	**0.208**	-0.011
Mean	1.156	1.038	0.151	-0.015	0.001	1.272	1.109	0.161	-0.039	-0.005
SD	0.518	1.075	0.107	0.196	0.009	0.618	0.875	0.107	0.213	0.011

Table B.2. Transformed items parameters (mean/mean and mean/sigma)

	Scale I converted to Scale J using MM					Scale I converted to Scale J using MS				
item	a2	b2	c2	d2	e2	a2	b2	c2	d2	e2
1	2.378	1.538	0.283	0.010	0.004	2.378	1.524	0.283	0.011	0.004
2	0.498	0.731	0.032	0.014	-0.016	0.498	0.717	0.032	0.012	-0.016
3	2.119	2.127	0.356	0.064	-0.007	2.119	2.113	0.356	0.063	-0.007
4	2.306	1.292	0.214	0.209	0.008	2.306	1.278	0.214	0.210	0.008
5	1.789	1.848	0.288	0.077	-0.010	1.789	1.834	0.288	0.075	-0.010
6	2.178	1.846	0.247	0.161	-0.009	2.178	1.832	0.247	0.159	-0.009
7	2.311	1.930	0.328	0.365	0.004	2.311	1.916	0.328	0.365	0.004
8	1.572	2.110	0.228	0.281	-0.004	1.572	2.096	0.228	0.280	-0.004
9	1.775	2.365	0.341	0.151	0.000	1.775	2.352	0.341	0.151	0.000
10	0.538	2.601	0.000	-0.051	0.007	0.538	2.587	0.000	-0.050	0.007
11	1.342	1.862	0.149	0.180	-0.005	1.342	1.849	0.149	0.179	-0.005
12	1.867	2.782	0.176	-0.014	-0.018	1.867	2.768	0.176	-0.016	-0.018
13	0.855	1.334	0.172	-0.211	0.003	0.855	1.320	0.172	-0.211	0.003
14	0.825	1.547	0.281	-0.153	-0.005	0.825	1.533	0.281	-0.153	-0.005
15	0.873	0.827	0.000	-0.113	0.014	0.873	0.813	0.000	-0.111	0.014
16	1.035	1.273	0.085	0.058	0.010	1.035	1.259	0.085	0.059	0.010
17	0.693	1.595	0.122	-0.052	-0.004	0.693	1.581	0.122	-0.052	-0.004
18	1.188	1.575	0.152	0.136	0.006	1.188	1.561	0.152	0.137	0.006
19	1.826	1.865	0.230	0.294	-0.012	1.826	1.851	0.230	0.292	-0.012
20	2.210	2.087	0.271	0.283	-0.013	2.210	2.073	0.271	0.281	-0.013
21	2.235	3.556	0.390	0.040	-0.007	2.235	3.542	0.390	0.039	-0.007
22	1.963	2.724	0.353	-0.248	-0.005	1.963	2.710	0.353	-0.248	-0.005
23	0.808	1.546	0.032	-0.227	0.003	0.808	1.532	0.032	-0.227	0.003
24	1.377	2.093	0.112	-0.014	-0.014	1.377	2.080	0.112	-0.016	-0.014
25	1.768	1.697	0.200	-0.039	0.012	1.768	1.684	0.200	-0.037	0.012
26	1.206	1.544	0.099	0.008	-0.008	1.206	1.530	0.099	0.007	-0.008
27	0.772	0.785	0.000	0.119	0.006	0.772	0.771	0.000	0.120	0.006
28	2.060	2.847	0.153	0.153	-0.023	2.060	2.833	0.153	0.150	-0.023
29	0.995	0.339	0.007	-0.087	-0.003	0.995	0.325	0.007	-0.087	-0.003
30	0.769	0.976	0.079	0.040	-0.001	0.769	0.962	0.079	0.040	-0.001
31	0.622	2.012	0.000	-0.234	-0.005	0.622	1.998	0.000	-0.235	-0.005
32	0.630	1.842	0.017	0.163	-0.001	0.630	1.828	0.017	0.163	-0.001
33	0.743	0.726	0.000	-0.056	-0.003	0.743	0.712	0.000	-0.057	-0.003
34	0.909	1.475	0.163	0.104	-0.011	0.909	1.461	0.163	0.102	-0.011
35	1.077	2.656	0.328	-0.172	0.000	1.077	2.642	0.328	-0.172	0.000
36	1.066	2.226	0.193	0.212	-0.026	1.066	2.213	0.193	0.208	-0.026
37	1.630	3.978	0.314	-0.146	-0.004	1.630	3.964	0.314	-0.146	-0.004
38	0.852	3.178	0.303	-0.051	-0.003	0.852	3.164	0.303	-0.051	-0.003
39	2.467	3.424	0.242	0.056	-0.004	2.467	3.410	0.242	0.055	-0.004
40	0.987	5.113	0.300	0.123	-0.017	0.987	5.099	0.300	0.121	-0.017
41	1.274	0.229	0.000	0.110	0.020	1.274	0.215	0.000	0.113	0.020
42	1.261	1.926	0.430	0.227	0.011	1.261	1.912	0.430	0.229	0.011
43	1.607	0.777	0.035	-0.022	0.003	1.607	0.763	0.035	-0.021	0.003

