MEASUREMENT INVARIANCE OF THE GERIATRIC DEPRESSION SCALE SHORT FORM ACROSS LATIN AMERICA AND THE CARIBBEAN

By

Ola Stacey Rostant

A DISSERTATION

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILSOPHY

Measurement and Quantitative Methods

2011

ABSTRACT

MEASUREMENT INVARIANCE OF THE GERIATRIC DEPRESSION SCALE SHORT FORM ACROSS LATIN AMERICA AND THE CARIBBEAN

By

Ola Stacey Rostant

This study examined the measurement invariance of the Geriatric Depression Scale Short Form (GDS-15) in older adults across five countries in Latin America and the Caribbean. Multiple group confirmatory factor analysis (MGCFA) and item response theory likelihood ratio tests (IRTLR-DIF) were used to test the measurement invariance of the GDS-15 by gender and cross-country comparisons. The sample for this study was made up of 7,573 older adults between the ages of 60 and 102. Data for the present study comes from the Survey on Health Well-Being and Aging in Latin America and the Caribbean (SABE)(Pelaez, et al., 2004). The underlying factor structure of the GDS-15 was examined within each country. A one-factor structure was found for the countries of Chile and Cuba and a two-factor was found for the countries of Argentina, Mexico and Uruguay. Results of the multiple group confirmatory factor analysis offered support for full measurement invariance of a one-factor structure by gender within the countries of Chile and Cuba. Partial measurement invariance was found for a two-factor structure by gender within the countries of Argentina, Mexico and Uruguay by gender. The IRTLR-DIF analyses by gender and cross-country comparisons revealed a lack of parameter invariance. Items exhibiting DIF by gender were on average more difficult for men to endorse than women. In summary the IRTLR-DIF procedure identified more non-invariant items than the MGCFA procedure. Implications and recommendations for future research are addressed.

DEDICATION

To, JC, MS, DR and SR, without your support this would not have been possible.

ACKNOWLEDGEMENTS

Thank you to my committee for your insightful feedback. I extend my heartfelt thanks to my many mentors who have encouraged and supported me through my dissertation, David Lounsbury, Renee Canady, Clare Luz, Karen P. Williams and Clarissa Shavers.

I would like to thank the elders who participated in the Survey on Health and Well Being in Latin America and the Caribbean for their time and generosity.

Finally, I would like to thank my husband Matthew Swayne, for his patience, love and support. Without you I would not have been able to complete this process. Thank you to my friends and family who provided humor and encouragement along the way.

TABLE OF CONTENTS

LIST OF TABLES..... vii

LIST OF FIGURES..... xiv

CHAPTER 1: INTRODUCTION..... 1

INTRODUCTION..... 1

STUDY RATIONALE..... 4

CHAPTER 2: LITERATURE REVIEW..... 5

THE GERIATRIC DEPRESSION SCALE..... 5

CONFIRMATORY FACTOR ANALYSIS..... 10

MEASUREMENT INVARIANCE..... 15

Configural Invariance..... 15

Metric Invariance..... 16

Scalar Invariance..... 17

Strict Invariance..... 17

Partial Measurement Invariance..... 18

ITEM RESPONSE THEORY..... 21

DIFFERENTIAL ITEM FUNCTIONING..... 28

NON-ITEM RESPONSE THEORY METHODS..... 28

Mantel-Haenszel..... 28

Logistic Regression..... 30

ITEM RESPONSE THEORY METHODS..... 32

DFIT Framework..... 32

Item Response Theory Likelihood Ratio Test..... 36

CHAPTER 3: METHODOLOGY..... 38

MEASURES..... 38

SAMPLE..... 38

ANALYSIS..... 43

Multiple-group CFA analysis procedure..... 44

Binary factorial invariance..... 47

IRTLR-DIF analysis procedure..... 55

Benjamini-Hochberg procedure..... 57

CHAPTER 4: RESULTS..... 59
INTRODUCTION..... 59
Multiple-group CFA analysis..... 62
Invariance models by gender within country..... 65
Invariance models by cross-country comparisons..... 85
Multiple-group CFA invariance summary..... 105
IRTLR-DIF analysis..... 108
IRTLR-DIF by gender..... 110
IRTLR-DIF by cross-country comparisons..... 117
IRTLR-DIF summary..... 133

CHAPTER 5: DISCUSSION..... 136
GENERAL OVERVIEW..... 136
Major Findings Multiple-Group CFA..... 137
Major Findings IRTLR-DIF..... 144
Research and Clinical Implications..... 148
Limitations and Future Directions..... 150

APPENDICES..... 295

REFERENCES..... 298

LIST OF TABLES

Table 1. Factor models 8

Table 2. Sample demographics by gender 42

Table 3. Configural invariance constraints 49

Table 4. Metric invariance constraints 50

Table 5. Scalar invariance constraints 51

Table 6. Residual invariance constraints 52

Table 7. Mean GDS-15 scores 61

Table 8. Demographic characteristics of countries 61

Table 9. One factor EFA Chile 153

Table 10. One factor EFA Cuba 154

Table 11. Two factor EFA Argentina 155

Table 12. Two factor EFA Mexico 156

Table 13. Two factor EFA Uruguay 157

Table 14. One factor CFA by country 158

Table 15. Two factor CFA by country 159

Table 16. Model fit for configural and nested models Chile by gender 160

Table 17. Invariance hypothesis tests Chile by gender 161

Table 18. Model fit for configural and nested models Cuba by gender 162

Table 19. Invariance hypothesis tests Cuba by gender 163

Table 20. Model fit for configural and nested models Argentina by gender 164

Table 21. Invariance hypothesis tests Argentina by gender 165

Table 22. Model fit for configural and nested models Mexico by gender 166

Table 23. invariance hypothesis tests Mexico by gender 167

Table 24. Model fit for configural and nested models Uruguay by gender 168

Table 25. Invariance hypothesis tests Uruguay by gender 169

Table 26. Model fit for configural and nested models Chile by Cuba 170

Table 27. Invariance hypothesis tests Chile by Cuba 171

Table 28. Model fit for configural and nested models Mexico by Uruguay 172

Table 29. Invariance hypothesis tests Mexico by Uruguay 173

Table 30. Model fit for configural and nested models Mexico by Argentina 174

Table 31. Invariance hypothesis tests Mexico by Argentina 175

Table 32. Model fit for configural and nested models Uruguay by Argentina 176

Table 33. Invariance hypothesis tests Uruguay by Argentina 177

Table 34. Item parameters and standard errors for anchor item Chile by gender 178

Table 35. Item parameters and standard errors for item Exhibiting DIF Chile by gender 179

Table 36. Item parameters and standard errors for anchor item Cuba by gender 180

Table 37. Item parameters and standard errors for item Exhibiting DIF Cuba by gender 181

Table 38. Item parameters and standard errors for anchor item Argentina by gender 182

Table 39. Item parameters and standard errors for item Exhibiting DIF Argentina by gender 183

Table 40. Item parameters and standard errors for anchor item Mexico by gender 184

Table 41. Item parameters and standard errors for item Exhibiting DIF Mexico by gender 185

Table 42. Item parameters and standard errors for anchor item Uruguay by gender 186

Table 43. Item parameters and standard errors for item Exhibiting DIF Uruguay by gender 187

Table 44.Summary of DIF analyses of the GDS-15: gender anchor item 188

Table 45.Summary of DIF analyses of the GDS-15: gender Type of DIF 189

Table 46. Summary of DIF analyses of the GDS-15: gender BH-Adjustment 190

Table 47. Item parameters and standard errors for anchor item Chile by Cuba 191

Table 48. Item parameters and standard errors for item Exhibiting DIF Chile by Cuba 192

Table 49. Item parameters and standard errors for anchor item Mexico by Uruguay 193

Table 50. Item parameters and standard errors for item Exhibiting DIF Mexico by Uruguay 194

Table 51. Item parameters and standard errors for anchor item Argentina by Mexico 195

Table 52. Item parameters and standard errors for item Exhibiting DIF Argentina by Mexico 196

Table 53. Item parameters and standard errors for anchor item Argentina by Uruguay 197

Table 54. Item parameters and standard errors for item Exhibiting DIF Argentina by Uruguay 198

Table 55. Item parameters and standard errors for anchor item Argentina by Chile 199

Table 56. Item parameters and standard errors for item Exhibiting DIF Argentina by Chile 200

Table 57. Item parameters and standard errors for anchor item Argentina by Cuba 201

Table 58. Item parameters and standard errors for item Exhibiting DIF Argentina by Cuba 202

Table 59. Item parameters and standard errors for anchor item Mexico by Chile 203

Table 60. Item parameters and standard errors for item Exhibiting DIF Mexico by Chile 204

Table 61. Item parameters and standard errors for anchor item Mexico by Cuba 205

Table 62. Item parameters and standard errors for item Exhibiting DIF Mexico by Cuba 206

Table 63. Item parameters and standard errors for anchor item Uruguay by Chile 207

Table 64. Item parameters and standard errors for item Exhibiting DIF Uruguay by Chile 208

Table 65. Item parameters and standard errors for anchor item Uruguay by Cuba 209

Table 66. Item parameters and standard errors for item Exhibiting DIF Uruguay by Cuba 210

Table 67. Summary of DIF analyses of the GDS-15: country by country anchor item 211

Table 68. Summary of DIF analyses of the GDS-15: country by country anchor item 212

Table 69. Summary of DIF analyses of the GDS-15: country by country Type of DIF 213

Table 70. Summary of DIF analyses of the GDS-15: country by country Type of DIF 214

Table 71: Summary of DIF analyses of the GDS-15: country by country BH-Adjustment 215

Table 72: Summary of DIF analyses of the GDS-15: country by country BH- Adjustment 216

Table 73 Descriptive statistics and correlations for the GDS-15 Argentina 217

Table 74 Descriptive statistics and correlations for the GDS-15 Chile 218

Table 75 Descriptive statistics and correlations for the GDS-15 Mexico 219

Table 76 Descriptive statistics and correlations for the GDS-15 Cuba 220

Table 77 Descriptive statistics and correlations for the GDS-15 Uruguay 221

Table 78 GDS-15 cutoff scores by country 222

Table 79 GDS-15 cutoff scores by gender and country 222

Table 80 Countries with the most difficulty endorsing items 223

LIST OF FIGURES

Figure 1. Chile scree-plot 224

Figure 2. Cuba scree-plot 225

Figure 3. Argentina scree-plot 226

Figure 4. Mexico scree-plot 227

Figure 5. Uruguay scree-plot 228

Figure 6. Chile by gender 229

Figure 7. Test information curve Chile by gender 230

Figure 8. Cuba by gender item 7 231

Figure 9. Cuba by gender item 12 232

Figure 10. test information curve Cuba by gender 233

Figure 11. Argentina by gender item 9 234

Figure 12. Argentina by gender item 11 235

Figure 13. Argentina by gender item 12 236

Figure 14. Argentina by gender item 15 237

Figure 15. Test information curve Argentina by gender 238

Figure 16. Mexico by gender item 6 239

Figure 17. Mexico by gender item 8 240

Figure 18. Test information curve Mexico by gender 241

Figure 19. Uruguay by gender item 5 242

Figure 20. Uruguay by gender item 6 243

Figure 21. Uruguay by gender item 12 244

Figure 22. Uruguay by gender item 14 245

Figure 23. Test information curve Uruguay by gender 246

Figure 24. Chile by Cuba item 4 247

Figure 25. Chile by Cuba item 8 248

Figure 26. Chile by Cuba test information curve 249

Figure 27. Mexico by Uruguay item 6 250

Figure 28. Mexico by Uruguay item 7 251

Figure 29. Mexico by Uruguay item 9 252

Figure 30. Mexico by Uruguay item 12 253

Figure 31. Mexico by Uruguay test information curve 254

Figure 32. Mexico by Argentina item 2 255

Figure 33. Mexico by Argentina item 7 256

Figure 34. Mexico by Argentina item 9 257

Figure 35. Mexico by Argentina item 11 258

Figure 36. Mexico by Argentina item 12 259

Figure 37. Mexico by Argentina test information curve 260

Figure 38. Argentina by Uruguay item 2 261

Figure 39. Argentina by Uruguay item 6 262

Figure 40. Argentina by Uruguay item 7 263

Figure 41. Argentina by Uruguay item 11 264

Figure 42. Argentina by Uruguay test information curve 265

Figure 43. Argentina by Chile item 2 266

Figure 44. Argentina by Chile item 7 267

Figure 45. Argentina by Chile test information curve 268

Figure 46. Argentina by Cuba item 2 269

Figure 47. Argentina by Cuba item 5 270

Figure 48. Argentina by Cuba item 7 271

Figure 49. Argentina by Cuba test information curve 272

Figure 50. Mexico by Chile item 7 273

Figure 51. Mexico by Chile item 8 274

Figure 52. Mexico by Chile item 11 275

Figure 53. Mexico by Chile item 12 276

Figure 54. Mexico by Chile item 15 277

Figure 55. Mexico by Chile test information curve 278

Figure 56. Mexico by Cuba item 2 279

Figure 57. Mexico by Cuba item 4 280

Figure 58. Mexico by Cuba item 7 281

Figure 59. Mexico by Cuba item 11 282

Figure 60. Mexico by Cuba item 12 283

Figure 61. Mexico by Cuba item 14 284

Figure 62. Mexico by Cuba test information curve 285

Figure 63. Uruguay by Chile item 2 286

Figure 64. Uruguay by Chile item 3 287

Figure 65. Uruguay by Chile item 10 288

Figure 66. Uruguay by Chile item 14 289

Figure 67. Uruguay by Chile test information curve 290

Figure 68. Uruguay by Cuba item 4 291

Figure 69. Uruguay by Cuba item 6 292

Figure 70. Uruguay by Cuba item 8 293

Figure 71. Uruguay by Cuba test information curve 294

CHAPTER 1: INTRODUCTION

INTRODUCTION

Health disparities research on the mental health of older adults faces multiple measurement challenges. Most mental health research relies on self-report measures, which represents the subjective reality of the individual. In addition, when self-report measures developed to understand psychological phenomena in one cultural group are applied to other groups, the conceptual and psychometric properties of said measures in these new environments are seldom tested. Since the mechanisms that contribute to mental disorders such as depression, are self-assessed and subjective, it is necessary to know whether the instruments being used to capture the phenomena are conceptually and psychometrically equivalent across different populations (Stalh & Hahn, 2006). Although, research by Hays, Ramirez, Stalh et al. (Hays, Morales, & Reise, 2000; Ramirez, Ford, Stewart, & Teresi, 2005; Stalh & Hahn, 2006) and others have recognized the need for addressing these issues in health measurement, a dearth of research still exists.

Health disparities research focuses on significant and persistent differences in disease rates and health outcomes between people of differing, race, ethnicity, socioeconomic position and area of residence (Eberhardt, 2004; Hartley, 2004). Constructs used in health disparities research tend to be abstract and hence not directly observable or measureable (Stewart & Napoles-Springer, 2000), and require statistical techniques that meet this structure, such as structural equation modeling and item response theory.

The accuracy or inaccuracy, with which health constructs are measured, can have an impact on study results, by producing biased estimates of symptoms and disorders, which can in turn lead to spurious conclusions, which impact health policy. The first step in addressing health disparities measurement among diverse populations would be to assess how self-reported health measures function. That is, for meaningful comparisons to be made cross-culturally, researchers must first determine whether measures developed among a majority group or in western society, perform the same way when used in non-majority groups or non-western societies.

In addition to the need for assessing how measures function between groups such as older Mexican adults versus older Mexican American adults, researchers need to be mindful of the "cultural homogeneity" pitfall. Cultural homogeneity assumes measurement equivalence based on ethnic populations sharing the same language. Ramirez (2005) states that the assumption of cultural homogeneity can actually

"*Exacerbate inaccurate cultural stereotypes and can lead to misleading conclusions in comparing prevalence of disorders, hindering the delivery of quality healthcare to different racial and ethnic groups. In addition, assuming cultural homogeneity based on shared language, is misleading because there are cultural and idiomatic nuances that can potentially exist within populations even though they share the same language" (Ramirez, et al., 2005). In short, failure to address between and within group measurement issues ultimately creates problems for researchers trying to draw research conclusions as well as end-users of said research such as health care providers (pg. 1643)."*

With a growing and diverse aged population within the United States and globally, there is a need for the validation of existing measures in order to establish cross-ethnic equivalence of health related assessment tools (Byrne & Watkins, 2003; Myers, Calantone, Page, & Taylor, 2000; Ramirez, et al., 2005).

The goal of the current study is to add to the cross-cultural and methodological literature by examining the within and between country differences in the manifestation of depression across Latin America and the Caribbean.

Present Study Rationale

The aim of this study is to assess measurement equivalence/invariance of the Geriatric Depression Scale Short Form (GDS-15) across five Spanish speaking countries in Latin America and the Caribbean .Two methodological approaches will be used (1) Study 1: multiple group confirmatory factor analysis (MGCFA) and (2) Study 2: multiple group item response theory and item response theory likelihood ratio tests (IRTLR) for the assessment of differential item functioning (DIF). These techniques will be compared to determine each methods consistency in assessing item and scale function across countries and gender.

Research Questions

Study 1 will be guided by the following questions:

1. What is the factorial structure of the GDS-15 in each of five Spanish speaking countries in Latin America?
2. Is the factor structure of the GDS-15 invariant across countries and gender?

Study 2 will be guided by the following questions:

1. How invariant are IRT-based item difficulty estimates across countries and gender?
2. How invariant are IRT-based item discrimination estimates across countries and gender?

CHAPTER 2: LITERATURE REVIEW

THE GERIATRIC DEPRESSION SCALE

The original geriatric depression scale (GDS) was developed 28 years ago by Brink et al. (1982) (Brink, et al., 1982). Prior to the development of the GDS, there were no depression screening instruments developed specifically for older adults. Previous instruments used with older adults contained items which referred to physical manifestations of depressive symptomatology. Research has shown that items referring to physical symptoms are not a good indicator of depression in older populations. For example, (Coleman, et al., 1981) found that sleep disturbances were a common symptom of depression but such disturbances were also common in older adults without depression, while rare in younger adults not suffering from depression (Yesavage, et al., 1983). For these reasons the GDS was developed and validated with older adults in the United States.

The original GDS consisted of 100 items which were tested on 46 older adults in San Francisco, CA. Thirty items were selected from the original 100, based on high item-total correlations. These items covered six qualitative domains (1) lowered affect, (2) inactivity, (3) irritability, (4) withdrawal, (5) distressing thoughts and (6) negative judgments of the past and present. Yesavage then repeated the process a year later (1983), selecting 30 items based on highest item-total correlations and identified nine qualitative domains (1) somatic complaints, (2) cognitive complaints, (3) motivation, (4) future/past

orientation, (5) self-image, (6) losses, (7) agitation, (8) obsessive traits and (9) mood itself.

From the GDS-30, the GDS-15 was developed to ease administration and lessen the time requirement for completing the instrument. In a subsequent validation study Sheikh and Yesavage (1986) (Sheikh & Yesavage, 1986) selected the 15 items from the GDS-30 which had the highest item-total correlations. Of the 15 items, 10 of them reflect the presence of depression when answered positively, while the rest indicate depression when answered negatively.

Research on the factorial structure of the GDS-30 and GDS-15 has been inconsistent across the literature (Adams, 2001; Adams, Matto, & Sanders, 2004; L. M. Brown & Schinka, 2005; P. J. Brown, Woods, & Storandt, 2007; Chau, Martin, Thompson, Chang, & Woo, 2006; Cheng & Chan, 2004; Friedman, Heisel, & Delavan, 2005; Ganguli, et al., 1999; D. W. L. Lai, Fung, & Yuen, 2005; Malakouti, Fatollahi, Mirabzadeh, Salavati, & Zandi, 2006; Parmelee & Katz, 1990; Parmelee, Lawton, & Katz, 1989; Salamero & Marcos, 1992; Schreiner, Morimoto, & Asano, 2001; Tang, Wong, Chiu, Ungvari, & Lum, 2005; Wrobel & Farrag, 2006; Yang, Small, & Haley, 2001).

For the GDS-30, Parmalee and colleagues found 6 factors (Parmelee, et al., 1989), Salamero and Marcos found 3 factors (Salamero & Marcos, 1992), Adams found 6 factors (Adams, 2001), Adams evaluated the GDS-30 again and found 5 factors (Adams, et al., 2004), the Arabic version of the GDS-30 revealed 7 factors (Chaaya, et al., 2008;

Wrobel & Farrag, 2006) while the Portuguese version had 3 factors (Pocinho, Farate, Dias, Yesavage, & Lee, 2009).

Studies that evaluated the factor structure of the GDS-15 include Mitchell et al. (1993) in which they found 3 factors (Mitchell, Matthews, & Yesavage, 1993), Brown et al. found 2 factors (P. J. Brown, et al., 2007), the Chinese version had 4 factors (Chau, et al., 2006; D. Lai, Tong, Zeng, & Xu, 2010; D. W. L. Lai, et al., 2005) and finally the Iranian version had 2 factors (Malakouti, et al., 2006). In addition to the inconsistent factorial structures, the definitions of these factors were also wide spread, from general depression and dsyphoria to lack of vigor and agitation. Across studies the factors referenced most often tended to be positive and negative affect, energy loss and life satisfaction.

Table 1 Factor models of the GDS-15

Study		Number of Factors	Factor Definitions
Sheikh et al.	(1986)	1	General Depressive Affect
Mitchell et al.	(1993)	3	General Depressive Affect, Life Satisfaction, Withdrawal
Schreiner et al.	(2001)	2	Positive Affect, Energy Loss and Depressed Mood
Incalzi et al.	(2003)	3	Positive Attitude Toward Life, Distressing Thoughts and Negative Judgement, Inactivity and Reduced Self-Esteem
Friedman et al.	(2005)	2	General Depressive Affect, Positive Affect
Lai et al.	(2005)	4	Negative Mood, Positive Mood, Inferiority and Disinterest, Uncertainty
Brown et al.	(2007)	2	General Depressive Affect, Life Satisfaction
Oinishi et al.	(2007)	4	Unhappiness, Apathy and Anxiety, Loss of Hope and Morale, Energy Loss

To date the instrument has been translated into 24 languages, with less than 10 of the studies that used the instrument, actually evaluating the factor structure. The primary forms of psychometric evaluation of the GDS-15 have been (1) exploratory factor analyses, (2) test-retest reliability and (3) sensitivity and specificity analyses. Only Brown et al. (2007) conducted multiple group confirmatory factor analysis.