44	2.436	2.159	0.235	0.417	-0.004	2.436	2.145	0.235	0.416	-0.004
45	2.221	1.540	0.293	-0.193	-0.020	2.221	1.526	0.293	-0.196	-0.020
46	1.013	1.622	0.301	-0.080	0.013	1.013	1.608	0.301	-0.078	0.013
47	1.540	2.357	0.352	0.089	-0.004	1.540	2.343	0.352	0.089	-0.004
48	1.874	1.719	0.244	-0.161	-0.003	1.874	1.705	0.244	-0.161	-0.003
49	1.573	1.940	0.185	0.271	0.005	1.573	1.926	0.185	0.272	0.005
50	1.957	2.700	0.228	-0.025	-0.002	1.957	2.686	0.228	-0.026	-0.002
51	1.571	2.117	0.179	-0.113	-0.006	1.571	2.103	0.179	-0.114	-0.006
52	1.656	2.871	0.314	0.211	-0.017	1.656	2.857	0.314	0.209	-0.017
53	1.397	1.992	0.328	-0.089	-0.009	1.397	1.978	0.328	-0.090	-0.009
54	0.683	3.061	0.348	-0.148	0.001	0.683	3.047	0.348	-0.147	0.001
55	0.986	1.184	0.364	0.102	0.011	0.986	1.170	0.364	0.103	0.011
56	1.228	0.973	0.124	0.037	0.004	1.228	0.959	0.124	0.037	0.004
57	2.673	2.456	0.245	-0.173	0.003	2.673	2.443	0.245	-0.173	0.003
58	0.562	1.837	0.121	-0.205	-0.017	0.562	1.823	0.121	-0.207	-0.017
59	0.572	1.985	0.098	0.236	0.002	0.572	1.971	0.098	0.236	0.002
60	1.181	2.351	0.097	-0.061	0.010	1.181	2.337	0.097	-0.059	0.010
61	1.431	3.177	0.163	0.105	-0.017	1.431	3.163	0.163	0.103	-0.017
62	0.985	3.114	0.147	-0.100	0.001	0.985	3.100	0.147	-0.100	0.001
63	1.167	0.725	0.000	0.143	0.006	1.167	0.711	0.000	0.144	0.006
64	0.916	0.371	0.000	0.182	0.006	0.916	0.357	0.000	0.183	0.006
65	1.738	2.235	0.345	-0.006	-0.018	1.738	2.221	0.345	-0.009	-0.018
66	0.526	1.823	0.001	-0.239	-0.001	0.526	1.809	0.001	-0.239	-0.001
67	0.532	1.149	0.000	0.328	0.007	0.532	1.136	0.000	0.329	0.007
68	1.189	2.666	0.105	0.095	-0.005	1.189	2.652	0.105	0.094	-0.005
69	0.674	0.307	0.000	0.256	0.015	0.674	0.294	0.000	0.258	0.015
70	0.812	0.016	0.000	0.336	-0.003	0.812	0.003	0.000	0.336	-0.003
71	0.704	1.540	0.000	0.066	0.018	0.704	1.526	0.000	0.069	0.018
72	0.951	1.263	0.171	-0.220	0.006	0.951	1.249	0.171	-0.219	0.006
73	0.371	1.075	0.000	0.271	0.003	0.371	1.061	0.000	0.271	0.003
74	0.881	2.988	0.246	-0.104	-0.006	0.881	2.974	0.246	-0.105	-0.006
75	0.816	1.656	0.031	0.232	0.001	0.816	1.642	0.031	0.232	0.001
76	1.054	2.853	0.263	0.097	-0.005	1.054	2.839	0.263	0.096	-0.005
77	-0.104	-7.942	0.035	0.039	0.013	-0.104	-7.956	0.035	0.040	0.013
78	1.168	2.346	0.303	0.080	-0.002	1.168	2.332	0.303	0.080	-0.002
79	0.154	4.150	0.004	-0.171	0.001	0.154	4.136	0.004	-0.171	0.001
80	2.119	2.127	0.356	0.064	-0.007	2.119	2.113	0.356	0.063	-0.007
Mean	1.263	1.828	0.173	0.039	-0.002	1.263	1.814	0.173	0.039	-0.002
SD	0.613	1.448	0.128	0.166	0.010	0.613	1.448	0.128	0.166	0.010

Table B.3. Transformed items parameters (Haebara and Stocking-Lord)

Scale I converted to Scale J by using H and SL					
Item	a2h	b2h	c2h	d2h	e2h
1	2.378	0.751	0.283	0.012	0.004
2	0..498	-0..056	0..032	0..008	-0..016
3	2.119	1.340	0.356	0.061	-0.007
4	2.306	0.505	0.214	0.212	0.008
5	1.789	1.061	0.288	0.073	-0.010
6	2.178	1.059	0.247	0.157	-0.009
7	2.311	1.143	0.328	0.366	0.004
8	1.572	1.323	0.228	0.279	-0.004
9	1.775	1.579	0.341	0.151	0.000
10	0.538	1.814	0.000	-0.049	0.007
11	1.342	1.075	0.149	0.178	-0.005
12	1.867	1.995	0.176	-0.021	-0.018
13	0.855	0.547	0.172	-0.210	0.003
14	0.825	0.760	0.281	-0.155	-0.005
15	0.873	0.040	0.000	-0.107	0.014
16	1.035	0.486	0.085	0.062	0.010
17	0.693	0.808	0.122	-0.053	-0.004
18	1.188	0.788	0.152	0.138	0.006
19	1.826	1.078	0.230	0.289	-0.012
20	2.210	1.300	0.271	0.277	-0.013
21	2.235	2.769	0.390	0.038	-0.007
22	1.963	1.937	0.353	-0.250	-0.005
23	0.808	0.759	0.032	-0.226	0.003
24	1.377	1.307	0.112	-0.020	-0.014
25	1.768	0.911	0.200	-0.034	0.012
26	1.206	0.757	0.099	0.005	-0.008
27	0.772	-0.002	0.000	0.121	0.006
28	2.060	2.060	0.153	0.144	-0.023
29	0.995	-0.448	0.007	-0.088	-0.003
30	0.769	0.189	0.079	0.040	-0.001
31	0.622	1.225	0.000	-0.236	-0.005
32	0.630	1.055	0.017	0.163	-0.001
33	0.743	-0.061	0.000	-0.058	-0.003
34	0.909	0.688	0.163	0.100	-0.011
35	1.077	1.869	0.328	-0.172	0.000
36	1.066	1.440	0.193	0.201	-0.026
37	1.630	3.191	0.314	-0.147	-0.004
38	0.852	2.391	0.303	-0.052	-0.003
39	2.467	2.637	0.242	0.054	-0.004
40	0.987	4.326	0.300	0.116	-0.017
41	1.274	-0.558	0.000	0.118	0.020
42	1.261	1.139	0.430	0.232	0.011
43	1.607	-0.010	0.035	-0.020	0.003