The aims of the present study are to (1) assess the factorial structure within each country, (2) conduct multiple group confirmatory factor analyses within each country by gender and (3) conduct multiple group confirmatory factor analyses between countries.

CONFIRMATORY FACTOR ANALYSIS

Confirmatory factor analysis (CFA) is a form of the factor analytical model which examines the covariation among manifest indicators in order to confirm the hypothesized underlying latent constructs or common-factor. CFA is a theory driven technique in which the researcher specifies (1) the number of factors and their inter-correlation, (2) which items load on which factor and (3) whether errors are correlated. Statistical tests can then be conducted to determine whether the data confirm the theoretical model, thus the model is thought of as confirmatory (Bollen, 1989). With CFA a researcher is able to simultaneously conduct multiple group analyses across time or samples, in order to evaluate measurement invariance/equivalence.

The following is a mathematical presentation of linear CFA for testing measurement invariance (Baumgartner & SteenKamp, 2001; Bollen, 1989; Jöreskog, 1971). In the CFA model, the observed response χ_i to an item i $(i = 1,\ldots p)$ is represented as a linear function of a latent construct ξ_j $(j = 1,..m)$, an intercept τ_i, and stochastic error term δ_i. Thus,

$$x_i = \tau_i + \lambda_{ij}\xi_j + \delta \quad \text{Equation 1}$$

Where τ_{ij} is the slope of the regression of x_i on ξ_j, the slope or factor loading, defines the metric of measurement, as it shows the amount of change in x_i due to a unit change in ξ_j. The intercept τ_i, in contrast, indicates the expected value of x_i when $\xi_j = 0$ (Sörbom, 1974).

Assuming p items and m latent variables, and specifying the same factor structure for each country g $(g = 1,...G)$ we get the following measurement model

$$x^g = \tau^g + \Lambda^g \xi^g + \delta^g$$ **Equation 2**

where x^g is a $p \times 1$ vector of observed variables in country g, ξ^g is a $m \times 1$ vector of latent variables, δ^g is a $p \times 1$ vector of errors of measurement, τ^g is a $p \times 1$ vector of item intercepts and Λ^g is a $p \times m$ matrix of factor loadings.

It is assumed that $\mathrm{E}\left(\delta^g\right) = 0$ and that $COV\left(\xi^g, \delta^g = 0\right)$. Equation (2) shows that observed scores on p items are a function of underlying factor scores, but

that observed scores may not be comparable across countries because of different intercepts $\left(\tau_i^g\right)$ and scale metrics λ_{ij}^g.

To identify the model, the latent constructs have to be assigned a scale in which they are measured. In multiple group analyses this is done by setting the factor loading of one item per factor to 1. Items for which loadings are fixed at unity are referred to as marker (or reference) items. The same items should be used as marker item(s) in each country.

Taking the expectations of equation (2) yields the following relationship between the observed item means and the latent means

$$\mu^g = \tau^g + \Lambda^g \mathrm{K}^g$$

Equation 3

where μ^g is the $p \times 1$ vector of item means and K^g is the $m \times 1$ vector of latent means (i.e. the means of ξ^g). The parameters K^g and τ^g cannot be identified simultaneously (Sörbom, 1982). In other words, there is no definite origin for the latent variables. To deal with this indeterminacy, constraints are placed on the parameters. There are two approaches to placing constraints. The first is to fix the intercept of each latent variable's marker item to zero in each country. This equates the means of the

latent variables to the means of their marker variables (i.e. $\mu_m^g = \kappa_m^g$, where m indicates that the item is a marker item).

A second approach is to fix the vector of latent means at zero in the reference country (i.e. $\kappa^r = 0$, where the superscript r indicates the reference country) and to constrain one intercept per factor to be invariant across countries. The latent means in the other countries are then estimated relative to the latent means in the reference country. These two approaches lead to an exactly identical model with respect to the item intercepts and latent construct means. If further restrictions are imposed on the model (e.g., all intercepts are specified to be invariant across countries), the intercepts and latent means are over-identified, and the fit of the means part of the model can be investigated.

In addition to the mean structure given by equation (3), the covariance structure has to be specified. As usual, the variance-covariance matrix of x in country g, Σ^g is given by:

Equation 4

$$\Sigma^g = \Lambda^g \Phi^g \Lambda'^g + \Theta^g$$

where Φ^g is the variance-covariance matrix of the latent variables in ξ^g and Θ^g is the variance-covariance matrix of δ^g which is usually constrained to be a diagonal matrix. The overall fit of the model is based on the discrepancy between the

observed variance-covariance matrices S^g and the implied variance-covariance matrices $\hat{\Sigma}^g$ and the discrepancy between the observed vectors of the means m^g and the implied vectors of $\hat{\mu}^g$.

.

MEASUREMENT INVARIANCE

Measurement invariance is an umbrella term that is really made up of various forms of invariance. For example, measurement invariance can refer to the invariance of factor loadings, intercepts, or errors (Meredith, 1993). These forms or levels of invariance are referred to as configural, metric/weak, scalar/strong and residual error/strict. Each of these levels is a successively more restrictive test of measurement invariance.

Configural invariance

Configural invariance (J.L. Horn, McArdle, & Mason, 1983), weak invariance (Meredith, 1993) or pattern invariance (R. E. Millsap, 1997) is the lowest or weakest level of measurement invariance that can be obtained. Configural invariance refers to the pattern of salient (non-zero) and non-salient (zero or near zero) loadings which define the structure of a measurement instrument. Configural invariance is supported if the specified model with zero-loadings on non-target factors fits the data well in all groups; all salient factor loadings are significantly below unity. Simply stated, configural invariance holds when the same items load on the same factors for both groups of interest (e.g. men vs. women). The configural model is also used as the baseline model to which the decrements in fit associated with more constrained nested models are compared.

Metric invariance

Metric (Thurstone, 1947), weak (Meredith, 1993), or factor pattern invariance (R. E. Millsap, 1995) is more restrictive than configural invariance. This level of invariance requires that the loadings in a CFA be constrained to be equivalent in each group while permitting the factor variances and covariances to vary across groups.

$$\Lambda^1 = \Lambda^2 = \ldots \Lambda^G \quad \text{Equation 5}$$

If the factor loadings are found to be invariant this means that the factor loadings in one group are proportionally equivalent to corresponding loadings in other groups (Bontempo, et al., 2008).

"Loadings standardized to the common-factor variance would each differ from the corresponding loading in another group by the same proportion—the ratio of the variance in each group. It is essential that the common-factor variances are freely estimated in all but the first group. This condition is what creates a test of proportionality when equality constraints are imposed on the loadings (pg.51)" (Bontempo, 2007). If metric invariance holds it allows a researcher to claim that there are similar interpretations of the factors across groups. However, others have suggested that higher levels of invariance provide greater evidence of the equivalent construct interpretation across groups (Meredith, 1993).

Scalar invariance

Scalar (SteenKamp & Baumgartner, 1998) or Strong (Meredith, 1993) invariance is more restrictive than metric/weak invariance because it constrains factor loadings as well as intercepts to be equal across groups.

Equation 6

$$\tau^1 = \tau^2 = \ldots \tau^G$$

"This requires the model to account for all mean differences in the items solely through the common-factor mean (pg. 52)" (Bontempo, 2007). If scalar invariance is obtained then comparison of factor means across groups is supported.

Strict invariance

Strict invariance is the most restrictive constraint; it requires that equality constraints for loadings, intercepts and unique-nesses (errors) be held equal across groups.

Equation 7

$$\Theta^1 = \Theta^2 = \ldots \Theta^G$$

This level of measurement requires that the specific and random error components of each item be equivalent across groups, such that differences in variance across groups can only take place at the latent variable level. If strict invariance is obtained, a researcher can be confident in making measurement comparisons based on factor mean and factor covariance structures across groups. There is however, a lack of consensus in

the literature as it relates to the necessity of acquiring strict invariance. Researchers such as Byrne and Vandenberg & Lance, refer to strict invariance as being too restrictive and not of import (Byrne, 2008) (R.J. Vandenberg & C.E. Lance, 2000).

Partial measurement invariance

The aforementioned measurement invariance tests build upon one another and with each level of invariance obtained a researcher can build support for the equivalence of a measure across groups and time. Initially researchers testing the invariance of an instrument had two options (1) obtain full measurement invariance or (2) if full measurement invariance is not obtained, abandon further invariance testing. Byrne and colleagues (Byrne, Shavelson, & Muthen, 1989) presented the idea that there could be a middle ground between full measurement invariance and no measurement invariance. That middle ground is partial measurement invariance.

Partial measurement invariance is the idea that some invariance is better than no-invariance. What this means is that if, for example, a researcher obtains configural invariance and then proceeds to test metric invariance and finds that they have a decrement in fit, future invariance analyses do not have to be abandoned. Instead they can investigate the source of the misfit and then relax the constraints for specific parameters that are exhibiting misfit. The relaxation of constraints then allows the researcher to recalibrate the model for fit and move on to another level of invariance testing, with the understanding that further analyses are based on the partial invariance of that level (e.g. partial metric invariance leads to partial scalar invariance and so on).

The implementation of partial measurement invariance allows analyses to proceed but it also introduces two additional issues (1) how much measurement invariance or lack thereof is acceptable and (2) how does one identify misfit? Byrne and colleagues address both of these issues in the context of metric invariance. Byrne et al. (Byrne, et al., 1989) state that the measurement invariance literature leaves researchers with the impression that "given a non-invariant pattern of factor loadings (metric invariance), further testing of invariance and the testing for differences in factor mean scores is unwarranted"(pg. 458).

This idea is unfounded when the model specification includes multiple indicators of a construct and at least one item (other than the one that is fixed to 1.00 for identification purposes) is invariant (Byrne, et al., 1989; Muthen & Christoffersson, 1981). Byrne and colleagues provide evidence that partial metric invariance only requires cross-group invariance of zero loadings and some, but not necessarily all of the salient loadings (Byrne, et al., 1989).

The second issue with partial measurement invariance focuses on how to decide which constraints need to be relaxed. Generally a researcher relies on substantive reasons when deciding which loadings or intercept constraints to relax across-groups. This information is not always available which means that the researcher must depend on modification indices when respecifying their model. Structural equation modeling software packages such as Mplus provide modification indices that identify parameters with the poorest fit. The values of the modification indices provide information on the

estimated decrease in the χ^2 value (misfit) that would occur if the constrained parameter in question was relaxed (freely estimated).

Caution must be taken when using modification indices in order to avoid over fitting of the model which would impair generalizability (Tomarken & Waller, 2003). Steenkamp et al. (SteenKamp & Baumgartner, 1998) state that "invariance constraints should be relaxed only when modification indices are highly significant (both in absolute magnitude and in comparison with the majority of other modification indices)" (pg.81). Changes in alternative indices of overall model fit (CFI, TLI, RMSEA) should be evaluated especially those that take model parsimony into account (Steiger, 1990). As a rule the use of modification indices should be kept to a minimum, by only implementing modifications that would correct severe problems with model fit, this would in turn minimize capitalization on chance and once again maximize the cross-validity of the model (MacCallum, Roznowski, & Necowitz, 1992).

The reason for proceeding with partial measurement invariance analyses can be summed up by Horn (1991, p. 125) (J.L. Horn, 1991; J.L. Horn, et al., 1983)

"*metric invariance is a reasonable ideal…a condition to be striven for, not one expected to be fully realized…and scientifically unrealistic"(pg.125).*

The use of partial measurement invariance allows researchers to deal with the realities of working with latent constructs measured by self-reported items all of which have some level of inherent bias, which ultimately influences the equivalence of measures.

ITEM RESPONSE THEORY

Item response theory (IRT) is less of a theory and more of a collection of mathematical models, mostly non-linear latent variable/trait models which attempt to explain how people respond to items. IRT models present a picture of the performance of each item on an instrument and how the instrument measures the construct of interest in the population being studied. Basic IRT definitions, models and assumptions are presented below:

Basic IRT definitions

Latent variable

In classical test theory the "latent variable/trait" is represented by the "true score" in structural equation modeling it is referred to as the "latent factor" in IRT the latent trait is represented by theta θ; theta is the unobservable construct being measured (e.g., depression).

Item threshold

The item threshold (b) (item location, item difficulty, item severity) provides information on the location of an item along the θ continuum indicating the level of the underlying variable (e.g. depression) needed to endorse an item (e.g., do you feel like your life is empty?) with a specified probability, typically set at .50 (Reise, 2005).

Item discrimination

Item discrimination (a, α item slope) describes the strength of an items ability to discriminate among people with trait θ levels below and above the item threshold- b. The a- parameter can also be interpreted as how related an item is to the trait measured by the instrument (Reise, 2005).

Item characteristic curve (ICC)

The item characteristic curve models the relationship between a person's probability for endorsing an item category (e.g., yes or no) and the level on the construct θ measured by the instrument.

Information function

The information function for items (IIF) or scales (SIF) is an index which indicates the range of trait level θ over which an item or scale is most useful for distinguishing among individuals. "For any item, the IIF or SIF characterizes the precision of measurement for persons at different levels of θ, with higher information denoting better precision (lower standard error)(pg.427) (Reise, 2005)".

Item response theory models for dichotomous data

The three commonly used IRT models are the 1(1PL), 2 (2-PL) and the 3-parameter logistic model (3-PL).

The 1-parameter model is specified as follows:

$$P(\Theta) = \frac{1}{1 + \exp\left[-1.7a(\Theta - b)\right]} \quad \text{Equation 8}$$

where -1.7 is a scaling factor and (a) is the slope/discrimination parameter which is constant for all items (and is often scaled to equal 1). Having "a" scaled to 1 implies that all items on the scale have equal discrimination, which is an assumption of the 1 parameter model.

Items can however, discriminate at different places along the θ continuum; with items that are easy to endorse discriminating among people who are low on the construct of interest (e.g. depression) and items that are hard to endorse discriminating among people who are high on the construct of interest. Theta θ is the unobservable construct being measured and (b) is the item difficulty/severity which provides information on the location of an item along the θ continuum indicating the level of the underlying variable needed to endorse an item.

Although, the 1 parameter-model is used in many settings, the strict assumption of equal discrimination rarely holds for health measures; instead the 2 parameter model is most commonly applied to dichotomous health data.

The 2-parameter model is specified as:

$$P_i(\Theta) = \frac{1}{1 + \exp\left[-1.7a_i\left(\Theta_j - b_i\right)\right]} \quad \text{Equation 9}$$

Where $P_i(\Theta)$ the probability of endorsing item (i), a_i is the slope/discrimination parameter for item (i) which indicates how related a particular item is to the construct being measured. The higher the slope, the more variability in item responses can be attributed to differences in the latent construct (Edwards, 2009). The discrimination parameter is analogous to a factor loading. Parameter b_i is the threshold or severity parameter for item (i). The threshold parameter indicates the point along the latent continuum where an individual would have a 50% chance of endorsing an item.

The higher the threshold, the higher an individual must be on the latent trait to have a 50% chance of endorsing that item. With the 1- parameter model the only parameter allowed to vary was the threshold (b), in the 2-parameter model both the discrimination (a) and threshold (b) are allowed to vary. The last model presented will be the 3 parameter model, which is primarily used in educational testing.

The 3-parameter model is specified as:

$$P(\Theta) = c + \frac{(1-c)}{1+\exp\left[-1.7a(\Theta - b)\right]} \quad \text{Equation 10}$$

where “a”, “b”, Θ are defined as in the 1 and 2 parameter models. The 3-parameter model introduces the “c” parameter or guessing parameter, which is a lower asymptote parameter. The 3-PL simultaneously, estimates the discrimination, difficulty/severity and lower asymptote parameters. With multiple-choice items, questions can be solved by “guessing” and because of this the probability of success is greater than zero, even for persons with lower trait/ability levels.

Finally, it should be noted that the (b) parameter, is interpreted differently in the 3-PL model. Item difficulty still occurs at the point of inflection, however, the probability of endorsement is no longer at .50, but rather the inflection point is shifted by the lower asymptote (Embretson, 2000). This type of model is primarily used in educational testing, but is not applicable to psychological or health measurement instruments. The reason the use of the 3-PL model is not used in health measurement is because it has a guessing parameter which is difficult to interpret in the context of health items.

Assumptions

There are four assumptions that IRT models make (1) monotonicity, (2) unidimensionality, (3) local independence and (4) invariance. *Monotonicity* implies that as levels of the latent construct increase (absence of depression to severe depression), individuals have a higher probability of endorsing item response categories indicating poorer mental health (higher depression). This assumption is evaluated by examining graphs of summed scale scores compared to item endorsement rates or item means.

The *unidimensionality* assumption means that there is only one common latent variable being measured. This means that no other variable except a person's level on "depression" accounts for the variation in responses to the 15 items on the GDS-15 scale. Currently there is no gold standard for assessing whether data is unidimensional for IRT application (Reise, 2005). With that said, the most commonly used approaches for evaluating unidimensionality is to use a combination of exploratory (EFA) and confirmatory (CFA) factor analyses.

With exploratory factor analysis a researcher is looking for a large ratio between the first and second Eigen values, which would indicate one primary dimension, with all items loading highly on a single common factor. In addition, after extracting one factor, residuals should be small which would indicate that one dimension accounts for a high percentage of item covariance. The measures of fit for a CFA (CFI, TLI, and RMSEA) provide sufficient evidence for unidimensionality.

The assumption of local independence means that once one common factor has been extracted from an item covariance matrix, the residuals are zero; or that after

accounting for the latent variable, item responses are independent of one another (Reise, 2005).

The fourth assumption of invariance follows the following tenants:

1. An individual's position on a latent-trait continuum can be estimated from their responses to any set of items with known item characteristic curves.
2. Item properties do not depend on the characteristics of a particular population.
3. The scale of the trait does not depend on an item set, but instead exits independently (Reise, Ainsworth, & Haviland, 2005).

DIFFERENTIAL ITEM FUNCTIONING

The purpose of this section is to provide a general overview of various differential item functioning (DIF) detection methods. DIF detection methods involve three elements: (1) item response which may be treated as observed or latent, (2) an estimate of ability (depression) level and (3) subgroup membership such as gender, country of origin, or race/ethnicity. The overarching question in a DIF analysis is how a person's item response is related to their level of ability (depression) based on their subgroup membership. To investigate this question, DIF analyses focus on differences in item parameters. In other words, a DIF analysis is concerned with whether or not the likelihood of item or category endorsement is the same across subgroups. DIF methods fall into two categories: (1) IRT based methods and (2) non-IRT based methods.

Non-IRT DIF Methods

Mantel-Haenszel

The Mantel-Haenszel (MH) procedure proposed by Holland and Thayer (Holland & Thayer, 1988) is one of the most widely used nonparametric methods for detecting differential item functioning (DIF). The MH procedure is based on the analysis of contingency tables. Subjects are matched on an observed variable (e.g., total score on the GDS-15), and then counts of subjects in the focal and reference groups endorsing or not endorsing the item are compared. The reference group consists of individuals for whom the test/instrument is expected to favor and the focal group consists of individuals who were at risk of being disadvantaged by the test/instrument. The MH procedure can be

used to examine whether within each depression score grouping, the odds of a symptom (endorsement of a particular item) is the same across groups.

A common odds ratio (which tests whether or not the likelihood of item symptom response is the same across the depression groupings) also can be used to construct a DIF magnitude measure (J. A. Teresi, Ramirez, Lai, & Silver, 2008). This is accomplished by converting the odds to log odds and applying transformations, which in turn provide interpretable measures of magnitude (J. A. Teresi, et al., 2008). The MH common odds ratio assesses the strength of association in three-way 2x2xJ contingency tables. It estimates how stable the association between two factors is in a series of "J" partial tables. This procedure tests the following null hypothesis

$$H_0 : \frac{P_{Ri}}{q_{Ri}} = \frac{P_{Fi}}{q_{Fi}} \qquad \text{Equation 11}$$

which reflects the odds of endorsing item i for the reference group R, $\frac{P_{Ri}}{q_{Ri}}$, are equal to the corresponding odds for the focal group F, $\frac{P_{Fi}}{q_{Fi}}$.

Note that $q*i = 1 - P*i$.

The alternative hypothesis is

$$H_1 : \frac{P_{Ri}}{q_{Ri}} = \alpha_i \frac{P_{Fi}}{q_{Fi}} \quad \text{Equation 12}$$

where $\alpha_i = \frac{P_{Ri} q_{Fi}}{P_{Fi} q_{Ri}}$ is the common odds ratio $\left(\alpha_{i \neq 1} \right)$. This procedure provides a χ^2 test statistic as well as an estimator of α_i across the 2x2xJ tables. The latter is a measure of effect size, how much the data depart from the null hypothesis.

Logistic regression

Another DIF detection procedure similar to the Mantel-Haenszel method is logistic regression (LR). The LR procedure predicts the probability of endorsement of an item based on

$$P(U = 1 | \theta) = \frac{e\left(\beta_0 + \beta_1 \theta \right)}{\left[1 + e\left(\beta_0 + \beta_1 \theta \right) \right]} \quad \text{Equation 13}$$

where U is the response to the item, θ is the observed ability or symptomatology (depression), β_0 is the intercept and β_1 is the slope parameter. This formula is the

standard LR model for predicting a dichotomous dependent variable from given independent indicators (Steele, et al., 2006; Swaminathan & Rogers, 1990).

This model can then be extended to evaluate differences between groups such that

Equation 14

$$y = \beta_0 + \beta_1 x_1 (totalscore) + \beta_2 x_2 (grp) + \beta_3 \left(x_1 x_2 \right)$$

where y is the item response, β_0 is the intercept, x_1 is the total score, x_2 is the group membership, β_1 is the coefficient for the total score, β_2 is the coefficient for group membership and β_3 is the interaction between the total score and group membership.