44	2.436	1.372	0.235	0.415	-0.004
45	2.221	0.753	0.293	-0.201	-0.020
46	1.013	0.835	0.301	-0.075	0.013
47	1.540	1.570	0.352	0.088	-0.004
48	1.874	0.932	0.244	-0.162	-0.003
49	1.573	1.153	0.185	0.273	0.005
50	1.957	1.913	0.228	-0.026	-0.002
51	1.571	1.330	0.179	-0.115	-0.006
52	1.656	2.084	0.314	0.205	-0.017
53	1.397	1.205	0.328	-0.093	-0.009
54	0.683	2.274	0.348	-0.147	0.001
55	0.986	0.397	0.364	0.106	0.011
56	1.228	0.186	0.124	0.038	0.004
57	2.673	1.670	0.245	-0.172	0.003
58	0.562	1.050	0.121	-0.212	-0.017
59	0.572	1.198	0.098	0.237	0.002
60	1.181	1.564	0.097	-0.057	0.010
61	1.431	2.390	0.163	0.098	-0.017
62	0.985	2.327	0.147	-0.100	0.001
63	1.167	-0.062	0.000	0.145	0.006
64	0.916	-0.416	0.000	0.185	0.006
65	1.738	1.448	0.345	-0.013	-0.018
66	0.526	1.036	0.001	-0.239	-0.001
67	0.532	0.363	0.000	0.331	0.007
68	1.189	1.879	0.105	0.093	-0.005
69	0.674	-0.479	0.000	0.262	0.015
70	0.812	-0.770	0.000	0.335	-0.003
71	0.704	0.753	0.000	0.073	0.018
72	0.951	0.476	0.171	-0.218	0.006
73	0.371	0.288	0.000	0.272	0.003
74	0.881	2.201	0.246	-0.107	-0.006
75	0.816	0.869	0.031	0.232	0.001
76	1.054	2.066	0.263	0.095	-0.005
77	-0.104	-8.729	0.035	0.044	0.013
78	1.168	1.559	0.303	0.080	-0.002
79	0.154	3.363	0.004	-0.170	0.001
80	0.675	2.265	0.273	-0.129	-0.004
Mean	1.263	1.041	0.173	0.038	-0.002
SD	0.613	1.448	0.128	0.166	0.010

Figure B.1. TI and TCC (mean-mean)

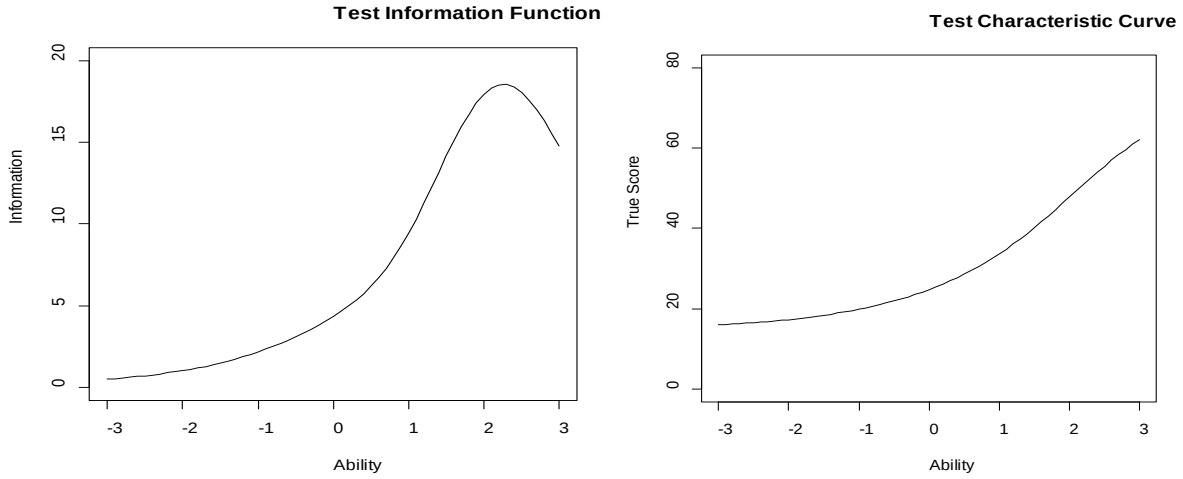


Figure B.2. TI and TCC (mean-sigma)

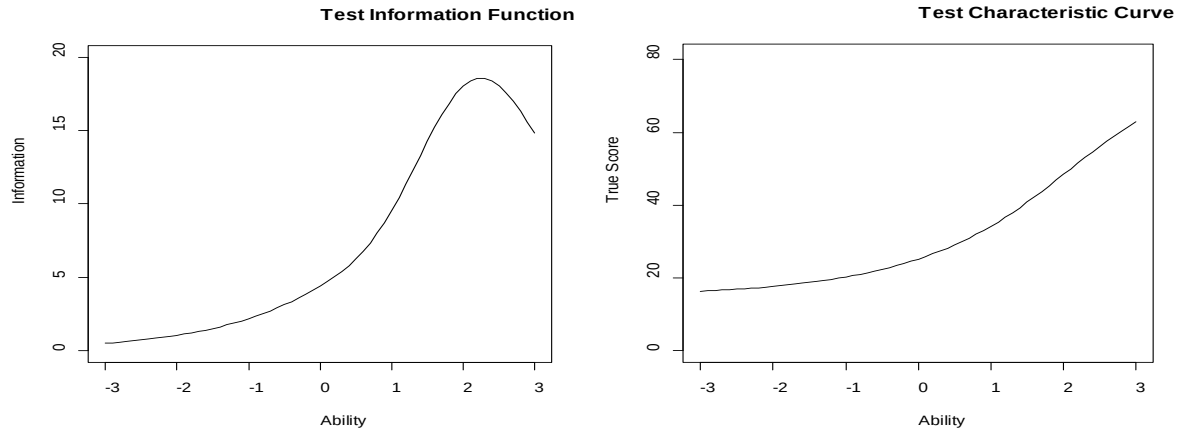


Figure B.3. TI and TCC (Haebara and Stocking-Lord)

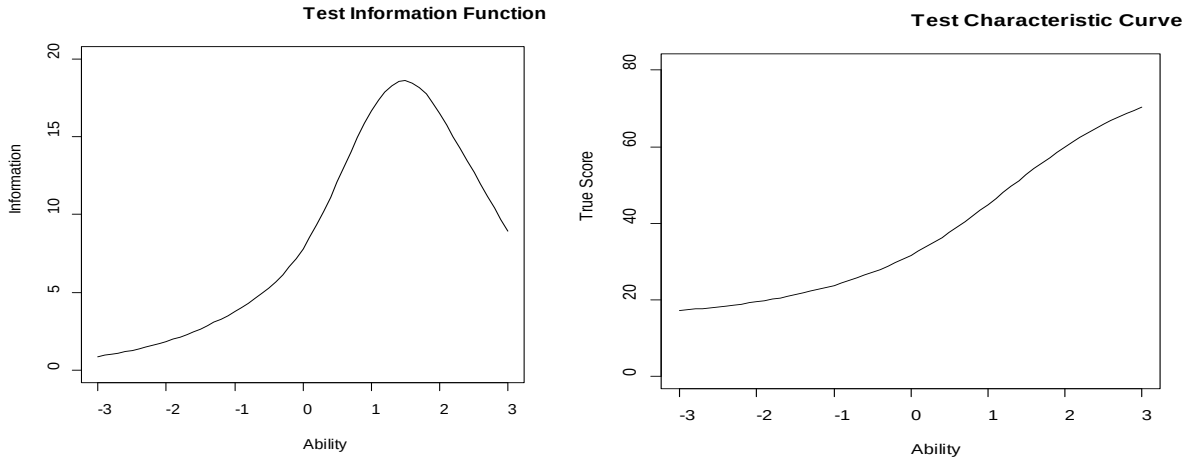


Figure B.4. TI and TCC (mean-mean, IRT without covariates)