DIF is said to be present if persons with the same level of depression but from different groups do not have the same probability of endorsing an item. There are two types of DIF: uniform and non-uniform. Looking at equation (14) DIF is not present if the LR curves for the two groups are the same, that is $\beta_{01} = \beta_{02}$, $\beta_{11} x_{11} = \beta_{12} x_{12}$. If $\beta_{11} x_{11} = \beta_{12} x_{12}$ but $\beta_{01} \neq \beta_{02}$ the curves will be parallel indicating uniform DIF. Uniform DIF is present when the probability of endorsing an item is greater for one group over the other, uniformly across all levels of the construct of interest.

Conversely if $\beta_{01} = \beta_{02}$ but $\beta_{11}x_{11} \neq \beta_{12}x_{12}$ the curves are not parallel, indicating non-uniform DIF. Non-uniform DIF is present when there is an interaction between the total score and group membership, which means that the difference in the probabilities of item endorsement for the two groups are not the same at all levels of the construct of interest.

ITEM RESPONSE THEORY METHODS

DFIT framework

The DFIT method developed by Raju and colleagues (Flowers, Oshima TC, & Raju, 1999; Raju, van der Linden WJ, & Fleer, 1995) is based on IRT. Raju's DFIT framework has several characteristics: (1) it can be used with both dichotomous and polytomous items, (2) it can handle both unidimensional and multidimensional IRT models, (3) it evaluates DIF at both the item and scale level and (4) it provides two types of DIF: compensatory DIF (CDIF) and non-compensatory DIF (NCDIF).

Within the DFIT framework, DIF is defined as the difference between the probabilities of a positive item response for individuals from different groups at the same level of the underlying attribute (Jeanne A. Teresi, 2006) . This framework uses a weighted difference in the conditional item probabilities which is the probability of people endorsing the item at each score for the reference group (women) and the focal group (men). Before the DFIT framework can be applied, however, separate IRT item parameter estimates for a reference group (women) and a focal group (men) must be obtained. Because the two sets of item parameters are obtained from separately estimated

IRT models, they must be placed on a common metric before comparisons to evaluate DIF can be made. This process is referred to as linking or equating. Once linked, the two sets of item parameters can be used to make item and scale level assessments of DIF using the DFIT framework.

Non-compensatory DIF (NCDIF)

To examine the magnitude of DIF Raju et al. (Raju, et al., 1995) and later Flowers et al. (Flowers, et al., 1999) developed the compensatory DIF index (CDIF) as well as the non-compensatory DIF index (NCDIF). According to Raju and colleagues (Oshima TC & Morris, 2008; Raju, et al., 1995), the NCDIF is defined as the average squared distance between the item characteristic functions for the focal and reference groups. For dichotomous IRT models, the gap is defined as the difference in the probability of a correct response,

$$d_i(\theta_S) = P_{iF}(\theta_S) - P_{iR}(\theta_S) \quad \text{Equation 15}$$

.

NCDIF is defined as the expected value of the square distance,

NCDIF Equation 16

$$_i = E_F\left[d_i(\theta_S)^2\right]$$

where E_F denotes the expectation taken over the θ distribution from the focal group. Squaring (d) is important so that differences in opposite directions will not cancel each other out. This allows NCDIF to assess both uniform and non-uniform DIF. When item characteristic functions (ICF's) differ based on the "b" parameter, this is referred to as uniform DIF. Uniform DIF is present when the probability of endorsing an item or getting an item correct is greater for one group over the other, uniformly across all levels of the construct of interest.

Non-uniform DIF occurs when the "a" parameters differ across groups. In this case non-uniform DIF is present when there is an interaction between the ability θ and group membership, which means that the difference in the probabilities of item endorsement or a correct response for the two groups are not the same at all levels of the construct of interest. When DIF is non-uniform, differences in both directions will contribute to NCDIF.

The DFIT framework is able to assess DIF at both the item and test level (DTF) differential test functioning. The DTF is similar to the NCDIF except the curves being compared are test characteristic functions instead of item characteristic functions.

$$D(\theta_S) = T_F(\theta_S) - T_R(\theta_S) \quad \text{Equation 17}$$

DTF is defined as the expected value of the squared difference between focal and reference groups, where the expectation is taken across the θ distribution from the focal group,

$$\text{DTF} = E_F\left[D(\theta_S)^2\right] \quad \text{Equation 18}$$

Despite the similarity in how NCDIF and DTF are defined, the relationship between the two is not straightforward. NCDIF assumes that all items other than the studied item are DIF free, while DTF depends not only on the level of DIF on each item, but also the pattern of DIF across items (Oshima TC & Morris, 2008). As a result, if you removed an item with a large NCDIF it would not necessarily result in a large decrease in DTF. However, the compensatory DIF index does a better job at reflecting an item's contribution to DTF.

Compensatory DIF (CDIF)

Compensatory DIF (CDIF) takes the item covariances into account. CDIF is defined as Equation 19

$$CDIF_l = \mathrm{E}_l = \left(d_{i,}D\right) = Cov\left(d_{i,}D\right) + \mu_{di}\mu_D$$

where COV stands for covariance. The CDIF index is additive such that:

$$\text{DTF} = \sum_{i=1}^{n} CDIF_i \qquad \text{Equation 20}$$

The additive nature of CDIF makes it possible for a researcher to investigate the net effect of removing particular items, on DTF.

Item response theory likelihood ratio tests

The item response theory log-likelihood ratio test (IRTLR) is another IRT-based approach to DIF detection procedure. The IRTLR involves the statistical comparison of two hierarchically nested item response models, (1) a constrained or compact model and (2) an unconstrained or augmented model. The unconstrained model contains all of the parameters of the constrained model, hence the constrained model is said to be hierarchically nested within the unconstrained model.

As explained by Thissen et al. (1993):

"*the goal of the procedure is to test whether the additional parameters in the unconstrained model is significantly different from zero….The IRTLR test takes the form of*

$$G^2(df) = 2\log\left[\frac{Likelihood[unconstrained]}{Likelihood[constrained]}\right] \qquad \text{Equation 21}$$

where Likelihood[.] represents likelihood of the data given the maximum likelihood estimates of the parameters of the model, and df is the difference between the number of parameters in the unconstrained model and the number of parameters in the constrained model (pg. 73)".

The value of G^2 is assumed to be distributed as $\chi^2(df)$ under the null hypothesis. Thus, if the value of $G^2(df)$ is large, representing an unlikely value from a $\chi^2(df)$ distribution, we reject the null hypothesis and the constrained model (Thissen, Steinberg, & Wainer, 1993).

The IRTLR method employs the Marginal Maximum Likelihood estimation algorithm developed by Bock & Aitkin (Bock & Aitkin, 1981), which makes it possible for the item parameters to be reliably estimated without using information about the ability distribution and uses likelihood ratio tests to evaluate the statistical differences of models between groups (Thissen, et al., 1993).

CHAPTER 3: METHODOLOGY

MEASURES

Geriatric Depression Scale-Short Form (GDS-15): A 15-item short scale version of the 30 item GDS, it consists of 10 items which indicate the presence of depression when answered positively and 5 items that indicate depression when answered negatively. Each item has a yes or no response format (1 = yes, 0 = no) and five items are reverse scored. A score on the GDS-15 which is less than 6 indicates no depression and scores greater than 6 indicate the presence of depressive symptoms.

SAMPLE

The sample for this study is made up of 7,573 older adults between the ages of 60 and 102. Data for the present study comes from the Survey on Health, Well-Being and Aging in Latin America and the Caribbean (SABE)(Pelaez, et al., 2004). The SABE study collected data during 1999 and 2000 with the primary purpose of examining health conditions and functional limitations of persons 60 and older in the countries of Argentina, Chile, Cuba, Mexico and Uruguay. The study was conducted in the official language of each country, which is Spanish.

The sample came from a population over 60 years of age that resided in private households in each of the urban areas in the respective countries, Argentina (Buenos Aires), Cuba (Havana), Mexico (Mexico City), Chile (Santiago) and Uruguay (Montivedo). The sampling framework for the SABE came from national employment surveys, household surveys, national census and national electoral registries in the

respective countries. Data was collected through face to face interviews. All countries in the SABE adhered to the same data collection protocol whereby subjects were only interviewed if they demonstrated that they were cognitively sound.

The universe of study was a population aged 60 and older who resided in private households occupied by permanent residents in urban areas of each country. A multistage clustered sample with stratification was employed in a three stage process. The plan for sampling in each country was the following: first the primary sampling unit (PSU) was established; the PSU is a cluster of independent households within predetermined geographic areas. PSU's were grouped into either geographic or socioeconomic strata. The sample distribution by geographic or socioeconomic strata was proportional to the size of the elderly population within each country.

Secondly, the PSU's were then divided into secondary sampling units, (SSU) each containing a smaller number of independent households. These SSU's were comprised of tertiary sampling units (TSU) formed by interviewees in the selected households or by single individuals in those countries where only one person was selected out of each household. As such, the household or target individuals constitute the last layer of aggregation in the sample.

The first stage in the sampling process led to sampling a predetermined number of PSU's which were each selected with probability proportional to the household distribution within each stratum. The second stage of sampling led to the selection of SSU's and the third stage of sampling consisted of the selection of households within each SSU. Finally, both secondary sampling units (SSU) and tertiary sampling units

(TSU) were selected with equal probabilities within each chosen primary sampling unit (PSU).

There were also some country specific sampling design differences. The first is the stratification of the clusters within each country. In some countries stratification was conducted in terms of geography only while in others the strata were defined by geography and aggregate indicators of socioeconomic conditions. The second area of difference was oversampling. In three countries (Cuba, Uruguay and Chile) the samples included oversamples for people age 80 and above.

In Cuba and Uruguay an individual in a household who was 80 or above was chosen with a probability of one. In Chile selection of a person among eligible household members was done randomly but if an individual aged 80 or above was present and not chosen by the random process, he/she was also interviewed. The third area of difference was with the secondary sampling unit. In three countries (Cuba, Argentina and Uruguay) only one target individual was selected per household. In Mexico all eligible individuals found in the household were interviewed.

The final sample observations broke down in the following way: Argentina (N = 1043), Chile (N = 1301), Cuba (N = 1905), Mexico (N = 1876) and Uruguay (N = 1450). These samples are proportional to the size of the elderly population within each country.

The SABE (across all countries) was made up of 54.7 % female and 45.3% male participants. Forty-six percent of the female sample was married while 44% were widows and 10% never married. At the time of the study 71% of males surveyed were married, 21% were widowers and 8% had never married. The men in the sample tended to be younger, with 44% between the ages of 50 and 65 while 27% of women fell within this range.

Women made up a larger portion of participants 66-85 years of age (66%) while men made up 52%. In subjects age 86-102, women represented 7% while men made up 4% of this grouping. The larger number of females at the upper end of the age grouping reflects earlier mortality among men. The average age across countries was 70. In the overall sample, 95% of participants had 12 years of education or less while 5% had more than 12 years of education. The average years of education across all countries are 7, Table 2 summarizes these statistics.

Table 2 Sample demographics by gender

	Total		Women		Men	
Variable	*n (%)*	*M (SD)*	*n (%)*	*M (SD)*	*n (%)*	*M (SD)*
Age		70 (9.01)		72 (8.26)		68 (9.31)
Unmarried	3308 (43.7%)		2301 (55.6%)		1006 (29.3%)	
Married	4267 (56.3%)		1838 (44.4%)		2428 (70.7%)	
No formal education	883 (11.8%)		451 (11.1%)		431 (12.8%)	
Elementary – middle school	4621 (61.0%)		2640 (63.8%)		1980 (57.7%)	
H.S.	1304 (17.5%)		659 (16.2%)		645 (19.1%)	
> H.S.	645 (8.7%)		328 (7.9%)		317 (9.4%)	

ANALYSIS

Prior to invariance testing exploratory factor analyses (EFA) will be run for each country individually. Because of the lack of consensus on the factorial structure of the GDS-15 and the absence of psychometric work on the instrument in Chile, Argentina, Cuba, Uruguay and Mexico, it is necessary to conduct EFA's to determine factor structures.

In order to establish measurement invariance a nested sequence of increasingly restrictive CFA models (invariance hypotheses) are tested. These levels of invariance are referred to as configural, metric, scalar and strict invariance

The sequence of nested invariance hypotheses are well established in the literature (Cheung & Rensvold, 2002; Reise, Widaman, & Pugh, 1993; Robert J. Vandenberg & Charles E. Lance, 2000). They are based on establishing a baseline model and additively testing hypotheses of metric, scalar, and strict invariance.

Multiple-Group CFA Analysis Procedures

Configural invariance

Configural invariance requires that the same number of factors and pattern of salient factor loadings be equivalent across groups. The baseline model tests the hypothesis of zero-loadings needed to specify a degree of simple structure. The principle of simple structure states that items comprising an instrument should exhibit the same configuration of salient and non-salient factor loadings across groups being compared (Beckstead, Yang, & Lengacher, 2008). The configural (baseline) model can be identified by giving the factor means and variances a scale for each group (men vs. women). This is done by fixing the mean of the factor(s) to zero and fixing a single factor loading to one. Note that there is more than one way to identify the baseline model.

A researcher may choose to fix the mean and variance of a factor, or fix the intercept and loading of a reference item. Either approach to identifying the baseline model will have the same degrees of freedom and model fit, only the scaling of the parameters will differ (Bontempo, 2007; Reise, et al., 1993). Configural invariance is used as a baseline model, to which the decrements in fit associated with more constrained nested models are compared against. In other words, the configural model is the model with freely estimated parameters against which subsequent nested models with constrained parameters will be compared against.

Summary of configural model constraints:

1. The same indicators are specified in each group.
2. The variance of the factor(s) is fixed to one in the 1st group.
3. The mean of the factor(s) is fixed to zero in the 1st group.
4. The intercept and loading of a reference item is constrained to be equal across groups.
5. All other non-fixed parameters are freely estimated

If configural invariance is obtained the next step is a test of metric invariance.

Metric invariance

Metric invariance implies that items are measured according to the same scale units, in that the factor loadings are equivalent across groups. When an items factor-loading is non-invariant, the regression slope relating a score on the item to a score on the latent construct differs across groups. Metric invariance requires that equality constraints be placed on factor loadings across groups, while allowing the factor variances and covariances to be free.

Note, for the metric of the factor to be identified, the factor variance or one of the factor loadings must be fixed to 1. If the factor loadings are found to be invariant, this does not mean that they are actually identical because the factor variances and covariances are allowed to vary across groups. Instead, invariant factor loadings in one group are said to be proportionally equivalent to corresponding loadings in the other

group (Bontempo & Hofer, 2007). “Loadings standardized to the common-factor variance would each differ from the corresponding loading in another group by the same proportion—the ratio of the variance in each group” (pg.51) (Bontempo, 2007).

The factor variances are freely estimated in all but the first group. This condition creates a test of proportionality when equality constraints are imposed on the loadings. Because the metric model is a more constrained model, the fit will be poorer than the configural model. The issue then becomes, ‘is the fit significantly worse’; if not, metric invariance has been obtained. If metric invariance is not obtained this implies that the factor(s) or groups of items have different meanings across groups.

Scalar invariance

Scalar invariance requires equality constraints on corresponding factor loadings and item intercepts across groups. This level of invariance requires the fitting of mean and covariance structure models. When assessing configural and metric invariance, only the covariance structures are examined. The scalar invariance model is compared against the metric invariance model and any significant worsening of fit suggests that the hypothesis of equal item intercepts is not supported. What this says, is that group comparisons of observed and factor means, factor variances and covariances may not be defensible.

Strict invariance

Strict (error) invariance requires that constraints be imposed on unique variances, unique means and factor loadings. Strict invariance implies that the item reliabilities and

therefore scale reliabilities are the same across groups (Beckstead, et al., 2008). Bontempo (2007) states that the strict invariance model forces the combined specific and random error components of each variable to be equivalent across groups such that differences in variance across groups are permitted only at the latent variable level. Thus, if a strict invariance model fits the data well, a researcher can be confident that measurement comparisons across groups involving factor mean and factor covariance structures are valid (Bontempo, 2007).

Binary Factorial Invariance

The instrument used in this study is the GDS-15 which has a yes/no response format. Because of this, multiple-group CFA measurement models with binary indicators require a different parameterization which requires modifications to the aforementioned procedures (Jöreskog & Moustaki, 2001; Roger E. Millsap & Yun-Tein, 2004; B. Muthén, & Asparouhov, T., 2002).

Each item on the measure is connected to its respective construct through a latent continuous response variable. This variable is cut by m-1 threshold parameters (where "m" is the number of response options) which produce observed response frequencies. Analyses are then based on a matrix of tetrachoric correlations. The latent response variables require additional scaling factors in order to assess group differences in the common factor mean and variance. To identify the model the following steps must be taken

1. The intercept parameters for all latent response variables must be fixed to zero in the first group.

2. Uniqueness variances need to be fixed to one in the first group.

3. The test of strict invariance will require the constraint of fixing uniqueness parameters in both groups.

As with any standard multiple-group confirmatory factor analysis, additional constraints are necessary in order to place the common-factor mean and variance on the same metric across groups.

There are two approaches presented for setting up these additional constraints (1) the Millsapp and Tein (2004) approach and (2) the Muthen and Asparouhov (2002) approach. The Millsap and Tein approach requires that

1. The first m-1 thresholds be constrained across all groups.
2. A second threshold or uniqueness (in the case of binary items, there would be no 2nd threshold) be constrained for one reference item in each group.

The Muthen and Asparouhov approach requires that thresholdsand loadings are constrained in a reduced model and that tests of selected items are conducted against a full model where thresholds and loadings for these items are freed while maintaining model identification through fixing the specific-variance to unity for the selected items (B. Muthén, & Asparouhov, T., 2002). MIMIC models (multiple indicator multiple causes) are suggested as a means of selecting items to be tested, because MIMIC models are sensitive to threshold invariance; modification indices produced by Mplus can also be used, tables 3 through 6 present the sequence of constraints used in the present study.

Table 3 *Configural invariance constraints*

Parameter Name	Constraints
Reference Group	
Loadings $\lambda(1)-\lambda(15)$	Free
Thresholds $\tau(1)-\tau(15)$	Free
Residuals $\theta(1)-\theta(15)$	Fixed to 1
Factor mean (α)	Fixed to 0 for factor 1 & factor 2
Factor variance (ψ)	Fixed to 1 for factor 1 & factor 2
Focal Group	
Loadings $\lambda(1)-\lambda(15)$	Free
Thresholds $\tau(1)-\tau(15)$	Free
Residuals $\theta(1)-\theta(15)$	Fixed to 1
Factor mean (α)	Fixed to 0 for factor 1 & factor 2
Factor variance (ψ)	Fixed to 1 for factor 1 & factor 2

(1)- (15) refer to item 1 through item 15

Table 4 Metric invariance constraints

Parameter Name	Constraints
Reference Group	
Loadings $\lambda(1)-\lambda(15)$	Held equal
Thresholds $\tau(1)-\tau(15)$	Free
Residuals $\theta(1)-\theta(15)$	Fixed to 1
Factor mean (α)	Fixed to 0 for factor 1 & factor 2
Factor variance (ψ)	Fixed to 1 for factor 1 & factor 2
Focal Group	
Loadings $\lambda(1)-\lambda(15)$	Held equal
Thresholds $\tau(1)-\tau(15)$	Free
Residuals $\theta(1)-\theta(15)$	Fixed to 1
Factor mean (α)	Fixed to 0 for factor 1 & factor 2
Factor variance (ψ)	Free for factor 1 & factor 2

(1)- (15) refer to item 1 through item 15

Table 5 Scalar invariance constraints

Parameter Name	Constraints
Reference Group	
Loadings $\lambda(1)-\lambda(15)$	Held equal
Thresholds $\tau(1)-\tau(15)$	Held equal
Residuals $\theta(1)-\theta(15)$	Fixed to 1
Factor mean (α)	Fixed to 0 for factor 1 & factor 2
Factor variance (ψ)	Fixed to 1 for factor 1 & factor 2
Focal Group	
Loadings $\lambda(1)-\lambda(15)$	Held equal
Thresholds $\tau(1)-\tau(15)$	Held equal
Residuals $\theta(1)-\theta(15)$	Fixed to 1
Factor mean (α)	Free for factor 1 & factor 2
Factor variance (ψ)	Free for factor 1 & factor 2

(1)- (15) refer to item 1 through item 15

Table 6 Residual invariance constraints

Parameter Name	Constraints
Reference Group	
Loadings $\lambda(1)-\lambda(15)$	Held equal
Thresholds $\tau(1)-\tau(15)$	Held equal
Residuals $\theta(1)-\theta(15)$	Fixed to 1
Factor mean (α)	Fixed to 0 for factor 1 & factor 2
Factor variance (ψ)	Fixed to 1 for factor 1 & factor 2
Focal Group	
Loadings $\lambda(1)-\lambda(15)$	Held equal
Thresholds $\tau(1)-\tau(15)$	Held equal
Residuals $\theta(1)-\theta(15)$	Free
Factor mean (α)	Free for factor 1 & factor 2
Factor variance (ψ)	Free for factor 1 & factor 2

(1)- (15) refer to item 1 through item 15

Assessing model fit

The Mplus program version 5.2, will be used to fit models to a matrix of tetrachoric correlations, using robust weighted least squares (RWLS) estimation (this is the weighted least squares mean and variance adjusted estimator (Flora & Curran, 2004; L. Muthén & B. O. Muthén, 2008). To assess the overall fit of a model to the data global fit indices such as CFI (comparative fit index), TLI (Tucker-Lewis index) chi-square goodness of fit test and RMSEA (root mean square error of approximation) will be used.

The comparative fit index (CFI) (Bentler, 1990) is a revised version of the Bentler and Bonnet normed fit index (Bentler & Bonnet, 1980) which adjusts for degrees of freedom and ranges in value of 0.00 to 1.00. The comparative fit index (CFI) compares the model with the baseline model. Hu and Bentler (Hu & Bentler, 1999) suggest that CFI be .95 or greater for good fit, although others have suggested that .90 is also considered to be acceptable fit (Byrne & Campbell, 1999).

The Tucker-Lewis index (TLI) indicates where a model lies on a continuum between a baseline model with unrelated observed variables and an ideal model that fits the data perfectly. A value of .95 or greater is also considered good fit for the TLI index. The root-mean-square-error-of approximation or RMSEA is an index of discrepancy between the model and the data per degree of freedom (P. J. Brown, et al., 2007). A value less than .06 suggests close fit between the model and the data (Browne & Cudeck, 1993; Hu & Bentler, 1999). The hypothesis of the chi-square goodness of fit test is that the model with the specified number of factors holds. With large samples, almost all models are rejected and with smaller samples model fit may go undetected. Because

of the chi-squares sensitivity to large sample sizes the aforementioned fit indices will be relied upon for this study.

IRTLR Analysis Procedure

DIF analyses with the IRTLR method (Thissen, 2001) consists of two parts: (1) anchor purification and (2) DIF detection. The presence of possible biased items can lead to inaccurate estimation of depression (ability), which in turn would contaminate the DIF investigation. As such prior to a DIF analysis a subset of DIF-free items should be identified and used to link the two groups being evaluated in the analysis. At large testing companies these anchor items are selected from a pool of established unbiased items (R.E. Millsap & Everson, 1993). When prior anchor items are not available items can be prescreened using procedures such as the Mantel-Haenszel test, logistic regression or MIMIC models.

Within IRTLR anchor items are identified through an iterative procedure. This process involves testing every item for DIF as a first step while treating all other items as anchor items. In other words if there were fifteen items on a depression instrument, item 1 would be evaluated for DIF while items 2 through 15 would be considered DIF free "temporarily" (temporary anchor set) in order to link items for both groups (women vs. men).

As mentioned before IRTLR involves the estimation of two hierarchically nested item response models which generate relevant χ^2 difference tests for DIF detection. As such during the first iteration in which all items are considered study items, IRTLR generates several nested model comparisons at least one for each item. For each item a model with all parameter estimates constrained to be equal for the reference and for the

focal groups is first compared with a model in which the parameters for the studied item are free to be estimated separately for the 2 groups.

For items modeled using the 2PL, the model comparison will have 2 *df*, one for each parameter constrained in the first model and freed in the second model. If the resulting model comparison has a χ^2 value greater than or equal to 3.84 (which indicates that at least one parameter might differ between groups at a nominal $\alpha = 0.05$), the item is classified as displaying DIF in one or more of its parameters, and the individual parameters would then be evaluated (Edelen, Thissen, Teresi, Kleinman, & Ocepek-Welikson, 2006).

Once all items on the instrument have been tested once using all other items as temporary anchor items, the items with potential DIF are removed and the process is repeated until no items are classified as potentially exhibiting DIF. This final set of items is then used as an anchor item set. With anchor items established, the studied items are now retested for DIF relative to the now-specified anchor. If the model comparison test indicates that at least one of the study item's parameters might differ between groups at a nominal $\alpha = 0.05$, IRTLR then generates parameter specific model comparison tests so that the source of the DIF can be identified.

The χ^2 values associated with these model comparison tests are generated first for the c-parameter if a 3PL is used, then for the a-parameter if the 2PL or graded response model is used and finally for the b-parameter. For each test the IRTLR lists the item parameter estimates associated with the less constrained model for the reference and

for the focal group and provides the focal group overall mean and standard deviation (relative to a N [0,1] distribution for the reference group) (Edelen, et al., 2006).

After item parameters with significant DIF have been identified, parameters for a final 2-group model that incorporates the identified DIF can be specified and estimated using MULTILOG (Edelen, et al., 2006; Jeanne A. Teresi, 2006). MULTILOG software generates item parameter estimates and standard errors, as well as summary statistics, item information, reliability estimates and the focal group overall mean and standard deviation (relative to a N [0,1] distribution for the reference group) (Jeanne A. Teresi, 2006; Thissen, W-H., & Bock, 2003). The results of this calibration can be used to interpret the DIF and to assess its impact at the item and scale levels. The IRTLR procedure will be used to assess DIF within each of five countries by gender and between countries.

Benjamini-Hochberg Procedure

Due to the multiple comparisons associated with the IRTLR procedure, the Benjamini-Hochberg procedure will be used to control the false discovery rate (type 1 error)(Thissen, Steinberg, & Kuang, 2002) . The Benjamini-Hochberg procedure has been used in the reporting of results from the National Assessment of Educational Progress (NAEP) (Braswell, et al., 2001) as well as other research contexts (Edelen, et al., 2006; Steinberg, 2001; Thissen, Steinberg, & Wainer, 1988; Thissen, et al., 1993) such as DIF analyses. The B-H procedure is a sequential approach which has greater power than the Bonferroni adjustment and an easier to implement.

As explained by (Steinberg, 2001) observed chi- square p-values are ranked from largest to smallest. The ranks (1 to 15) are used to adjust the critical p-values for statistical inference, according to the following formula: Equation 22

$$\left(\frac{rank-of-observed-p-value}{number-of-comparisons}\right)\times\left(level-of-significance\right)$$

A level of .05 will be used for all comparisons; this procedure controls the false positive rate, so that no more than 5% of the results marked significant for DIF may be in the wrong direction. The total number of comparisons is 15 (15 items) for generating the critical p-values used for each statistical test.

For a more detailed treatment of the procedure see (Thissen, et al., 2002) & (Benjamini & Hochberg, 1995).

CHAPTER 4: RESULTS

INTRODUCTION

In this section, findings of the measurement invariance analyses performed on the Geriatric Depression Scale Short Form (GDS-15) are described in detail. The overall goal of this study was to compare two psychometric techniques (1) multiple group confirmatory factor analysis for binary indicators and (2) item response theory likelihood ratio tests, in their ability to evaluate measurement invariance across gender and country of origin. The results are sequenced as follows: (1) descriptive statistics with country gender group comparisons, (2) multiple group confirmatory factor analysis and (3) item response theory likelihood ratio test analyses.

Descriptive Statistics: Country-Gender Comparisons

The GDS-15 is a 15 item yes/no response format depression screening instrument with scores that can range from 0 to 15. The scoring protocol states that individuals with scores less than 6, no active depressive symptomatology, 6 and above further screening with a medical professional is needed. Across all five countries the average depression score fell below a score of six, which would indicate no active depressive symptomatology see Table 7. Within all countries and across all countries, the mean score on the GDS-15 was higher for women; additional demographics are presented in Table 8.

Across all five countries the depression scores were positively skewed, which based on the cutoff value for depression being greater than 6, indicates that a large

portion of each sample would not be classified as having depression and only a small group of people would be classified as exhibiting depressive symptoms. GDS-15 total test scores had a skewness of 1.565 in Argentina, 1.403 in Cuba, 1.178 in Chile, 1.345 in Mexico and 1.649 in Uruguay.

The skewness of the data was addressed via the Mplus program and the robust weighted least squares (RWLS) estimation procedure. Research by (Flora & Curran, 2004) has shown that the (RWLS) is a "theoretically appropriate estimation procedure for binary response items and it produces accurate test statistics, parameter estimates and standard errors under both normal and non-normal latent response distributions across all sample sizes and model complexities" (p.489) (P. J. Brown, et al., 2007; Flora & Curran, 2004).

Table 7 Mean GDS-15 scores

	Women		Men	
Country	*M*	*SD*	*M*	*SD*
Argentina	2.53	2.9	2.29	2.72
Mexico	2.94	3.21	2.35	2.63
Chile	3.96	3.58	3.24	3.10
Cuba	3.27	3.41	1.81	2.41
Uruguay	2.70	3.08	2.02	2.55
Across all countries	3.11	3.30	2.27	2.69

Table 8 Demographic characteristics of countries

	N	Age	Female	Male
Argentina	1043	71 (7.2)	63%	37%
Mexico	1876	65 (9.8)	73%	27%
Uruguay	1450	71 (7.3)	63%	37%
Chile	1301	72 (8.0)	66%	34%
Cuba	1905	72 (8.9)	63%	37%

Multiple-Group Confirmatory Factor Analysis

Preliminary exploratory factor analyses

The first stage of these analyses focused on establishing what the underlying factor structure of the GDS-15 was by fitting exploratory factor analysis models across all five countries (Argentina, Mexico, Cuba, Uruguay and Chile) separately. The Mplus program (Version 5.2) (L. Muthén & B. O. Muthén, 2008), was used to fit models to a matrix of tetrachoric correlations, using robust weighted least squares estimation (RWLS) (Flora & Curran, 2004).

An initial EFA analysis was necessary because to date there have only been four factor analytical studies of the GDS-15 (P. J. Brown, et al., 2007; Friedman, et al., 2005; Incalzi, Cesari, Pedone, & Carbonin, 2003; Mitchell, et al., 1993) (only one used multiple group analysis, Brown & Woods), with two studies identifying a 2 factor model and 2 studies identifying a 3 factor model. Brown and colleagues replicated two of the 3 factors that Mitchell and colleagues found initially. With such discordant reporting of the factor structure across studies an EFA was a necessary first step.

One, two and three factors were extracted for the GDS-15 within each country, eigenvalues were (Argentina: 7.53, 1.88, 0.90, Mexico: 7.87, 1.42, 0.89, Cuba: 8.57, 1.05, 0.93, Chile: 8.11, 1.01, 0.95, and Uruguay: 8.06, 1.19, 0.93 respectively. Representative scree-plots for each country are presented in Figure 1, Figure 2, Figure 3, Figure 4 and Figure 5. Factor loading patterns are presented in Table 9, Table 10, Table 11, 12 and 13. A third factor was also extracted, however, the 1 and 2 factors were selected for further

evaluation because they provided the most parsimonious, substantively meaningful solution and because each factor was well represented (more than 3 items loaded on each factor). Across countries either no items loaded on a third factor or there were less than three items loading on the third factor.

Examination of the screeplots by country indicated that there was 1 dominant factor and possibly a second factor across countries. To rule out the existence of a one dimensional construct, a one-factor CFA model was estimated for all countries (N=7575) simultaneously and fit was assessed. The one-factor CFA fit the data poorly, with a χ^2 (328) = 2167.856, p<.00001, RMSEA = .062, CFI=.935 and TLI=.968. Based on the results of the exploratory factor analyses and the 1-factor total group CFA, separate CFA's were run within each country to confirm a final factor structure.

A summary of the results of individual 1 and 2 factor CFA's within each country can be found in Table 14 and Table 15. These results indicate that for the countries of Argentina, Mexico and Uruguay a one- factor model fit poorly; however for the countries of Chile and Cuba, the one- factor model fit the data adequately. There was an increase in overall global model fit, between the one- factor CFA and the two- factor CFA for the countries of Argentina, Mexico and Uruguay. Based on these results, for the countries of Chile and Cuba a one-factor structure was qualitatively defined as general depressive affect and in the countries of Argentina, Mexico and Uruguay a two-factor structure was qualitatively defined as life satisfaction (v1, v5, v7, v11 and v13) and general depressive affect (v2-v4, v6, v8-v10, v12 and v14-v15). This two-factor pattern of loadings replicates the results of Brown (2007) and two of the three original factors that Mitchell

and colleagues found (Mitchell, et al., 1993) and is defined in the same way. After establishing the underlying factor structure, within country invariance analyses by gender was evaluated.

Invariance Models by Gender within Country

The following section presents the results of the measurement invariance hypotheses with respect to the equivalence of factor loadings (metric invariance), thresholds (scalar invariance), and uniqueness' (strict invariance) by gender within the countries of Chile, Cuba, Argentina, Mexico and Uruguay.

Chile

The extent to which an item factor model measuring geriatric depression (with 15 observed items) exhibited measurement invariance between women and men in the country of Chile was examined using Mplus v. 5.2 (L. Muthén & B. O. Muthén, 2008). WLSMV estimation including a probit link and the THETA parameterization was used to estimate all models (L. Muthén & B. Muthén, 2008). Missing data was handled by maximum likelihood estimation assuming missing at random. WLSMV provides weighted least squares parameter estimates using a diagonal weighted matrix with standard errors and mean- and- variance adjusted chi-squared test statistic that use a full weight matrix (B. Muthén, du Toit, & Spisic, 1997).

For nested models the difference between the fit functions is not distributed as chi-square and the degrees of freedom is not the simple difference in free parameters. However, an appropriate chi-square statistic and degrees of freedom are calculated according to formulas in the MPlus Technical Appendices (www.statmodel.com).Thus, model fit statistics describe the fit of the item factor model to the polychoric correlation matrix among the items for each group.

Nested model comparisons were conducted using the Mplus chi-square DIFFTEST procedure. Model fit was evaluated with relative fit indices CFI, TLI, and RMSEA. For the CFI and TLI indices values above .95 indicate a good fit. For the RMSEA, a value less than .06 is considered to indicate good fit. For the Chilean gender model global fit indices were adequate and ranged from a CFI of .966 to .983, TLI of .983 to .990 and RMSEA of .037 to .047, all global fit indices are summarized in Table 16.

Invariance hypotheses tests

Configural

A configural invariance model was initially specified in which one factor was estimated simultaneously for Chilean women and men. The factor variance was fixed to 1 and the factor mean was fixed to 0 in each group for identification, such that all item factor loadings and thresholds (1 per item given a binary response option) were then estimated. The residual variances are not uniquely identified in the configural invariance model and as such were all constrained to 1 in both groups. As shown in $\chi^2(119) = 287.15, p < .0000$ Table 10, the configural invariance model fit adequately across groups, , RMSEA = .047, CFI = .966, TLI = .983. The analysis proceeded by applying parameter constraints in successive models to examine potential decreases in fit resulting from measurement non-invariance between Chilean men and Chilean women, with women as the reference group and men as the focal group.

Metric

Equality of the unstandardized item factor loadings between groups was then examined in a metric invariance model. The factor variance was fixed to 1 in women for identification but was freely estimated in men; the factor mean was fixed to 0 in both groups for identification. All factor loadings were constrained equal across groups, all item thresholds were estimated, and all residual variances were constrained to 1 across groups.

The metric invariance model did not result in a decrement in fit, DIFFTEST (12) = 4.640, p = .9689. Modification indices did not suggest any points of localized misfit for the constrained loadings. The fact that the metric invariance hypothesis was supported indicates that the items were related to the latent factor equivalently across groups, or more simply, that the same latent factor was being measured in each group.

Scalar

Equality of the unstandardized item thresholds across groups was then examined in a scalar invariance model. The factor variance and mean were fixed to 1 and 0, respectively, in women for identification, but the factor variance and mean were then estimated for men. All factor loadings and item thresholds were constrained equal across groups; all residual variances were still constrained equal to 1 in both groups. The full scalar invariance model did not fit significantly worse than the metric invariance model, DIFFTEST (13) = 18.922, p < .1256.

The fact that scalar invariance (i.e., “strong invariance”) held indicates that all items have the same expected response for each item threshold at the same absolute level of the trait, or more simply, that the observed differences in the proportion of responses in each category for those items was due to factor mean differences only.

Strict

Equality of the unstandardized residual variances across groups was then examined in a residual variance invariance model. The model comparison at this step proceeded backwards, such that a model with all residual variances freely estimated in the men was fitted first, and then compared with a model in which the residual variances for the invariant items (v1-v15) were fixed to 1 in the men. The residual variances in the women were all fixed to 1 for identification in both models, and the rest of the model parameters were estimated as described for the last scalar invariance model. The model with the residual variances for invariant items constrained to 1 (to be equal to the women) did not fit significantly worse than the model with those residual variances freed, DIFFTEST (13) = 6.670, p = .9183, indicating that residual variance invariance held for all items.

Residual variance invariance (i.e., “strict invariance”) being supported indicates that the amount of item variance not accounted for by the factor was the same across Chilean men and women for all items.

Summary for Chile

In conclusion, these analyses showed that full measurement invariance was obtained across Chilean men and women – that is, the relationships of the items to the latent factor of general depressive affect was equivalent between men and women. In addition, full scalar invariance held which indicates that all items have the same expected response for each item threshold at the same absolute level of the trait, or more simply, that the observed differences in the proportion of responses in each category for those items was due to factor mean differences only. Finally, full residual variance invariance (i.e., "strict invariance") being supported indicated that the amount of item variance not accounted for by the factor was the same across groups for all items. Invariance hypotheses tests are summarized in Table 17.

.

Cuba

For the Cuban gender model, global fit indices were adequate and ranged from a CFI of .966 to .981, TLI of .981 to .987 and an RMSEA of .041 to .049, all global fit indices are summarized in Table 18.

Invariance hypotheses tests

Configural

A configural invariance model with a one-factor structure (general depressive affect) was estimated simultaneously for Cuban women and men. The previous protocol for model identification was followed for these analyses. As shown in Table 18, the configural invariance model had adequate fit $\chi^2(113) = 348.55, p < .0000$, RMSEA = .049, CFI = .966, TLI = .981. The analysis proceeded by applying parameter constraints in successive models to examine potential decreases in fit resulting from measurement non-invariance between Cuban men and women, with women as the reference group and men as the focal group.

Metric

Equality of the unstandardized item factor loadings between groups was then examined in a metric invariance model. All factor loadings were constrained equal across groups, all item thresholds were estimated, and all residual variances were constrained to 1 across groups. The metric invariance model did not result in a decrement in fit, DIFFTEST (12) = 11.06, p = .5236. Support for the metric invariance hypothesis

indicates that the items were related to the latent factor equivalently across groups, or more simply, that the same latent factor was being measured in each group.

Scalar

Equality of the unstandardized item thresholds across groups was then examined in a scalar invariance model. The factor variance and mean were fixed to 1 and 0, respectively, in women for identification, but the factor variance and mean were then estimated for men. All factor loadings and item thresholds were constrained equal across groups; all residual variances were still constrained equal to 1 in both groups. The full scalar invariance model did not fit significantly worse than the metric invariance model, DIFFTEST (13) = 21.95, $p < .0561$.

The fact that scalar invariance (i.e., "strong invariance") held indicates that all items have the same expected response for each item threshold at the same absolute level of the trait, or more simply, that the observed differences in the proportion of responses in each category for those items was due to factor mean differences only.

Strict

Equality of the unstandardized residual variances across groups was then examined in a residual variance invariance model. The model with the residual variances for invariant items constrained to 1 (to be equal to the women) did not fit significantly worse than the model with those residual variances freed, DIFFTEST (13) = 14.60, $p = .3326$, indicating that residual variance invariance held for all items.

Residual variance invariance (i.e., “strict invariance”) being supported indicates that the amount of item variance not accounted for by the factor was the same across groups in all items.

Summary for Cuba

In conclusion, these analyses indicate that full measurement invariance was obtained across Cuban men and women – that is, the relationships of the items to the latent factor of general depressive affect was equivalent between men and women. In addition, full scalar invariance held which indicates that all items have the same expected response for each item threshold at the same absolute level of the trait. Finally, full residual variance invariance (i.e., “strict invariance”) being supported indicated that the amount of item variance not accounted for by the factor was the same across groups for all items. Invariance hypotheses tests are summarized in Table 19.

Argentina

For the Argentinean gender model, global fit indices were adequate and ranged from a CFI of .949 to .962, TLI .972 to .978 and RMSEA .046 to .052. All global fit indices are summarized in Table 20.

Invariance hypotheses tests

Configural

A configural invariance model with a two-factor structure (Life Satisfaction and General Depressive Affect) was estimated simultaneously for Argentinean women and men. As shown in Table 20, the configural invariance model had marginal fit $\chi^2(92 = 217.88, p\langle.0000)$, RMSEA = .052, CFI = .949, TLI = .972. The analysis proceeded by applying parameter constraints in successive models to examine potential decreases in fit resulting from measurement non-invariance between Argentinean men and women, with women as the reference group and men as the focal group.

Metric

Equality of the unstandardized item factor loadings between groups was then examined in a metric invariance model. The metric invariance model did not result in a decrement in fit, DIFFTEST (10) = 10.24, p = .5245. Modification indices did not suggest any points of localized misfit for the constrained loadings. The fact that the metric invariance hypothesis was supported indicates that the items were related to the

latent factor equivalently across groups, or more simply, that the same latent factor was being measured in each group.

Scalar

Equality of the unstandardized item thresholds across groups was then examined in a scalar invariance model. The factor variance and mean were fixed to 1 and 0, respectively, in women for identification, but the factor variance and mean were then estimated for men. All factor loadings and item thresholds were constrained equal across groups; all residual variances were still constrained equal to 1 in both groups. The full scalar invariance model fit significantly worse than the metric invariance model, DIFFTEST (12) = 25.700, $p < .001$.

The modification indices suggested that the threshold of item 9 ("do you prefer to stay at home rather than going out and doing new things?") was the largest source of misfit and should be freed. After doing so, the partial scalar invariance model did not fit significantly worse than the full metric invariance model, DIFFTEST (11) = 25.7, P<.2799.

Support for partial scalar invariance indicates that items 1-8 and 10-15 have the same expected response for each item threshold at the same absolute level of the trait, or more simply, that the observed differences in the proportion of responses in each category for those items was due to factor mean differences only.

Strict

Equality of the unstandardized residual variances across groups was then examined in a residual variance invariance model. The model with the residual variances for invariant items constrained to 1 (to be equal to the women) did not fit significantly worse than the model with those residual variances freed, DIFFTEST (12) = 15.478, p = .2163, indicating that residual variance invariance held for items 1-8 and 10-15.

Partial residual variance invariance (i.e., "strict invariance") being supported indicates that the amount of item variance not accounted for by the factor was the same across groups in items 1-8 and 10-15. The residual variance for item 9 was assumed non-invariant because of lack of threshold/scalar invariance found previously, and was not tested.

Summary for Argentina

In conclusion, these analyses indicate that partial measurement invariance was obtained across Argentinean men and women – that is, full metric invariance was obtained, meaning that the relationships of the items to the latent factor of general depressive affect were equivalent between men and women. Partial scalar invariance was obtained for items 1-8 and 10-15 indicating that these items had the same expected response for each item threshold at the same absolute level of the trait across Argentinean men and women. Finally, partial strict invariance was obtained for items 1-8 and 10-15, indicating that the amount of item variance accounted for by the factor was the same across groups.

Based on the lack of full scalar invariance, the observed values for item 9 will differ between men and women in Argentina at a given level of the latent factors. Invariance hypotheses tests are summarized in Table 21.

.