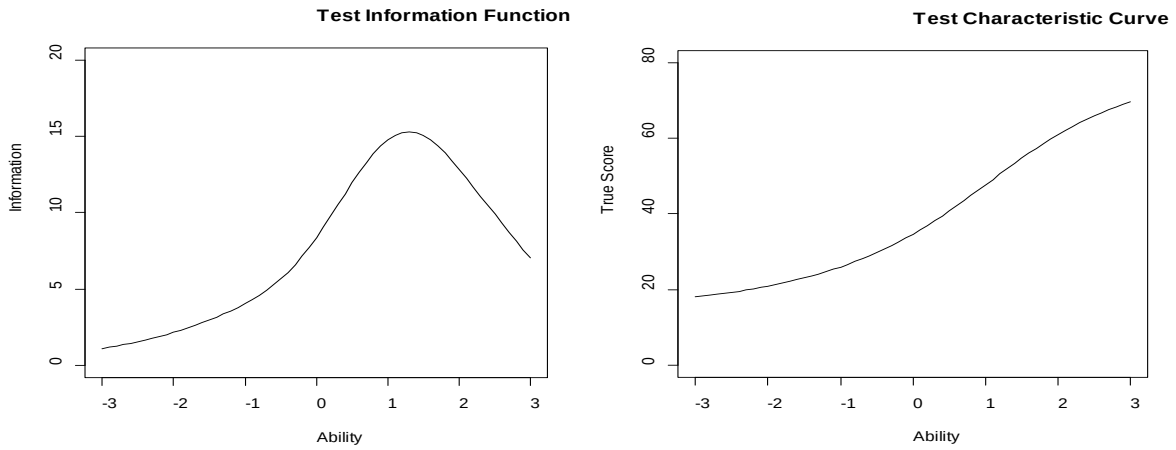


Figure B.5. TI and TCC (mean-sigma, IRT without covariates)

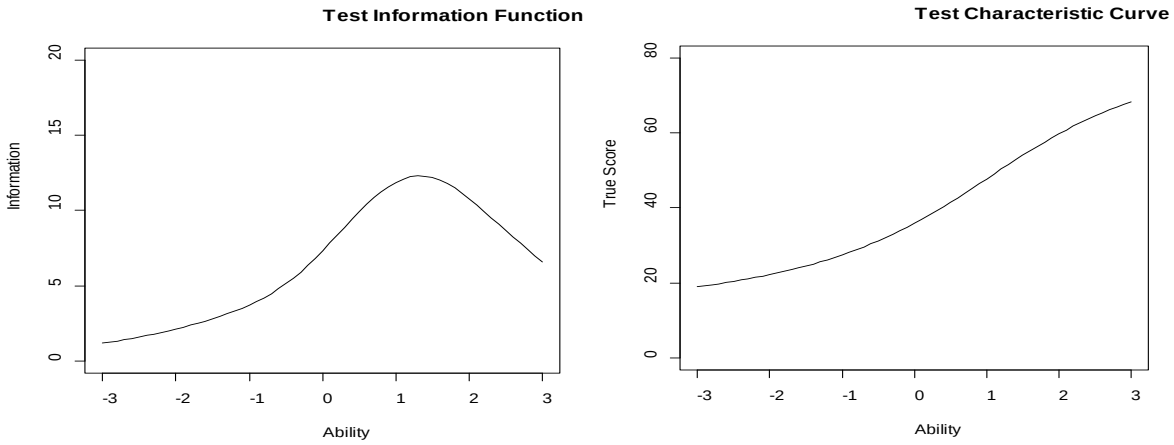


Figure B.6. TI and TCC (Haebara, IRT without covariates)