Mexico

For the Mexican gender model, global fit indices were adequate and ranged from a CFI of.969 to .980, TLI .982 to .986 and RMSEA .038 to .043; these results are summarized in Table 22.

Invariance hypotheses tests

Configural

A configural invariance model with a two-factor structure (Life Satisfaction and General Depressive Affect) was estimated simultaneously for Mexican women and men. As shown in Table 16, the configural invariance model had adequate fit $\chi^2(108) = 295.67, p < .0000$, RMSEA = .043, CFI = .969, TLI = .982. The analysis proceeded by applying parameter constraints in successive models to examine potential decreases in fit resulting from measurement non-invariance between Mexican men and women, with women as the reference group and men as the focal group.

Metric

Equality of the unstandardized item factor loadings between groups was then examined in a metric invariance model. The metric invariance model did not result in a decrement in fit, DIFFTEST (11) = 13.427, p = .2663. The fact that the metric invariance hypothesis was supported indicates that the items were related to the latent factor equivalently across groups, or more simply, that the same latent factor was being measured in each group.

Scalar

Equality of the unstandardized item thresholds across groups was then examined in a scalar invariance model. The factor variance and mean were fixed to 1 and 0, respectively, in women for identification, but the factor variance and mean were then estimated for men. All factor loadings and item thresholds were constrained equal across groups; all residual variances were still constrained equal to 1 in both groups. The full scalar invariance model fit significantly worse than the metric invariance model, DIFFTEST (12) = 37.123, $p < .0002$.

The modification indices suggested that the threshold of item 8 ("do you often feel helpless?") was the largest source of misfit and should be freed. After doing so, the partial scalar invariance model still had significantly worse fit than the full metric invariance model, DIFFTEST (11) = 19.978, P<.0456.

The modification indices suggested that the threshold of item 6 ("are you afraid that something bad is going to happen to you?") was the largest remaining source of misfit and should be freed. After doing so, the new partial scalar invariance model (with thresholds for items 8 and 6 freed) did not fit significantly worse than the full metric invariance model, DIFFTEST (10) = 12.809, $p = .2346$. Support for partial scalar invariance indicates that items 1-5, 7 and 9-15 have the same absolute level of the trait, or more simply, that the observed differences in the proportion of responses in each category for those items was due to factor mean differences only.

Strict

Equality of the unstandardized residual variances across groups was then examined in a residual variance invariance model. The model with the residual variances for invariant items constrained to 1 (to be equal to the women) did not fit significantly worse than the model with those residual variances freed, DIFFTEST (10) = 15.106, p = .1282, indicating that residual variance invariance held for items 1-5, 7 and 9-15.

Partial residual variance invariance (i.e., "strict invariance") being supported indicates that the amount of item variance not accounted for by the factor was the same across groups in items 1-5, 7 and 9-15. The residual variance for items 8 and 6 was assumed non-invariant because of lack of threshold/scalar invariance found previously, and was not tested.

Summary for Mexico

In conclusion, these analyses indicate that partial measurement invariance was obtained across Mexican men and women – that is, full metric invariance was obtained, meaning that the relationships of the items to the latent factors of life satisfaction and general depressive affect were equivalent between men and women. Partial scalar invariance was obtained for items 1-5, 7 and 9-15 indicating that these items had the same expected response for each item threshold at the same absolute level of the trait across Mexican men and women. Finally, partial strict invariance was obtained for items 1-5, 7 and 9-15; indicating that the amount of item variance not accounted for by the factor was the same across groups.

Based on the lack of full scalar invariance, the observed values for items 6 and 8 will differ between men and women in Mexico at a given level of the latent factors. Invariance hypotheses tests are summarized in Table 23.

.

Uruguay

For the Uruguayan gender model, global fit indices were adequate and ranged from a CFI of .968 to .978, TLI of .983 to .987 and an RMSEA of .039 to .043, global fit indices are summarized in Table 24.

Invariance hypotheses tests

Configural

A configural invariance model with a two-factor structure (Life Satisfaction and General Depressive Affect) was estimated simultaneously for Uruguayan women and men. As shown in Table 24, the configural invariance model had adequate fit $\chi^2(113) = 262.91, p < .0000$, RMSEA = .043, CFI = .968, TLI = .983. The analysis proceeded by applying parameter constraints in successive models to examine potential decreases in fit resulting from measurement non-invariance between Uruguayan men and women, with women as the reference group.

Metric

Equality of the unstandardized item factor loadings between groups was then examined in a metric invariance model. The metric invariance model did not result in a decrement in fit, DIFFTEST (11) = 14.469, p = .2081. The fact that the metric invariance hypothesis was supported indicates that the items were related to the latent factor equivalently across groups, or more simply, that the same latent factor was being measured in each group.

Scalar

Equality of the unstandardized item thresholds across groups was then examined in a scalar invariance model. The factor variance and mean were fixed to 1 and 0, respectively, in women for identification, but the factor variance and mean were then estimated for men. All factor loadings and item thresholds were constrained equal across groups; all residual variances were still constrained equal to 1 in both groups. The full scalar invariance model fit significantly worse than the metric invariance model, DIFFTEST (12) = 22.296, p = .0343.

The modification indices suggested that the threshold of item 15 (“do you think that most people are better off than you?”) was the largest source of misfit and should be freed. After doing so, the new partial scalar invariance model did not fit significantly worse than the full metric invariance model, DIFFTEST (11) = 15.647, P<.1547.

Support for partial scalar invariance indicates that items 1-14 have the same expected response for each item threshold at the same absolute level of the trait, or more simply, that the observed differences in the proportion of responses in each category for those items was due to factor mean differences only.

Strict

Equality of the unstandardized residual variances across groups was then examined in a residual variance invariance model. The model with the residual variances for invariant items constrained to 1 (to be equal to the women) did not fit significantly worse than the model with those residual variances freed, DIFFTEST (12) = 14.503, p = .2698, indicating that residual variance invariance held for items 1-14.

Partial residual variance invariance (i.e., "strict invariance") being supported indicates that the amount of item variance not accounted for by the factor was the same across groups for items 1-14. The residual variance for item 15 was assumed non-invariant because of lack of threshold/scalar invariance found previously, and was not tested.

Summary for Uruguay

In conclusion, these analyses indicate that partial measurement invariance was obtained across Uruguayan men and women – that is, full metric invariance was obtained, meaning that the relationships of the items to the latent factors of life satisfaction and general depressive affect were equivalent between men and women. Partial scalar invariance was obtained for items 1-14 indicating that these items had the same expected response for each item threshold at the same absolute level of the trait across Uruguayan men and women. Finally, partial strict invariance was obtained for items 1-14, indicating that the amount of item variance not accounted for by the factor was the same across groups.

Based on the lack of full scalar invariance, the observed values of item15 will differ between men and women in Uruguay at a given level of the latent factors. Invariance hypotheses tests are summarized in Table 25.

.

Invariance Models by Cross-Country Comparisons

The following section presents the results of the measurement invariance testing with respect to the equivalence of factor loadings '***metric invariance***', thresholds ***'scalar invariance'***, and uniqueness' ***'strict invariance'*** by cross-country comparisons between, Chile and Cuba, Mexico and Uruguay, Mexico and Argentina and Argentina and Uruguay. The cross-country comparisons are based on factor structures within each country, for example in the countries of Chile and Cuba there is a one-factor structure and in Mexico, Cuba and Uruguay there is a two-factor structure with the same pattern of loadings.

Chile by Cuba

For the Chile by Cuba model, global fit indices were adequate and ranged from a CFI of .963 to .978, TLI of .982 to .988 and RMSEA of .042 to .05. These results are summarized in Table 26.

Invariance hypotheses tests

Configural

A configural invariance model was initially specified in which a single factor (General Depressive Affect) was estimated simultaneously for the countries of Chile and Cuba. The factor variance was fixed to 1 and the factor mean was fixed to 0 in each group for identification, such that all item factor loadings and thresholds (1 per item given a binary response option) were then estimated.

The residual variances are not uniquely identified in the configural invariance model and as such were all constrained to 1 in both groups. As shown in Table 26, the configural invariance model had adequate fit across countries, $\chi^2(139) = 637.50, p < .0000$, RMSEA = .049, CFI = .963, TLI = .983. The analysis proceeded by applying parameter constraints in successive models to examine potential decreases in fit resulting from measurement non-invariance between the countries of Chile and Cuba, with Chile as the reference group and Cuba as the focal group.

Metric

Equality of the unstandardized item factor loadings between groups was then examined in a metric invariance model. The factor variance was fixed to 1 in Chile for identification but was freely estimated in Cuba; the factor mean was fixed to 0 in both groups for identification. All factor loadings were constrained equal across groups, all item thresholds were estimated, and all residual variances were constrained to 1 across groups.

The metric invariance model did result in a decrement in fit, DIFFTEST (12) = 31.31, p = .0018. Modification indices suggested that the lack of a correlation between the errors of item 5 and 7 was the largest source of misfit, so a correlation was added. After doing so the partial metric invariance model did not fit significantly worse than the full configural invariance model, DIFFTEST (11) 17.49, p=.0941. Support for partial metric invariance indicates that items 1-4, 6, and 8-15 were related to the latent factor equivalently across countries (with the exception of items 5 and 7).

Scalar

Equality of the unstandardized item thresholds across groups was then examined in a scalar invariance model. The factor variance and mean were fixed to 1 and 0, respectively, in Chile for identification, but the factor variance and mean was then estimated for Cuba. All factor loadings and item thresholds were constrained equally across groups; all residual variances were still constrained equal to 1 in both groups. The full scalar invariance model fit significantly worse than the partial metric invariance model, DIFFTEST (13) = 232.88, $p < .0000$.

Modification indices suggested that the thresholds of item 1 and item 15 were the largest source of misfit and should be freed. After doing so, the partial scalar invariance model still had significantly worse fit than the partial metric invariance model, DIFFTEST (11) = 98.76, $p <.0000$. The modification indices then suggested that the thresholds of item 5 and 7 were the largest source of misfit and should be freed. After doing so, the new partial scalar invariance model (with the thresholds of items 1, 5, 7 and 15 freed) still had significantly worse fit than the partial metric invariance model, DIFFTEST (19) = 19.05, p=.0248. The modification indices then suggested that the threshold of item 9 was the largest source of misfit and should be freed. After doing, so the new partial scalar invariance model (with the thresholds of 1, 5, 7, 9 and 15 freed) did not fit significantly worse than the partial metric invariance model, DIFFTEST (8) =15.15, p=.0562.

Support for partial scalar invariance indicates that items 2-4, 6, 8 and 10-14 have the same expected response for each item threshold at the same absolute level of the trait,

or more simply, that the observed differences in the proportion of responses in each category for those items was due to the factor mean differences only.

Strict

Equality of the unstandardized residual variances across groups was then examined in a residual variance invariance model. The model comparison at this step proceeded backwards, such that a model with all residual variances freely estimated in the country of Cuba was fitted first, and then compared with a model in which the residual variances for the invariant items (2-4, 6, 8 and 10-14) were fixed to 1 in Cuba.

The residual variances in the country of Chile were all fixed to 1 for identification in both models, and the rest of the model parameters were estimated as described for the last scalar invariance model. The model with the residual variances for invariant items constrained to 1 (to be equal in Cuba) fit significantly worse than the model with those residual variances freed, DIFFTEST (9) = 22.92, p =0.0064. The modification indices suggested that the residual variance for items 2 and 4 were the largest source of remaining misfit and should be freed.

After doing so, the new fixed partial residual invariance model did not fit significantly worse than the freed residual invariance model DIFFTEST (7) = 12.22, p = 0.0935. Support for partial residual variance invariance (i.e., "strict invariance") indicates that the amount of item variance not accounted for by the factor was the same across groups in items 3, 6, 8 and 10-14; the residual variance for items 1, 5, 7, 9 and 15

were assumed non-invariant because of lack of threshold invariance found previously, and were not tested.

Summary for Chile by Cuba

In conclusion, these analyses indicate that partial measurement invariance was obtained across Chile and Cuba – that is, the relationships of items 1-4, 6 and 8-15 were related to the latent factor of general depressive affect equivalently between the countries of Chile and Cuba or that the same latent factor was being measured in each group. In addition, for items 2-4, 6, 8 and 10-14 the observed differences in the proportion of responses in each category for these items was due to differences in the factor mean only. Finally, the amount of item variance not accounted for by the factor was the same across groups for items 3, 6, 8 and 10-14.

The lack of full metric and scalar invariance means that the observed values of items 1, 5, 7, 9 and 15 will differ between older adults in Chile and Cuba at a given level of the latent factor. Invariance hypotheses tests are summarized in Table 27.

Mexico by Uruguay

For the Mexico by Uruguay model, global fit indices were adequate and ranged from a CFI of .966 to .978, TLI of .983 to .987 and RMSEA of .037 to .044. These results are summarized in Table 28.

Invariance hypotheses tests

Configural

A configural invariance model was initially specified in which two factors (Life Satisfaction and General Depressive Affect) were estimated simultaneously for the countries of Mexico and Uruguay. The factor variance was fixed to 1 and the factor mean was fixed to 0 in each group for identification, such that all item factor loadings and thresholds (1 per item given a binary response option) were then estimated.

The residual variances are not uniquely identified in the configural invariance model and as such were all constrained to 1 in both groups. As shown in Table 28, the configural invariance model fit well across groups, $\chi^2(132) = 521.54, p < .0000$ RMSEA = .044, CFI = .966, TLI = .983. The analysis proceeded by applying parameter constraints in successive models to examine potential decreases in fit resulting from measurement non-invariance between the countries of Mexico and Uruguay, with Uruguay as the reference group and Mexico as the focal group.

Metric

Equality of the unstandardized item factor loadings between groups was then examined in a metric invariance model. The factor variance was fixed to 1 in Uruguay for identification but was freely estimated in Mexico; the factor mean was fixed to 0 in both groups for identification. All factor loadings were constrained equal across groups, all item thresholds were estimated, and all residual variances were constrained to 1 across groups.

The metric invariance model did result in a decrement in fit, DIFFTEST (12) = 29.88, p = .0029. Modification indices suggested that adding a correlation between the errors for item 1 and item 3 would improve fit. After doing so the partial metric invariance model still had significantly worse fit than the full configural invariance model, DIFFTEST (11) 19.87, p=.0471.

The modification indices suggested that adding a correlation between the errors for item 10 and item 15 would improve fit. After doing so the new partial metric invariance model (with error correlations for items 1, 3, 10 and 15) did not have significantly worse fit than the configural invariance model DIFFTEST (10) 15.16, p=.1261. Support for partial metric invariance indicates that items 2-9 and 11-14 were related to the latent factor equivalently across countries (with the exception of items 1, 3, 10 and 15).

Scalar

Equality of the unstandardized item thresholds across groups was then examined in a scalar invariance model. The factor variance and mean were fixed to 1 and 0, respectively, in Uruguay for identification, but the factor variance and mean was then estimated for Mexico. All factor loadings and item thresholds were constrained equal across groups; all residual variances were still constrained equal to 1 in both groups. The full scalar invariance model fit significantly worse than the partial metric invariance model, DIFFTEST (12) = 93.70, $p < .0001$.

Modification indices suggested that the thresholds of items 9 and 10 were the largest source of misfit and should be freed. After doing so, the partial scalar invariance model still had significantly worse fit than the partial metric invariance model, DIFFTEST (10) = 48.608, $p < .0001$. The modification indices then suggested that the thresholds of items 2 and 4 were the largest source of misfit and should be freed. After doing so, the new partial scalar invariance model (with the thresholds of items 9, 10, 2 and 4 freed) still fit significantly worse than the partial metric invariance model, DIFFTEST (9) = 18.46, p=.0301.

The modification indices then suggested that the threshold for item 15 was the largest remaining source of misfit and should be freed. After doing so, the new partial scalar invariance model (with the thresholds for items 9, 10, 2, 4 and 15 freed) did not fit significantly worse than the partial metric invariance model, DIFFTEST (8) = 12.88, p=.1158.

Support for partial scalar invariance indicates that items 1, 3, 5-8, and 11-14 have the same expected response for each item threshold at the same absolute level of the trait, or more simply, that the observed differences in the proportion of responses in each category for those items was due to the factor mean differences only.

Strict

Equality of the unstandardized residual variances across groups was then examined in a residual variance invariance model. The model comparison at this step proceeded backwards, such that a model with all residual variances freely estimated in the country of Mexico was fitted first, and then compared with a model in which the residual variances for the invariant items (1, 3, 5-8, and 11-14) were fixed to 1 in Mexico.

The residual variances in the country of Uruguay were all fixed to 1 for identification in both models, and the rest of the model parameters were estimated as described for the last scalar invariance model. The model with the residual variances for invariant items constrained to 1 (to be equal in Uruguay) fit significantly worse than the model with those residual variances freed, DIFFTEST (8) = 16.45, p =0.0363. The modification indices suggested that relaxing the constraints on the residual variance for item 3 would improve fit. After doing so, the new fixed partial residual invariance model did not fit significantly worse than the freed residual invariance model DIFFTEST (7) = 11.487, p = 0.1187. Support for partial residual variance invariance (i.e., "strict invariance") indicates that the amount of item variance not accounted for by the factor was the same across groups in items 1, 5-8, and 11-14); the residual variance for items 2,

4, 9, 10 and 15 were assumed non-invariant because of lack of threshold invariance found previously, and were not tested.

Summary for Mexico by Uruguay

In conclusion, these analyses indicate that partial measurement invariance was obtained across Mexico and Uruguay – that is, the relationships of items 2, 4, 9 and 11-14 (sans the error correlations of items 1 and 3 and 10 and 15), to the latent factors of life satisfaction and general depressive affect were equivalent between the countries of Mexico and Uruguay. The same expected response for the item thresholds of questions 1, 3, 5-8, and 11-15 were the same at absolute level of the trait. Finally, the amount of item variance not accounted for by the factor was the same across groups for items 1, 5-8, and 11-14.

The lack of full metric and scalar invariance means that the observed values of items 1, 2, 3, 4,9,10 and 15 will differ between older adults in Mexico and Uruguay at a given level of the latent factor. Invariance hypotheses tests are summarized in Table 29.

Mexico by Argentina

For the Mexico by Argentina model global fit indices were adequate and ranged from a CFI of .961 to .974, TLI of .979 to .984 and RMSEA of .040 to .046, these results are summarized in Table 30.

Invariance hypotheses tests

Configural

A configural invariance model was initially specified in which two factors (Life Satisfaction and General Depressive Affect) were estimated simultaneously for the countries of Mexico and Argentina. The factor variance was fixed to 1 and the factor mean was fixed to 0 in each group for identification, such that all item factor loadings and thresholds (1 per item given a binary response option) were then estimated.

The residual variances are not uniquely identified in the configural invariance model and as such were all constrained to 1 in both groups. As shown in Table 30, the configural invariance model had adequate fit across countries, $\chi^2(122) = 483.26, p < .0000$ RMSEA = .046, CFI = .961, TLI = .979. The analysis proceeded by applying parameter constraints in successive models to examine potential decreases in fit resulting from measurement non-invariance between the countries of Mexico and Argentina, with Argentina as the reference group and Mexico as the focal group.

Metric

Equality of the unstandardized item factor loadings between groups was then examined in a metric invariance model. The factor variance was fixed to 1 in Argentina for identification but was freely estimated in Mexico; the factor mean was fixed to 0 in both groups for identification. All factor loadings were constrained equal across groups, all item thresholds were estimated, and all residual variances were constrained to 1 across groups.

The metric invariance model did result in a decrement in fit, DIFFTEST (11) = 22.26, p = .0224. Modification indices suggested that adding a correlation between the errors for item 14 and item 15 would improve fit. After doing so the partial metric invariance model did not fit significantly worse than the full configural invariance model, DIFFTEST (11) 17.30, p=.0993. Support for partial metric invariance indicates that items 1-13 were related to the latent factor equivalently across countries (with the exception of items 14 and 15).

Scalar

Equality of the unstandardized item thresholds across groups was then examined in a scalar invariance model. The factor variance and mean were fixed to 1 and 0, respectively, in Argentina for identification, but the factor variance and mean was then estimated for Mexico. All factor loadings and item thresholds were constrained equally across groups; all residual variances were still constrained equal to 1 in both groups. The

full scalar invariance model fit significantly worse than the partial metric invariance model, DIFFTEST (12) = 134.60, $p < .0001$.

Modification indices suggested that the threshold of item 10 was the largest source of misfit and should be freed. After doing so, the partial scalar invariance model still had significantly worse fit than the partial metric invariance model, DIFFTEST (11) = 88.74, $p < .0000$. The modification indices then suggested that the thresholds of items 6 and 15 were the largest source of misfit and should be freed. After doing so, the new partial scalar invariance model (with the thresholds of items 6, 10, and 15 freed) still fit significantly worse than the partial metric invariance model, DIFFTEST (10) = 33.21, $p=.0003$.

The modification indices then suggested that the thresholds for items 5 and 8 were the largest remaining source of misfit and should be freed. After doing so, the new partial scalar invariance model (with the thresholds for items 5, 6, 8, 10 and 15 freed) did not fit significantly worse than the partial metric invariance model, DIFFTEST (8) = 9.51, $p=.3011$.

Support for partial scalar invariance indicates that items 1-4, 7, 9 and 11-14 have the same expected response for each item threshold at the same absolute level of the trait, or more simply, that the observed differences in the proportion of responses in each category for those items was due to the factor mean differences only.

Strict

Equality of the unstandardized residual variances across groups was then examined in a residual variance invariance model. The model comparison at this step proceeded backwards, such that a model with all residual variances freely estimated in the country of Mexico was fitted first, and then compared with a model in which the residual variances for the invariant items (1-4, 7, 9, and 11-14) were fixed to 1 in Mexico. The residual variances in the country of Argentina were all fixed to 1 for identification in both models, and the rest of the model parameters were estimated as described for the last scalar invariance model. The model with the residual variances for invariant items constrained to 1 (to be equal in Argentina) fit significantly worse than the model with those residual variances freed, DIFFTEST (12) = 22.35, p =0.0337.

The modification indices suggested that relaxing the constraints on the residual variance for item 13 would improve fit. After doing so, the new fixed partial residual invariance model did not fit significantly worse than the freed residual invariance model DIFFTEST (12) = 16.78, p = 0.1144.

Support for partial residual variance invariance (i.e., “strict invariance”) indicates that the amount of item variance not accounted for by the factor was the same across groups in items 1-4, 7, and 11-12 and 14); the residual variance for items 5, 6, 8,10 and 15 were assumed non-invariant because of lack of threshold invariance found previously, and were not tested.

Summary for Mexico by Argentina

In conclusion, these analyses indicate that partial measurement invariance was obtained across Mexico and Argentina – that is, the relationships of items 1-13 to the latent factors of life satisfaction and general depressive affect were equivalent between the countries of Mexico and Argentina. In addition, for items 1-4, 7, 9 and 11-14 the observed differences in the proportion of responses in each category for these items was due to differences in the factor mean only. Finally, the amount of item variance not accounted for by the factor was the same across groups for items 1-4, 7, and 11-12 and 14.

The lack of full metric and scalar invariance means that the observed values of items 5, 6, 8, 10, 14 and 15 will differ between older adults in Mexico and Argentina at a given level of the latent factor. Invariance hypotheses tests are summarized in Table 31.

.

Uruguay by Argentina

For the Uruguay by Argentina model, global fit indices were adequate and ranged from a CFI of .963 to .976, TLI of .980 to .984 and RMSEA of .040 to .045, these results are summarized in Table 32.

Invariance hypotheses tests

Configural

A configural invariance model was initially specified in which two factors (Life Satisfaction and General Depressive Affect) were estimated simultaneously for the countries of Uruguay and Argentina. The factor variance was fixed to 1 and the factor mean was fixed to 0 in each group for identification, such that all item factor loadings and thresholds (1 per item given a binary response option) were then estimated.

The residual variances are not uniquely identified in the configural invariance model and as such were all constrained to 1 in both groups. As shown in Table 32, the configural invariance model had adequate fit across countries, $\chi^2(123) = 434.28, p < .0000$, RMSEA = .045, CFI = .963, TLI = .980. The analysis proceeded by applying parameter constraints in successive models to examine potential decreases in fit resulting from measurement non-invariance between the countries of Uruguay and Argentina, with Uruguay as the reference group and Argentina as the focal group.

Metric

Equality of the unstandardized item factor loadings between groups was then examined in a metric invariance model. The factor variance was fixed to 1 in Uruguay for identification but was freely estimated in Argentina; the factor mean was fixed to 0 in both groups for identification. All factor loadings were constrained equal across groups, all item thresholds were estimated, and all residual variances were constrained to 1 across groups.

The metric invariance model did result in a decrement in fit, DIFFTEST (12) = 50.16, p = .0000. Modification indices suggested that adding a correlation between the errors for item 3 and item 4 would improve fit. After doing so the partial metric invariance model did not fit significantly worse than the full configural invariance model, DIFFTEST (10) 11.50, p=.3198. Support for partial metric invariance indicates that items 1, 2 and 5-15 were related to the latent factor equivalently across countries (with the exception of items 3 and 4).

Scalar

Equality of the unstandardized item thresholds across groups was then examined in a scalar invariance model. The factor variance and mean were fixed to 1 and 0, respectively, in Uruguay for identification, but the factor variance and mean was then estimated for Argentina. All factor loadings and item thresholds were constrained equal across groups; all residual variances were still constrained equal to 1 in both groups. The full scalar invariance model fit significantly worse than the partial metric invariance model, DIFFTEST (12) = 55.60, $p < .0000$.

Modification indices suggested that the threshold of item 15 was the largest source of misfit and should be freed. After doing so, the partial scalar invariance model still had significantly worse fit than the partial metric invariance model, DIFFTEST (11) = 28.95, $p < .0023$. The modification indices then suggested that the threshold of item 12 was the largest source of misfit and should be freed. After doing so, the new partial scalar invariance model (with the thresholds of items 12 and 15 freed) did not fit significantly worse than the partial metric invariance model, DIFFTEST (10) = 16.38, $p = .0893$.

Support for partial scalar invariance indicates that items 1-11 and 13-14 have the same expected response for each item threshold at the same absolute level of the trait, or more simply, that the observed differences in the proportion of responses in each category for those items was due to the factor mean differences only.

Strict

Equality of the unstandardized residual variances across groups was then examined in a residual variance invariance model. The model comparison at this step proceeded backwards, such that a model with all residual variances freely estimated in the country of Argentina was fitted first, and then compared with a model in which the residual variances for the invariant items (1-11 and 13-14) were fixed to 1 in Argentina. The residual variances in the country of Uruguay were all fixed to 1 for identification in both models, and the rest of the model parameters were estimated as described for the last scalar invariance model. The model with the residual variances for invariant items constrained to 1 (to be equal in Argentina) fit significantly worse than the model with those residual variances freed, DIFFTEST (11) = 20.94, p =0.0340. The modification indices suggested that relaxing the constraints on the residual variance for item 9 would improve fit. After doing so, the new fixed partial residual invariance model did not fit significantly worse than the freed residual invariance model DIFFTEST (11) = 14.92, p = 0.1849. Support for partial residual variance invariance (i.e., "strict invariance") indicates that the amount of item variance not accounted for by the factor was the same across groups in items 1-8, 10, 11, 13 and 14); the residual variance for items 12 and 15 were assumed non-invariant because of lack of threshold invariance found previously, and were not tested.

Summary for Uruguay by Argentina

In conclusion, these analyses indicate that partial measurement invariance was obtained across Uruguay and Argentina – that is, the relationships of items 1, 2 and 5-15 to the latent factors of life satisfaction and general depressive affect were equivalent between the countries of Uruguay and Argentina. In addition, for items 1-11and 13-14 the observed differences in the proportion of responses in each category for these items was due to differences in the factor mean only. Finally, the amount of item variance not accounted for by the factor was the same across groups for items 1-8, 10, 11, 13 and 14.

The lack of full metric and scalar invariance means that the observed values of items 3, 4, 12 and 15 will differ between older adults in Uruguay and Argentina at a given level of the latent factor. Invariance hypotheses tests are summarized in Table 33.

.

Multiple-Group CFA Invariance Summary

The analyses here represent the most robust examination of the measurement properties of the GDS-15 in older adults in Latin America and the Caribbean. The GDS-15's binary response format was appropriately modeled with EFA and CFA models in Mplus. Multiple group binary CFA models were used to test hypotheses metric, scalar and strict invariance across gender and country of origin in older adults.

EFA and CFA analyses found support for a one-dimensional structure of general depressive affect in the countries of Chile and Cuba. While in the countries of Argentina, Mexico and Uruguay a two-factor model defined as life satisfaction and general depressive affect best reflected the data, and replicated previous work by (Brown and Woods, 2007).

Across all countries by gender metric invariance was supported which indicates that the relationships of items to their latent factors were equivalent between men and women. Full measurement invariance by gender was obtained in the countries of Chile and Cuba, which indicates that within these countries the GDS-15 performs equivalently across men and women at all, levels of invariance. In addition, because full metric and scalar invariance was obtained, group comparisons of the latent mean of "general depressive affect" can be conducted. Across the countries of Argentina, Mexico and Uruguay partial scalar invariance was obtained, which means that for items 9 (Argentina), 6 & 8 (Mexico) and 15 (Uruguay) there is "something" other than the factors that is causing the thresholds to differ and that "something" is related to gender. Although full metric invariance was obtained across the countries of Argentina, Mexico

and Uruguay, a lack of full scalar invariance indicates that men and women within these countries will differ in the threshold parameter of the item exhibiting misfit. This means that all predicted observed scores will differ at various levels of the latent factors of life satisfaction and general depressive affect.

There were no country by country comparisons that obtained full measurement invariance. For the country by country comparison of Cuba by Chile, 5 out of 15 items contributed to the lack of scalar invariance (1, 5, 7, 9, and 15), which indicates that comparisons based on these items would be suspect. In the country by country comparison of Mexico and Uruguay 5 out 15 items contributed to a lack of scalar invariance (2, 4, 9, 10 and 15) also indicating that group comparisons which include these items would be suspect.

The comparison of Mexico and Argentina also had 5 items that contributed to lack of scalar invariance (5, 6, 8, 10, and 15) while the comparison of Argentina and Uruguay had 2 items that contributed to lack of scalar invariance (12 and 15). Across the country by country comparisons, item 9 ("do you prefer to stay at home, rather than going out and doing things") and 15 ("do you think that most people are better off than you are") presented most often as having lack of threshold invariance. A direct cause of the invariance in items 9 and 15 is difficult to ascertain because there are countless factors that may influence the responses to these items. However, the wording of these items may tap into constructs that are unrelated to depression; for example with item 9 ("do you prefer to stay at home, rather than going out and doing things") could relate to a lack of security/safety in an individuals' environment, such as living in a high crime

neighborhood, thus endorsement of this item might be a reflection of the community you live in and not you level of depression. Similarly, with item 15 ("do you think that most people are better off than you are") persons might interpret this item as being related to socioeconomic position and not depression.

IRTLR-DIF Analysis

In the current study, IRT likelihood-ratio tests were used to test for invariance of item response parameters across gender and country of origin for the Geriatric Depression Scale Short Form (GDS-15). The IRTLR procedure uses a 2-parameter logistic item response model. This procedure requires that members from each group under study are matched on an estimate of theta (i.e. depression). In order to get a valid estimate of theta, it is best to match the groups using items that are DIF-free. This set of items is referred to as an anchor set and, if there is no prior knowledge that there are scale items that are DIF-free, then an item purification process must be used.

The IRTLR-DIF procedure was used to first identify anchor items and then to test the studied items for DIF relative to these anchor items. The Benjimini-Hochberg procedure was used to control for multiple comparisons in determining statistically significant DIF. MULTILOG software was then used to obtain the final parameter estimates for the groups under study, while modeling the DIF that was significant.

The first step in the analysis was to ensure that the measure under study was sufficiently unidimensional. There is currently no single gold-standard procedure for determining whether a data-set is sufficiently unidimensional for IRT analyses (Reise, 2005). With that said, there are several commonly used approaches for evaluating unidimensionality: (1) examination of screeplots for a dominant first factor, (2) large ratio of 1^{st} to 2^{nd} eigenvalues (3 to 1) and (3) dominant first factor accounts for at least 20% of the variance in the survey items and (4) the factor with the second largest

eigenvalue only explains a small amount of the variance present (Reckase, 1979; Reise, 2005).

Based on the previous MGCFA analysis, screeplots for all five countries indicate that there is a dominant first factor, and the ratio of 1^{st} to 2^{nd} eigenvalues meets the requisite ratio of (3 to 1) (Argentina ratio of 1^{st} (7.53) to 2^{nd} (1.88) = 4.0) (Mexico ratio of 1^{st} (7.87) to 2^{nd} (1.42) = 5.5) (Cuba ratio of 1^{st} (8.57) to 2^{nd} (1.05) = 8.1) (Chile ratio of 1^{st} (8.11) to 2^{nd} (1.01) = 8.0) (Uruguay ratio of 1^{st} (8.06) to 2^{nd} (1.19) = 6.7), so the GDS-15 can be considered sufficiently unidimensional in all countries.

IRTLR-DIF by Gender

The following section presents the results of the IRTLR-DIF analysis within each country by gender and then between countries.

Chile

DIF was assessed within the country of Chile with respect to gender, with women as the reference group and men as the focal group. Using all other items as a tentative anchor, the initial IRTLR-DIF procedure identified 14 DIF-free items using a p-value of .05, and one item was identified as having potential non-uniform DIF with respect to gender. **Item 13** "do you feel full of energy", showed significant non-uniform DIF using IRTLR, prior to, but not after the BH-Adjustment. Although, item 13 lacked significance after the BH-Adjustment, the discrimination and location parameters were estimated freely (not constrained) for men and women in the MULTILOG analysis.

Final 2PL model estimates showed that for the anchor items in Table 34,, items 1-8 and 10-14 discriminated well with a high value of 2.63 for item 8 and a low value of .54 for item 9, as a group they covered a wide range of depression severity (***b-parameters range from -.05 to 1.56***) implying that they were well suited to serve as an anchor set. The right most column of Table 35, lists the chi-square values and associated probabilities for the IRTLR nested model comparison tests of ***a*** and ***b*** DIF for item 13, Figure 6 displays the item characteristic curve for item 13 by gender. The test information function for Chile, suggests that the GDS-15 is best differentiating for individuals between ($-.5 < \Theta < 2$) on the depression continuum and is less differentiating

for individuals on the lower end, evidenced by the large spread of measurement errors for individuals with very low scores on the latent trait, see Figure 7.

Cuba

DIF was assessed within the country of Cuba with respect to gender, with women as the reference group and men as the focal group. Using all other items as tentative anchor, the initial IRTLR-DIF procedure identified 13 DIF-free items, while two items 7 and 12, exhibited potential uniform DIF: **item 7** "do you feel happy most of the time" and **item 12** "do you feel worthless the way you are now", were then retested for DIF relative to the anchor item set. After the second item purification with IRTLR-DIF, items 7 and 12 remained significant for uniform DIF prior to the BH-Adjustment, but not after the BH-Adjustment. Parameters for both items were estimated freely despite their non-BH-Adjustment significance.

Final 2PL model estimates showed that the anchor items in Table 36 discriminated well (i.e. were strongly related to the underlying construct) with a high value of 2.91 for item 3 and low value of .92 for item 9 and as a group the anchor set covered a wide range of depression severity (-.01 to 1.26). The right most column of Table 37, lists chi-square values and associated probabilities for the IRTLR nested model comparison tests of non-uniform and uniform DIF for the two studied items 7 and 12. Figure 8 and Figure 9 display item characteristic curves for items 7 and 12 by gender.

The test information function indicates that this scale better differentiates between individuals who are in the middle range of $-1 < \Theta < 2$ on the latent trait continuum and is less differentiating at either extreme. Figure 10 also shows that there are relatively large measurement errors for individuals with very low scores on the latent trait.

Argentina

DIF was assessed by gender in the country of Argentina with women as the reference group and men as the focal group. The anchor items, for the gender DIF analysis in the country of Argentina were moderately discriminating in a narrow range of difficulty parameters (***a parameters range from 1.13 to 2.94 and b parameters ranged from .44 to 2.00)***; results are summarized in Table 38.

The analysis based on gender in the country of Argentina initially identified 10 DIF-free anchor items and 5 with potential DIF; however, after purification 4 items with DIF were identified, and all but one with uniform DIF. Application of the BH-Adjustment revealed that none of the four items evidenced DIF.

Results, are summarized in Table 39 and Figure 11 through Figure 14 , display the item characteristic curves for items 9, 11, 12 and 15 by gender. The test information function indicates that the scale differentiates best for individuals who are above average on the latent trait, and least for those on the lower end of the continuum as evidenced by the large spread of measurement errors; see Figure 15.

Mexico

DIF was assessed by gender in the country of Mexico with women as the reference group and men as the focal group. The analysis based on gender in the country of Mexico initially identified 11 DIF-free anchor items and 4 with potential DIF. The anchor item set was moderately discriminating with a high value of 2.76 for item 3 and a low value of 1.22 for item 9, in addition there was a wide range of severity (difficulty) parameters, see Table 40.

After item purification there were only two items exhibiting uniform DIF, **item 6** "are you afraid something bad is going to happen to you" and **item 8** "do you often feel helpless". After the BH-Adjustment both items 6 and 8 remained statistically significant for uniform DIF. As shown in Figure 16 and Figure 17, items 6 and 8 were both more severe indicators (difficult to endorse) for men than women, so it would require a higher amount of depression for men to endorse either of these items (**item 6** ***Men: a=1.54, b=.39, Women: a=1.54, b=.00***), (**item 8** ***Men a=3.16, b=.72, Women: a=3.16, b=.00***); results are summarized in Table 40 and Table 41.

In addition, the test information plot shows that the scale differentiates better for individuals who are just above average on the depression continuum and is less differentiating at either extreme as evidenced by the spread of the measurement errors; see Figure 18.

Uruguay

DIF was assessed by gender in the country of Uruguay with women as the reference group and men as the focal group. The analysis based on gender in the country of Uruguay initially identified 10 DIF-free anchor items and 5 with potential DIF; see Table 42. However, after item purification there were only 4 items exhibiting non-uniform DIF, ***item 5*** *"are you in good spirits most of the time",* ***item 6*** *"are you afraid something is going to happen to you",* ***item 12*** *"do you feel worthless the way you are now" and* ***item 14*** *"do you feel that your situation is hopeless".* Items 5,6,12 and 14 evidenced DIF prior to the BH-Adjustment, but after the adjustment only item 5 and 14 continued to exhibit non-uniform DIF, see Table 43.

For **item 5** men had a higher difficulty parameter than women ***(men a=1.76, b=1.07 and women a=2.81, b=1.02)*** indicating that it was more difficult or required slightly more depression for men to endorse this item, while women had a larger a parameter, indicating a stronger relationship with the item and the underlying construct relative to gender, see Table 43.

Item 5 is one of five negatively coded items (1, 5, 7, 11, 13) on the GDS-15; what this means is that all but five items on the GDS-15 are coded as 0=no and 1=yes; for the other five items the coding is 0=yes and 1=no. So in the case of item 5, if an individual endorses this item with the GDS-15 coding scheme, the interpretation of item 5 is the following: ***men find it harder to endorse "I am not in good spirits most of the time" relative to women, however based on the b parameter for women, they also find it difficult to endorse this item,*** see Table 43 and Figure 19.

For **item 14**, men also have a harder time endorsing "*do you feel that your situation is hopeless"* relative to women ***(men a=2.34, b=1.17, women a=2.90, b=1.14)***, see Table 43 and Figure 22. The test information plot, indicates that for men and women in Uruguay, the GDS-15 does a better job of differentiating between individuals in the middle to the upper end of the continuum $0 < \Theta < 2$; in addition, measurement errors were larger for individuals on the lower end of the continuum, see Figure 23.

With respect to the anchor item set used, all of the items provided modest discrimination with the highest value being 3.33 for item 7 and the lowest value being .89 for item 9. Difficulty parameters ranged from -.16 to 1.93, but tended to cluster between .72 and 1.05, see Figure 42.

A summary of the aforementioned DIF analyses of the GDS-15 by gender is provided in Table 44 to Table 46.

IRTLR-DIF by Cross Country Comparisons

Chile by Cuba

DIF was assessed between the countries of Chile and Cuba, with Chile as the reference group and Cuba as the focal group. Using all other items as a tentative anchor, the initial IRTLR-DIF procedure identified 7 DIF-free anchor items see Table 47.

Two out of 8 items originally identified with DIF evidenced DIF after item purification, but before the BH-Adjustment. After the BH-Adjustment, both of items showed DIF: **item 4** "do you often get bored" (uniform) and **item 8** "do you often feel helpless" (uniform), see Table 48. For both indicators older adults in the country of Chile found it more difficult to endorse these items ***(item 4 Chile: a=3.19, b=.08 Cuba: a=3.19, b=.00) (item 8 Chile: a=2.83, b=.56, Cuba: a=2.83, b=.00).*** In other words, the location parameter is lower for older adults in Chile, thus the DIF in this item indicates that given the same level of overall depression, it takes more depression for Chileans to endorse these items, see Figure 24 and Table 48.

The test information plot showed that across the countries of Chile and Cuba the GDS-15 is better at differentiating older adults between $-1 < \Theta < 2$ on the depression continuum and less differentiating for individuals with lower scores on the latent trait, evidenced by the larger spread of measurement errors; see Figure 26.

The final anchor item set of 13 indicators, were moderately discriminating (strong relationship to the underlying construct), with a high value of 2.89 for item 3 and a low

value of .83 for item 9 having the lowest value; in addition the location parameters clustered in a narrow range -.82 to 1.22; see Table 47.

Mexico by Uruguay

DIF was assessed between the countries of Mexico and Uruguay, with Uruguay as the reference group and Mexico as the focal group. Using all other items as a tentative anchor, the initial IRTLR-DIF procedure identified 6 DIF-free anchor items, see Table 49. Five out of 9 items originally identified with DIF evidenced DIF after item purification, but before the BH-Adjustment. After the BH-Adjustment 4 of the 5 items showed DIF, all uniform DIF: **item 6** "are you afraid something bad is going to happen to you, **item 7** "do you feel happy most of the time, **item 9** "do you prefer to stay at home, rather than going out and doing things and **item 12** "do you feel worthless the way you are now"; see Table 50.

All DIF items were highly discriminating relative to item 9. For **items 6 and 12** older adults from Uruguay had to have more depression to endorse these items than older adults from Mexico ***(item 6: Uruguay a=1.66, b=.29, Mexico a=1.66, b=.00; item 12: Uruguay a=3.11, b=.66, Mexico a=3.11, b=.00)***. Alternatively, given the same level of overall depression, DIF in these items (6 and 12) indicates that older adults in Mexico find it easier to endorse these items. For **item 7**, the interpretation of the question is based on the reverse coding scheme for items (1, 5, 7, 11 and 13), with that said, item 7 should be interpreted as "***I do not feel happy most of the time***". Consequently, older adults from Uruguay find it more difficult to endorse this item ***(item 7: Uruguay a=3.39, b=.21, Mexico a=3.39, b=.00)***. Older adults from Mexico found it more difficult to endorse

item 9, while persons from Uruguay found it easier to endorse ***(item 9 Uruguay a=.85, b= -.70, Mexico a=.85, b=.00)***; see Figure 27 through Figure 30 and Table 50.

Examination of the test information plot for the GDS-15 across the countries of Mexico and Uruguay, shows that the instrument differentiates best for older adults in the middle of the distribution and not at either extreme $-1.50 < \Theta < 1.75$; see Figure 31. In addition the measurement errors for individuals scoring lower on the continuum are large, indicating poor differentiation. With respect to the anchor item set, all ten items were moderately discriminating with a high value of 3.21 for item 8 and a low value of 1.27 for items 10 and 15; in addition the location parameters clustered around the upper range of the distribution .10 to 1.03; see Table 49.