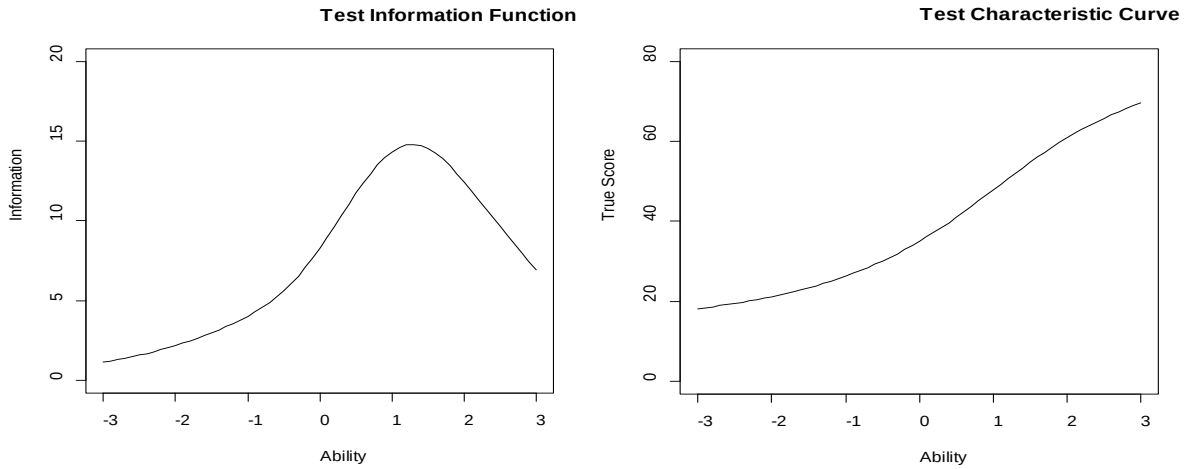


Figure B.7. TI and TCC (mean-mean, concurrent calibration)

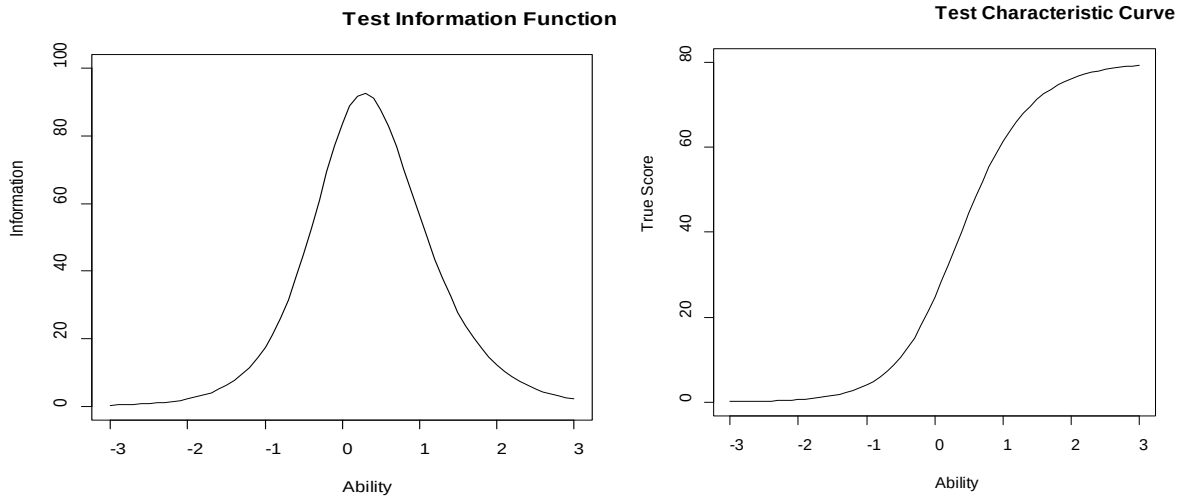


Figure B.8. TI and TCC (mean-sigma, concurrent calibration)