Mexico by Argentina

DIF was assessed between the countries of Mexico and Argentina, with Argentina as the reference group and Mexico as the focal group. Using all other items as a tentative anchor, the initial IRTLR-DIF procedure identified 6 DIF-free anchor items; see Table 51. Five out of 9 items originally identified with DIF evidenced DIF after item purification, but before the BH-Adjustment. After the BH-Adjustment all five items showed DIF, two with non-uniform and three with uniform: **item 2** "do you feel that your life is empty" (uniform DIF), **item 7** "do you feel happy most of the time (non-uniform DIF), **item 9** "do you prefer to stay at home, rather than going out and doing things (uniform DIF), **item 11** "do you think it is wonderful to be alive now" (non-uniform DIF) and **item 12** "do you feel worthless the way you are now (uniform DIF); see Table 52.

For older Mexicans with average standing on the latent trait, it was easier to endorse **items 2 and 12** ***(item 2 Mexico a=1.88, b=.00, Argentina a=1.88, b=.29; item 12 Mexico a=2.54, b=.00, Argentina a=2.54, b=.69)***, while **item 9** was easier for persons from Argentina to endorse ***(item 9 Mexico a=.98, b=.00, Argentina a=.98, b=-46)***; see Table 52. **Item 7** ***"do you Not feel happy most of the time" (reverse scored and interpreted)***, had higher discrimination (stronger relationship to the construct) but a smaller location parameter for older adults in Argentina. In Mexico, item 7 had smaller discrimination but a higher location parameter than Argentina. Thus, older adults in Mexico find it more difficult to endorse item 7 than persons from Argentina ***(Mexico a=2.27, b=.72, Argentina a=2.60, b=33)***; see Table 52. .

For **item 11** "do you think it is wonderful to be alive now", persons from Argentina had higher discrimination but lower b-parameters relative to persons from Mexico. Older adults from Mexico had lower discrimination but a larger location parameter relative to persons from Argentina ***(Argentina a=1.83, b=1.01, Mexico a=1.78, b=1.62)***. **Item 11** is interpreted under the same coding scheme as item 7, so the interpretation is actually ***"do you <u>not</u> think it is wonderful to be alive now"***, and persons from Mexico find it more difficult to endorse this item than people from Argentina; see Table 52 and Figure 32 to Figure 36.

Examination of the test information plot for the GDS-15 across the countries of Mexico and Argentina, shows that the instrument differentiates best for older adults between $-1 < \Theta < 1.75$ and not for persons on the lower end of the continuum, as evidenced by large measurement errors, see Figure 37.

With respect to the anchor item set, all ten items were strongly related to the underlying construct, with a high value of 3.01 for item 8 and a low value of 1.16 for item 15; in addition the location parameters were widely spread across the continuum .16 to .93; see Table 51.

Argentina by Uruguay

DIF was assessed between the countries of Argentina and Uruguay, with Argentina as the reference group and Uruguay as the focal group. Using all other items as a tentative anchor, the initial IRTLR-DIF procedure identified 8 DIF-free anchor items; see Table 53. Four out of 7 items originally identified with DIF evidenced DIF after item purification, but before the BH-Adjustment. After the BH-Adjustment four items showed DIF: **item 2** "do you feel that your life is empty" (uniform), **item 6** "are you afraid something bad is going to happen to you" (non-uniform), **item 7** "do you *not* feel happy most of the time" (non-uniform) and **item 11** "do you *not* think it is wonderful to be alive now" (non-uniform); see Table 54.

For ***item 2,*** older adults from Uruguay with average standing on the latent trait depression find it easier to endorse this item relative to persons from Argentina ***(Uruguay a=1.67, b=.00 Argentina a=1.67, b=.70)***. For ***item 6***, the discrimination parameter was larger in the country of Uruguay, indicating a stronger relationship to the construct than in the country of Argentina, but a smaller location parameter. While in Argentina the discrimination parameter is smaller and the location parameter is larger, indicating that persons from Argentina require larger amounts of theta/depression to endorse this item ***(Argentina: a=1.07, b=1.23, Uruguay: a=1.54, b=.77). Item 7*** "do you *not* feel happy most of the time" is highly discriminating especially in the country of Uruguay ***(a=3.37, b=.66)***, but the location parameter is smaller relative to Argentina ***(a=2.66, b=.68)***, indicating that it is easier for older adults from Uruguay to endorse ***item 7***. Finally, ***item 11*** "do you *not* think it is wonderful to be alive now" is more discriminating in the

country of Uruguay ***(a=2.45, b=1.28)*** than Argentina ***(a=1.86, b=1.36),*** but has a smaller location parameter which indicates that endorsement of ***item 11*** was easier for older adults from Uruguay, see Figure 38 to Figure 41 and Table 54..

The test information plot for the countries of Uruguay and Argentina differentiates best for individuals higher on the latent continuum and less so for individuals lower on the latent trait continuum, evidenced by large measurement errors; see Figure 42.

The anchor item set used had large discrimination parameters, indicating a strong relationship between the items and the underlying construct, with item 9 and 15 having the smallest discrimination values. The location parameters clustered around a narrow range -.20 to 1.94; see Table 53.

Argentina by Chile

DIF was assessed between the countries of Argentina and Chile, with Argentina as the reference group and Chile as the focal group. Using all other items as a tentative anchor, the initial IRTLR-DIF procedure identified 7 DIF-free anchor items, see Table 55. Two out of 8 items (item 2 and item 7) originally identified with DIF, evidenced DIF after item purification, but before the BH-Adjustment. After the BH-Adjustment only one item showed DIF: ***item 2*** "do you feel that your life is empty" (uniform), see Table 56.

Item 2 was easier to endorse for older adults from Chile ***(a=1.39, b=.00)*** relative to Argentina ***(a=1.39, b=.09)***, evidenced by the item characteristic curves for this item, see Figure 43. The item characteristic curve for item 7 is presented in Figure 44; item 7 did not have significant DIF after the BH- Adjustment. Across the countries of Argentina and Chile, the GDS-15 best differentiates for individuals above average on the depression continuum, and less so for individuals on the lower end of the continuum, as evidenced by large measurement errors; see Figure 45.

The anchor item set used for this analysis, had discrimination parameters which were strongly related to the underlying construct, with a high value of 2.66 for item 8 and a low value of .83 for item 9. Location parameters also reflect a wide range of values -.04 to 1.45; see Table 55.

Argentina by Cuba

DIF was assessed between the countries of Argentina and Cuba, with Argentina as the reference group and Cuba as the focal group. Using all other items as a tentative anchor, the initial IRTLR-DIF procedure identified 9 DIF-free anchor items; see Table 57. Three out of 6 items originally identified with DIF, evidenced DIF after item purification, but before the BH-Adjustment. After the BH-Adjustment all three items showed DIF: **item 2** "have you dropped many of your activities and interests" (uniform), **item 5** "are you *not* in good spirits most of the time" (non-uniform) and **item 7** "do you *not* feel happy most of the time" (non-uniform); see Table 58.

Item 2 was easier for older adults from Cuba to endorse ***(a=1.66, b=.00)*** than older adults from Argentina ***(a=1.66, b=.65)***. **Item 5** "are you *not* in good spirits most of the time", was more discriminating for older adults from Cuba, but also had a smaller location parameter indicating that it was easier for Cubans to endorse **item 5** than Argentineans ***(Cuba: a=2.47, b=.75; Argentina a=2.09, b=.89)***. Finally, for **item 7** "do you *not* feel happy most of the time", Cubans had larger discrimination parameters and smaller location parameters indicating that the item was easier to endorse than for older adults from Argentina ***(Cuba: a= 2.90, b=.49; Argentina a=2.63, b=.62);*** see Figure 46 to Figure 48 and Table 58. The test information plot for the GDS-15 across the countries of Cuba and Argentina indicates that the instrument differentiates best for persons above average on the latent trait continuum and less so, for persons on the lower end, evidenced by larger measurement errors; see Figure 49.

With respect to the anchor item set, all items had modest discrimination values, with a high value of 2.80 (item 8) and a low value of .99 (item 9) and as group the location parameters covered a wide range -.35 to 1.64; see Table 57.

Mexico by Chile

DIF was assessed between the countries of Mexico and Chile, with Chile as the reference group and Mexico as the focal group. Using all other items as a tentative anchor, the initial IRTLR-DIF identified 9 DIF-free anchor items see Table 59. Five out of 6 items originally identified with DIF, evidenced DIF after item purification, but before the BH-Adjustment. After the BH-Adjustment only 3 of the 5 items showed DIF: **item 7** "do you *not* feel happy most of the time" (uniform), **item 8** "do you often feel helpless" (non-uniform) and **item 12** "do you feel worthless the way you are now (uniform); see Table 60.

Older adults from Mexico found both items 7 and 12 easier to endorse ***(item 7 Mexico: a=3.15, b=.00, Chile: a=3.15, b=.40), (item 12 Mexico: a=3.09, b=.00, Chile a=3.09, b=.51)*** than older adults from Chile; see Figure 50, Figure 53 and Table 60. For item 8, Chileans had smaller discrimination parameters but larger location parameters ***(item 8 Chile a=2.97, b=.44)***, indicating that this item was difficult to endorse relative to Mexicans. Mexicans had higher discrimination parameters for item 8 but smaller location parameters, indicating that this item was easier to endorse ***(item 8 Mexico: a=3.46, b=.14***), see Figure 51 and Table 60.

The test information plot for the GDS-15 across the countries of Cuba and Chile indicates that the instrument differentiates best for individuals in the middle of the distribution but not at either extreme see Figure 55. With respect to the anchor item set, all items had modest discrimination with a high value of 2.98 for item 3 and a low value of 1.01 for item 9; as a group the location parameters were clustered in a narrow range -.69 to .67; see Table 59.

Mexico by Cuba

DIF was assessed between the countries of Mexico and Cuba, with Mexico as the reference group and Cuba as the focal group. Using all other items as a tentative anchor, the initial IRTLR-DIF identified 8 DIF-free anchor items; see Table 61. Six out of 7 items originally identified with DIF, evidenced DIF after item purification, but before the BH-Adjustment. After the BH-Adjustment only 5 of the 6 items showed DIF: **item 2** "have you dropped many of your activities and interests" (uniform), **item 4** "do you often get bored" (uniform), **item** 7 "do you *not* feel happy most of the time" (uniform), **item 12** "do you feel worthless the way you are now" (uniform) and **item 14** "do you feel that your situation is hopeless" (uniform); see Table 62.

Older adults in Cuba found items 2, 4, 7, 12 and 14 easier to endorse than older adults from Mexico ***(item 2 Cuba a=1.91, b=.00, Mexico a=1.91, b=.10), (item 4 Cuba a=2.98, b=.00, Mexico a=2.98, b=.03), (item 7 Cuba a=2.95, b=.00, Mexico a=2.95, b=.39), (item 12 Cuba a=2.99, b=.00, Mexico a=2.99, b=.35)*** and ***(item 14 Cuba a=3.15, b=.00, Mexico a=3.15, b=.23)***; see Figure 56 to Figure 61 and Table 62.

The test information plot for the GDS-15 across the countries of Mexico and Cuba indicates that the instrument differentiates best for individuals in the middle of the distribution but not at either extreme, see Figure 62.

With respect to the anchor item set, all items had modest discrimination with a high value of 3.21 for item 8 and a low value of 1.16 for item 9, as a group the location parameters were clustered in a narrow range -.68 to .64; see Table 62.

Uruguay by Chile

DIF was assessed between the countries of Uruguay and Chile, with Chile as the reference group and Uruguay as the focal group. Using all other items as a tentative anchor, the initial IRTLR-DIF identified 7 DIF-free anchor items; see Table 63. Four out of 8 items originally identified with DIF evidenced DIF after item purification, but before the BH-Adjustment. After the BH-Adjustment, only 3 of the 4 items showed DIF: **item 2** "have you dropped many of your activities and interests" (non-uniform), **item 10** "do you feel that you have more problems with memory than most" (non-uniform) and **item 14** "do you feel that your situation is hopeless" (uniform), see Table 64.

Item 2 had a higher discrimination parameter for older adults from Uruguay and a smaller location parameter indicating that this item was easier to endorse, while older adults from Chile had lower discrimination parameters but higher location parameters indicating more difficulty in endorsing this item ***(Uruguay a=1.71, b=.30, Chile a=1.27, b=.36)***; see Table 64 and Figure 63. **Item 10** was easier to endorse for Chileans than persons from Uruguay, in addition item 10 more discriminating in Chile than Uruguay ***(Chile a=1.39, b=.92, Uruguay a=1.15, b=1.45)*** see Table 64 and Figure 65. Finally, for **item 14** older adults from Uruguay found this item easier to endorse than Chileans ***(Uruguay a=2.61, b=.00, Chile a=2.61, b=.57)***, see Figure 66 and Table 64.

The test information plot for the GDS-15 across the countries of Uruguay and Chile indicates that the instrument differentiates best for individuals above average on the latent trait distribution but not at the lower end see Figure 67.

With respect to the anchor item set, all items had modest discrimination with a high value of 3.05 for item 7 and a low value of 1.02 for item 15; as a group the location parameters were clustered in a narrow range of -.73 to .80; see Table 63.

Uruguay by Cuba

DIF was assessed between the countries of Uruguay and Cuba, with Uruguay as the reference group and Cuba as the focal group. Using all other items as a tentative anchor, the initial IRTLR-DIF identified 9 DIF-free anchor items see Table 65. Three out of 6 items originally identified with DIF evidenced DIF after item purification, but before the BH-Adjustment. After the BH-Adjustment all three items showed DIF: item 4 "do you often get bored" (uniform), **item 6** "are you afraid something bad is going to happen to you" (non-uniform) and **item 8** "do you often feel helpless" (non-uniform), see Table 66. **Item 4** was easier for older adults from Cuba to endorse than Uruguay ***(Cuba a=2.51, b=.00, Uruguay a=2.51, b=.39).*** For **item 6,** older adults form Uruguay had higher discrimination parameters but lower location parameters indicating that this item was easier to endorse relative to Cubans ***(item 6 Uruguay a=1.54, b=.65, Cuba a=1.52, b=.67)***. Finally for **item 8,** older adults from Uruguay had higher discrimination parameters but lower location parameters indicating that this item was easiest to endorse relative to Cubans ***(item 8 Uruguay a=2.96, b=.72, Cuba a=2.77, b=.90);*** see Figure 68 to Figure 70 and Table 66.

The test information plot for the GDS-15 across the countries of Uruguay and Cuba indicates that the instrument differentiates best for individuals above average on the latent trait distribution but not at the lower end; see Figure 71. With respect to the anchor item set, all items had modest discrimination with a high value of 3.08 for item 7 and a low value of .92 for item 9; as a group the location parameters were widely spread -.44 to

.1.52, see Table 65. A summary of the aforementioned DIF analyses of the GDS-15 by country of origin is provided in Table 67 through Table 72.

IRTLR-DIF Summary

DIF analyses by gender showed that men had more difficulty endorsing item 5 "are you not in good spirits most of the time", item 6 "are you afraid that something bad is going to happen to you", item 8 "do you often feel helpless" and item 14 "do you feel that your situation is hopeless", than women across countries. With the exception of Mexico, item 9 "do you prefer to stay at home, rather than going out and doing things", consistently had the lowest discrimination values across Chile, Cuba, Argentina and Uruguay. This indicates that this item consistently provided lower levels of information and if the GDS-15 were to be revised or shortened, item 9 would be a candidate item to be removed from the scale. In addition, test information plots across countries indicate a need for items that can differentiate between individuals on the lower end of the continuum.

DIF analyses with country to country comparisons showed that item 2 "have you dropped many of your activities and interests", item 4 "do you often get bored", item 6 "are you afraid something bad is going to happen to you", item 7 "do you not feel happy most of the time" and item 12 "do you feel worthless the way you are now", were consistently more difficult for older adults from Mexico to endorse. In addition, older adults from Argentina found it difficult to endorse item 2 "have you dropped many of your activities and interests", item 6 "are you afraid something bad is going to happen to you", item 7 "do you not feel happy most of the time", item 9 "do you prefer to stay at home, rather than going and doing things and item 11 "do you not think it is wonderful to be alive now", across several country comparisons.

Like the gender DIF analyses, item 9 tended to have the lowest discrimination values across country to country comparisons, which means this item provided little information, and were the scale to be revised, consideration should be given to its removal. In addition just like in the gender DIF analyses, the test information plots for the country to country comparisons indicate overall, that the GDS-15 is best at differentiating between individuals who are high on the continuum of depression, but could benefit from items that can differentiate between individuals across the complete continuum of depression.

Full invariance on discrimination and difficulty parameters were not obtained within any country. With that said, full invariance is generally not a realistic expectation. Instead, these results can be seen as a means of understanding how the scale behaves at the item and scale level across gender and country of origin.

At the item level, the results provide researchers with information on how items are related to the underlying construct via high or low *a-parameter* estimates, and over what levels of the latent trait the item best discriminates among groups. This DIF analysis provides insight into what items on the GDS-15 might be problematic based on gender or country of origin; in addition these results can be used to help in the interpretation of the depressive etiology underlying group differences in response patterns to particular items. At the scale level, end users such as clinicians, nurses or social workers, can gain insight into the measurement precision of the scale over the levels of the depressive continuum. Because the GDS-15 was designed to identify the presence of depression, i.e. individuals with high levels of depression, the test information curves for

the scale is expected to function in the upper level of Θ. This pattern was observed across gender and cross-comparison analyses.

Finally, because of the relationship between physical and mental health evidenced by the work of (Ayotte, Yang, & Jones, 2010; A. T. F. Beekman, Kriegsman, Deeg, & Tilburg, 1995; A.T.F. Beekman, et al., 1997; Berkman, et al., 1986; Braam, et al., 2005; Ormel, et al., 1997) and others, future DIF work on the GDS-15 should assess item and test functioning in relationship to chronic disease and functional disability associated with aging. Finding out which items differentiate best among older adults with functional disabilities and chronic illnesses can inform the development of group specific depression interventions.

CHAPTER 5: DISCUSSION

This chapter presents a summary of the major findings of the current study, an evaluation of practical implications of these findings as well as suggestions for future research. Limitations of the current study are also discussed.

GENERAL OVERVIEW

Many disciplines such as psychology, sociology, education and epidemiology use group comparisons to investigate differences in psychological phenomena, social systems, learning styles and chronic disease. These comparisons explicitly or not are predicated on the assumption that the measures used to evaluate differences are the same conceptually across groups of interest. Unfortunately measurement invariance is often assumed and not tested. This can lead to spurious conclusions, which can ultimately affect educational outcomes, health outcomes or policy decisions. Testing the assumption of measurement invariance provides support for the suitability of an instrument for different populations. The assessment of measurement invariance can improve the quality of research in many disciplines. The need for making sure that an instrument is functioning in the same way for different groups was the driving force behind the current study.

The current study used two commonly used frameworks for assessing measurement invariance (1) structural equation modeling (SEM) and (2) item response theory (IRT). The two approaches used to test the measurement invariance of the GDS-15 within the countries of Chile, Cuba, Argentina, Mexico and Uruguay were multiple

group confirmatory factor analysis (MGCFA) and item response theory likelihood ratio tests (IRTLR-DIF). The multiple group approach was used to evaluate the invariance of factor loadings, thresholds and uniqueness's and the IRTLR-DIF approach was used to assess the invariance of discrimination and difficulty parameters.

Major Findings Multiple Group Confirmatory Factor Analyses

Gender

Previous psychometric research on the GDS-15 was limited and provided a lack of consensus on the underlying structure of the instrument (P. J. Brown, et al., 2007; Friedman, et al., 2005; Incalzi, et al., 2003; D. W. L. Lai, et al., 2005; Mitchell, et al., 1993; Onishi, et al., 2006; Schreiner, et al., 2001). As such, the current study first evaluated the underlying structure of the instrument within each country.

Exploratory factor analyses and CFA's within each country showed that a one-factor model provided the best fit in the countries of Chile and Cuba and this factor was defined as general depressive affect. Within the countries of Argentina, Mexico and Uruguay, a two-factor model fit the data best and these factors were defined as general depressive affect and life satisfaction. These two factors replicate the work of (P. J. Brown, et al., 2007) and (Mitchell, et al., 1993).

As discussed in chapters 2 and 3, in order to establish measurement invariance a nested sequence of increasingly restrictive CFA models (invariance hypotheses) are tested. These levels of invariance are referred to as configural, metric, scalar and strict invariance. The sequence of nested invariance hypotheses are well established in the literature (Cheung & Rensvold, 2002; Reise, et al., 1993; Robert J. Vandenberg &

Charles E. Lance, 2000). They are based on establishing a baseline/configural model and additively testing hypotheses of metric, scalar, and strict invariance.

MGCFA models of the GDS-15 did not support full measurement invariance of all parameters for men and women across countries. A summary of results from invariance testing is presented in Table 16 through Table 24. There was adequate fit by gender across countries for the baseline/configural models. Within the countries of Chile and Cuba a one-factor model fit the data best and full measurement invariance was obtained. These results provide support for the measurement equivalence of the GDS-15 by gender in the countries of Chile and Cuba, which means that group comparisons of the latent factor mean can be supported.

Across the countries of Argentina, Mexico and Uruguay by gender, a two-factor model fit the data best; however, support for full measurement invariance by gender was not supported. Within each of the three countries, metric invariance held, indicating that that there are similar interpretations of the factors across men and women within the countries of Argentina, Mexico and Uruguay. For the countries of Argentina, Mexico and Uruguay, partial scalar invariance was obtained. In Argentina the threshold for ***item 9 "do you prefer to stay at home rather than going out and doing new things"*** was non-invariant, in Mexico the thresholds for ***items 6 "are you afraid that something bad is going to happen to you" and Item 8 "do you often feel helpless"*** were non-invariant and in Uruguay the threshold for ***item 15 "do you think that most people are better off than you are"*** was non-invariant.

Why are these particular items non-invariant?

Substantive reasons for non-invariance are difficult to ascertain and speculative at best. In the MGCFA gender analysis items 9, 6 and 8, and 15 displayed non-invariance in Argentina, Mexico and Uruguay respectively. An examination of the item content suggests that these items speak to a sense of vulnerability with phrases and words like "being afraid", "helpless", "prefer to stay at home" and "people are better off than you". Cultures have their own gender roles/expectations, in cultures with rigid gender roles that do not allow or support men being "vulnerable" or women being "independent", questions with content that go against these cultural norms may be difficult for individuals to endorse.