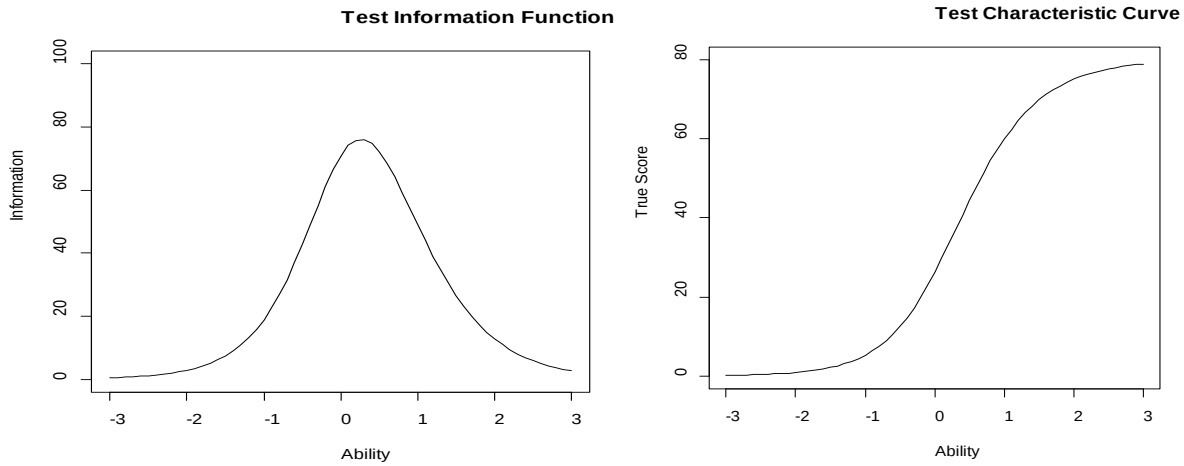


Figure B.9. TI and TCC (Haebara, concurrent calibration)

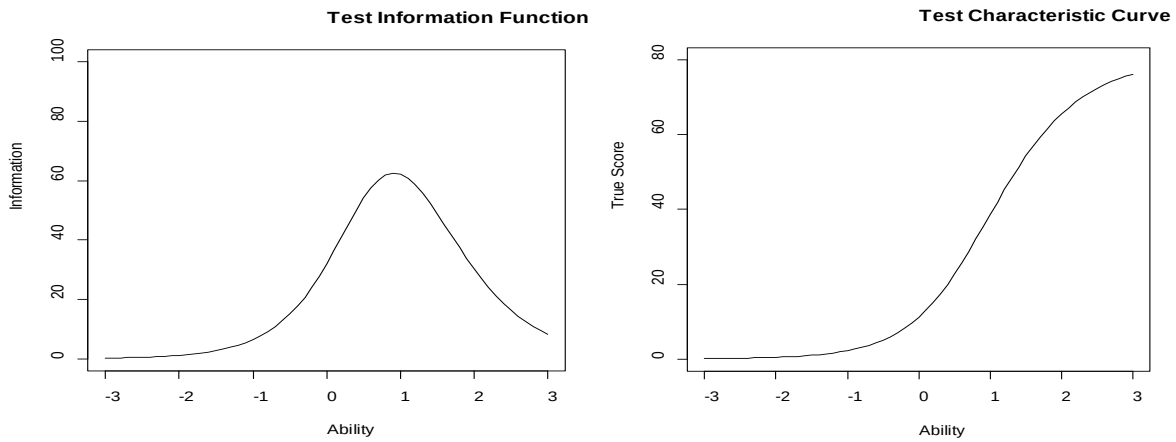


Figure B.10. TI and TCC (Stocking-Lord, concurrent calibration)