Work by Djernes, Zunzunegui and colleagues (Djernes, 2006; Zunzunegui, Alvarado, Beland, & Vissandjee, 2009) have found that older women in Latin America and the Caribbean had a 63% higher odds for depressive symptoms compared with older men and one of the main predictors of depressive disorders and depressive symptom cases is gender. In Latin America women generally have lower levels of education than men and they are not encouraged to be socially or economically independent, which leads to economic disadvantages in later life; the sociocultural context that creates financial stress in later life may also be a predictor of depression (Zunzunegui, et al., 2009). In addition, the content of these items could also be tapping into constructs that are unrelated to depression; for example with item 9 ("do you prefer to stay at home, rather than going out and doing things") could relate to living in an unsafe community which would influence whether a person felt safe enough to go out by themselves. Thus,

endorsement of this item might be a reflection of the community you live in and not your level of depression.

Cross-country comparisons

Country by country comparisons were based on having the same configural pattern of loadings. With that said, country by country comparisons were the following: *Chile by Cuba* (1-factor model general depressive affect), *Mexico by Uruguay* (2-factor model general depressive affect and life satisfaction), *Mexico by Argentina* (2-factor model general depressive affect and life satisfaction) and *Uruguay by Argentina* (2-factor model general depressive affect and life satisfaction).

Full measurement invariance was not obtained in the Chile by Cuba comparison. *Partial metric invariance* was obtained for the cross-country comparison of *Chile and Cuba*. To obtain *partial metric invariance* the errors of ***item 5 "are you in good spirits most of the time"*** and ***item 7 "do you feel happy most of the time"*** needed to be correlated. This indicates that there is some additional shared multidimensionality between these two items. *Partial scalar invariance* was obtained by freeing the thresholds for ***items 1 "are you basically satisfied with your life", item 5 "are you in good spirits most of the time"*** and ***item 7 "do you feel happy most of the time"*** and ***item 15* "do you think that most people are better off than you are".** The cross-country comparison of *Mexico by Uruguay* did not obtain *full measurement invariance*.

To obtain *partial metric invariance* the errors of ***items 1 "are you basically satisfied with your life", item 3 "do you feel that your life is empty", item 10 "do feel***

you have more problems with memory than most" and ***item 15* "do you think that most people are better off than you are"** needed to be correlated; this indicates that there is additional multidimensionality amongst these items.

Partial scalar invariance was obtained by relaxing the threshold constraints of ***items 2 "have you dropped many of your activities and interests"***; ***item 4 "do you often get bored", item 9 "do you prefer to stay at home rather than going out and doing new things", item 10 "do you feel you have more problems with memory than most"*** and ***item 15* "do you think that most people are better off than you are"**.

The cross-country comparison of <u>*Mexico by Argentina*</u> obtained *partial metric invariance* by correlating the errors of ***item 14 " do you feel that your situation is hopeless"*** and ***item 15* "do you think that most people are better off than you are"**. Partial scalar invariance was obtained by relaxing the threshold constraints of ***items 5 "are you in good spirits most of the time", item 6 "are you afraid that something bad is going to happen to you", item 8 "do you often feel helpless", item 10 "do you feel you have more problems with memory than most" and item 15 "do you think that most people are better off than you are".***

Finally the cross-country comparison of <u>*Uruguay by Argentina*</u> did not obtain *full measurement invariance.* Partial metric invariance was obtained by correlating the errors of ***items 3 "do you feel that your life is empty***" and ***4 "do you often get bored"***. *Partial Scalar invariance* was obtained by relaxing the threshold constraints of ***items 12 "do you feel pretty worthless the way you are now"*** and ***15 "do you think that most people are better off than you are".***

Comparing the results of the gender MGCFA and the cross-group comparison MGCFA revealed that item 9 "do you prefer to stay at home rather than going out" and item 15 "do you think that most people are better off than you displayed non-invariance across both testing situations. It is difficult to interpret why these two items exhibit misfit within countries by gender and with cross-group comparisons, they don't share similar wording and the content of the items are getting at two different things (1) withdrawal and (2) self-esteem.

Finally, the results of the cross-country invariance testing between Mexico, Argentina, Uruguay, Chile and Cuba reflect an important point made by Byrne and Watkins (2003), which is that ***"although the factorial structure of a measuring instrument may yield a similar pattern when tested within each of two or more groups, such findings represent no guarantee that the instrument will operate equivalently across these groups"***(pg. 156, pg. 556) (Byrne & Campbell, 1999; Byrne & Watkins, 2003). This statement can be expanded to also include the idea that although an instrument such as the GDS-15 may be administered in the same language, in this case Spanish, there can be linguistic and cultural nuances that contribute to a lack of invariance.

Non-invariance in the cross-country comparisons and gender comparisons is most likely due to translation errors. The use of the GDS-15 in the SABE study, involved no back-translation protocol, which means that the English version of the GDS-15 would have had to have been translated into Spanish and then back into English in order to assess whether the content of the back translation matched the original instrument

conceptually. These five countries share a common language but their cultures may be very different; based on history, economics and ethnic make-up, all of which can influence the interpretation and meaningfulness of constructs and items.

Major Findings Item Response Theory Likelihood Ratio Tests-DIF

The second framework used in the assessment of measurement invariance was IRTLR-DIF. Item response theory likelihood ratio tests were used to assess the invariance of discrimination and difficulty parameters across gender and cross-country comparisons. In other words the primary research question to be answered in this DIF analysis is *"how is item response related to level of depression and subgroup membership"* (i.e. gender or country of origin). The GDS-15 was put through an item purification process with the goal of identifying a core group of DIF free items (anchor items). After the anchor items were identified, a 2PL model which accounted for items with DIF was estimated for groups of interest simultaneously.

The first step in the analysis was to evaluate whether the assumption of unidimensionality had been met. For the GDS-15 exploratory factor analyses in each of the five countries found a dominant first factor which accounted for the majority of the covariance of item responses, which was satisfactory for meeting the assumption of essential unidimensionality. This indicated that the item content within the dominant factor reflected the scale content.

The two-parameter logistic IRT model was used to model item responses for the GDS-15. The 2PL model allows for variation in both the discrimination and difficulty parameters among items on a scale. Items on a scale are evaluated by examining item characteristic, information and standard error of measurement curves. Examination of the item characteristic and information curves, allows a researcher to identify low and

high discriminating items as well as identify the range over the underlying trait in which an item is most effective.

The test information curve and standard error of measurement curve provide a picture of the measurement precision of a scale across the latent trait continuum. The GDS-15 is a depression screening instrument designed to identify the presence of depression over the past week. As such the test information curves are expected to function in the upper end of the theta distribution. This pattern was observed across gender and cross-country comparisons in this study.

An item is said to exhibit DIF if two individuals from distinct groups have the same standing on the underlying trait, but have different probabilities of endorsing an item or getting an item correct. For example a Cuban man and Cuban woman with the same level of depressive symptomatology (theta) should have the same chance of responding "yes" to the item *"I believe it is wonderful to be alive now"*. If this is not the case, the item is said to be exhibiting DIF. What this means is that responses to an item with DIF are not equivalent across the groups under study, which can lead to potentially misleading group comparisons.

An item purification procedure was employed in this study to identify DIF-free items or anchor items as well as items exhibiting DIF after the purification procedure and the BH-Adjustment. After item parameters with significant DIF have been identified, parameters for a final 2-group model that incorporates the identified DIF can be specified and estimated using MULTILOG.

Gender

Depression items that showed DIF with respect to gender, even after the BH-Adjustment indicate that the instrument is performing differently based on gender. In the gender DIF analysis items 5, 6, 8 and 14 were classified as having significant DIF most often after the BH-Adjustment. In the IRTLR-DIF analysis Items 6 and 8 were both more severe indicators (difficult to endorse) for men than women, so it would require a higher amount of depression for men to endorse either of these items (**item 6** ***Men: a=1.54, b=.39, Women: a=1.54, b=.00***), (**item 8** ***Men a=3.16, b=.72, Women: a=3.16, b=.00***).

For **item 5** men had a higher difficulty parameter than women ***(men a=1.76, b=1.07 and women a=2.81, b=1.02)*** indicating that it was more difficult or required slightly more depression for men to endorse this item, while women had a larger a parameter, indicating a stronger relationship with the item and the underlying construct relative to gender. For **item 14**, men also find it more difficult to endorse "*do you feel that your situation is hopeless"* relative to women ***(men a=2.34, b=1.17, women a=2.90, b=1.14)***. Of note is that item 6 and item 8 were also classified as having gender non-invariance in the MGCFA analysis. Results of the gender DIF analysis indicate that just like the MGCFA analysis, items with content that could be interpreted as indicating vulnerability is also more difficult for men to endorse. Research on the relationship between gender and depression has found that even when men and women have similar levels of depression; women tend to report having more depressive symptoms (Angst, 1992).

Cross-country comparison

The two cross-country comparisons with the largest number of items exhibiting DIF after the BH-Adjustment were Argentina by Mexico with five items (2,7,9,11,12) and Mexico by Cuba with five items (2,4,7,12,14). There was no meaningful pattern of DIF found across the ten cross-group comparisons, for example in Chile by Cuba items 4 and 8 were significant, with Mexico by Uruguay items 6, 7, 9 and 12 were significant and in Argentina by Uruguay items 2, 6, 7 and 11 were significant. What is evident is that DIF is context specific, in other words just because item 2 is significant for DIF in the Uruguay by Argentina comparison this does not mean that it will be significant in the Mexico by Uruguay comparison (which it was not).

Overall the IRTLR-DIF procedure identified the same items as exhibiting DIF as those in the MGCFA analysis; however, they were not necessarily in the same cross-group comparison as in the MGCGA procedure and after adjusting for multiple comparisons they were not always significant for DIF. In the cross-country comparison IRTLR-DIF analysis, items 2, 7 and 12 were classified as having significant DIF most often after the BH-Adjustment. With such a disparate pattern of items exhibiting DIF the only interpretable reason for item misfit might be attributable to instrument translation issues as well as a range of other possible cultural, historical, economic and ethnic differences in these countries.

Research and Clinical Implications

For each country in the study, items were identified as exhibiting differential item functioning. There are several actions a researcher can take in order to deal with DIF items. The first step would be to qualitatively assess why the item is functioning differently between groups with a content expert. This could be in the form of back-translating the DIF items, to determine whether people from different countries might be interpreting the item content differently, or are they responding to items based on cultural mores' or is the construct ill-defined, does it lack meaningfulness in certain groups.

The second step would be to remove the problematic items, rewrite the items or score the item(s) differently based on group membership. For researchers in the instrument development stage, dropping DIF items may be the most efficient thing to do, but for an existing measure, dropping items may not be a good alternative. With respect to the GDS-15 and other health measures they tend to be short scales to begin with, so dropping items may reduce reliability and validity. For researchers who are strictly interested in group comparisons, dropping the items that exhibit DIF and using the DIF-free items to make group comparisons may also be a viable option.

If DIF with respect to gender was found, then a researcher may consider accounting for that DIF by adjusting scores with respect to gender. The issue with this approach is that if we used a gender adjusted score that accounted for men not endorsing items that indicate vulnerability, the DIF effect is an average over all men. So that means that there will be some men who are more sensitive and aware of their feelings and others who are not. So the "correction" for gender may be correct on average, but in adjusting

scores this way we may be running the risk of creating an illusion of a "gender fair" depression instrument, while in reality we might have just replaced one form of bias with another form of bias.

So where does this leave the clinician who wants to use this instrument with a population of adults in Latin America and the Caribbean? There is no clear guidance, but based on the results of the current study caution should be used in any cross-country or gender comparisons. Further research needs to be done with GDS-15 in Latin America and Caribbean before firm recommendations for its use can be made.

Limitations and Future Directions

Several of the more important study limitations merit attention. Of central concern is that there may be alternative factor structures for the GDS-15 that have not been explored, but might fit the data equally well for other sub-groups within the data, such as age-cohort groupings. In addition, the sampling framework for the SABE study focused specifically on older adults who resided in urban centers. With that said, the results presented in the current study may only be generalizable to individuals from urban communities and not rural communities. Limitations with respect to the study design are that cross-sectional data such as the SABE does not allow us to measure and document changes in depression longitudinally which could help researchers better understand and illustrate the complexities of depression in later life.

Limitations with respect to the factor structure(s) of the GDS-15 are that the GDS-15 has five items that must be reverse scored, because they are negatively worded items. The two-factor model found in the countries of Argentina, Mexico and Uruguay was defined as (1) general depressive affect and (2) life satisfaction. The life satisfaction factor was comprised of items 1, 5, 7, 11 and 13. All of these items are reverse scored (negatively worded) items that loaded on the second factor (life satisfaction). Research by (Chen, Rendina-Gobioff, & Dedrick, 2010; Roszkowski & Soven, 2010; Schriesheim, Eisenbach, & Hill, 1991) suggests that the inclusion of positive and negative items in the same scale can make constructs conceptualized as unidimensional appear multidimensional (e.g., positively and negatively worded items may form two separate factors). As such the second factor found in Argentina, Mexico and Uruguay may be a

result of the negative wording of the items on the GDS-15 scale. Further analyses, such as parallel analysis may be necessary to investigate how many factors should be retained. Parallel analysis is a method based on the generation of random variables to determine the number of factors to retain. Parallel analysis, compares the observed eigenvalues extracted from the correlation matrix to be analyzed with those obtained from uncorrelated normal variables (J. L. Horn, 1965).

Finally, the sensitivity and specificity of the GDS-15 accounting for DIF and not accounting for DIF could not be evaluated with the current data. Sensitivity and Specificity analyses require a comparison to a gold standard criterion of depression. The SABE study only has one measure of depression, the GDS-15. Future research should involve the assessment of sensitivity and specificity based on DIF analyses, cross-tabulating the GDS-15 scaled score against a gold standard depression scale score such as the CESD. Assessing the sensitivity and specificity of the GDS-15 with respect to a DIF analyses can inform health policy. It is important to find out how well the GDS-15 is at identifying individuals with and without depression, in order to appropriately treat depression in the elderly. This information would aid in avoiding under treatment or overtreatment of depressive conditions, which would ultimately save money and lives.

Future directions for work with the GDS-15 would be to move away from the assumption that the items that comprise the existing instrument represent the "best 15 items". Initial development of the GDS-15 did not involve rigorous psychometric work. Moving forward I suggest that the original 100 item GDS go through a factor analytical

and IRT analysis in order to select items that will be most effective along the depression continuum, in order to ultimately, build a better GDS-15.

Although invariance analyses of the kind used in this study should be applied at the time of translation, when changes can be made to items in order to eliminate or minimize DIF, this study adds to the methodological literature by illustrating both SEM and IRT based procedures for examining measurement invariance with an existing instrument. Finally, these analyses reflect the first rigorous psychometric assessment of the GDS-15 in the countries of Argentina, Cuba, Uruguay, Chile and Mexico.

Table 9 One factor EFA Chile

GDS-15 Items	Factor
	General Depressive Affect
V7: Do you feel happy most of the time	0.85
V8: Do you feel helpless	0.84
V12: Do you feel pretty worthless the way you are now	0.83
V14: Do you feel that your situation is hopeless	0.82
V3: Do you feel that your life is empty	0.81
V4: Do you often get bored	0.80
V1: Are you basically satisfied with your life	0.79
V5: Are you in good spirits most of the time	0.79
V13: Do you feel full of energy	0.75
V11: Do you think its wonderful to be alive now	0.74
V10: Do you feel you have more problems with memory than most	0.61
V6: Are you afraid that something bad is going to happen to you	0.59
V2:Have you dropped many of your activities and interests	0.59
V15: Do you think that most people are better off than you	0.49
V9: Do you prefer to stay at home rather than going out and doing new things	0.32

Table 10 One factor EFA Cuba

GDS-15 Items	Factor
	General Depressive Affect
V4: Do you often get bored	0.87
V7: Do you feel happy most of the time	0.87
V3: Do you feel that your life is empty	0.84
V5: Are you in good spirits most of the time	0.82
V8: Do you feel helpless	0.82
V1: Are you basically satisfied with your life	0.82
V14: Do you feel that your situation is hopeless	0.80
V12: Do you feel pretty worthless the way you are now	0.79
V13: Do you feel full of energy	0.76
V11: Do you think its wonderful to be alive now	0.73
V2:Have you dropped many of your activities and interests	0.66
V15: Do you think that most people are better off than you	0.64
V6: Are you afraid that something bad is going to happen to you	0.63
V10: Do you feel you have more problems with memory than most	0.54
V9: Do you prefer to stay at home rather than going out and doing new things	0.46

Table 11 Two factor EFA Argentina

	Factors	
GDS-15 Items	Life Satisfaction	General Depressive Affect
V11: Do you think its wonderful to be alive now	0.92	
V7: Do you feel happy most of the time	0.87	
V5: Are you in good spirits most of the time	0.79	
V13: Do you feel full of energy	0.71	
V1: Are you basically satisfied with your life	0.47	
V3: Do you feel that your life is empty		0.80
V12: Do you feel pretty worthless the way you are now		0.80
V2:Have you dropped many of your activities and interests		0.76
V4: Do you often get bored		0.75
V9: Do you prefer to stay at home rather than going out and doing new things		0.70
V14: Do you feel that your situation is hopeless		0.65
V8: Do you feel helpless		0.64
V6: Are you afraid that something bad is going to happen to you		0.60
V15: Do you think that most people are better off than you		0.54
V10: Do you feel you have more problems with memory than most		0.48

Table 12 Two factor EFA Mexico

	Factors	
GDS Items	Life Satisfaction	General Depressive Affect
V7: Do you feel happy most of the time	0.95	
V5: Are you in good spirits most of the time	0.83	
V1: Are you basically satisfied with your life	0.66	
V13: Do you feel full of energy	0.64	
V11: Do you think its wonderful to be alive now	0.54	
V14: Do you feel that your situation is hopeless		0.80
V10: Do you feel you have more problems with memory than most		0.74
V8: Do you feel helpless		0.73
V9: Do you prefer to stay at home rather than going out and doing new things		0.69
V12: Do you feel pretty worthless the way you are now		0.69
V15: Do you think that most people are better off than you		0.66
V3: Do you feel that your life is empty		0.61
V6: Are you afraid that something bad is going to happen to you		0.55
V4: Do you often get bored		0.52
V2:Have you dropped many of your activities and interests		0.48

Table 13 Two factor EFA Uruguay

	Factors	
GDS Items	Life Satisfaction	General Depressive Affect
V7: Do you feel happy most of the time	0.92	
V5: Are you in good spirits most of the time	0.85	
V1: Are you basically satisfied with your life	0.63	
V11: Do you think its wonderful to be alive now	0.62	
V13: Do you feel full of energy	0.45	
V12: Do you feel pretty worthless the way you are now		0.80
V9: Do you prefer to stay at home rather than going out and doing new things		0.67
V3: Do you feel that your life is empty		0.67
V2:Have you dropped many of your activities and interests		0.66
V4: Do you often get bored		0.65
V14: Do you feel that your situation is hopeless		0.60
V8: Do you feel helpless		0.59
V10: Do you feel you have more problems with memory than most		0.59
V15: Do you think that most people are better off than you		0.51
V6: Are you afraid that something bad is going to happen to you		0.43

Table 14 One factor CFA by Country

Country	Chi-Square	*df*	CFI	TLI	RMSEA
Argentina	377.89	52	0.89	0.93	0.08
Chile	269.943	70	0.96	0.98	0.04
Cuba	367.103	69	0.96	0.98	0.05
Mexico	518.617	64	0.93	0.96	0.06
Uruguay	349.851	66	0.94	0.97	0.05

Table 15 Two- factor CFA by country

Country	Chi-Square	*df*	CFI	TLI	RMSEA
Argentina	200.334	57	0.95	0.97	0.05
Mexico	287.968	66	0.96	0.98	0.04
Uruguay	234.107	66	0.97	0.98	0.04

****Two factor models were not estimated for the countries of Chile and Cuba because a one factor model was more parsimonious and provided better fit***

Table 16 Model fit for configural and nested models Chile by gender

Model	# Free Parms	Chi-Square Value	Chi-Square DF	Chi-Square p-value	CFI	TLI	RMSEA Estimate
1a. Configural	60	287.156	119	<.0000	0.966	0.983	0.047
2a. Metric	46	176.826	93	<.0000	0.983	0.989	0.038
3a. Scalar	32	189.152	102	<.0000	0.982	0.99	0.037
4a. Residuals Free	47	267.718	123	<.0000	0.969	0.985	0.044
4b. Residuals Fixed	32	189.152	101	<.0000	0.982	0.99	0.037

Table 17 Invariance hypothesis tests Chile by gender

Model	Chi-Square DIFFTEST· Value	*df*	Chi-Square p-value
1a. Metric vs. Configural	4.640	12	0.9689
2a. Scalar vs. Metric	18.922	13	0.1256
3a. Residual fixed vs. Residual Free	6.670	13	0.9183

Table 18 Model fit for configural and nested models Cuba by gender

Model	# Free Parm	Chi-Square Value	Chi-Square DF	Chi-Square p-value	CFI	TLI	RMSEA Estimate
1a. Configural	60	348.551	113	<.0000	0.966	0.981	0.049
2a. Metric	46	222.800	88	<.0000	0.981	0.986	0.042
3a. Scalar	32	235.256	95	<.0000	0.980	0.987	0.041
4a. Residuals Free	47	304.784	109	<.0000	0.972	0.984	0.046
4b. Residuals Fixed	32	235.256	95	<.0000	0.980	0.987	0.041

Table 19 Invariance hypothesis tests Cuba by gender

Model	Chi-Square DIFFTEST-Value	*df*	Chi-Square p-value
1a. Metric vs. Configural	11.062	12	0.5236
2a. Scalar vs. Metric	21.955	13	0.0561
3a. Residual fixed vs. Residual Free	14.606	13	0.3326

Table 20 Model fit for configural and nested models Argentina by gender

Model	# Free Parm	Chi-Square Value	Chi-Square DF	Chi-Square p-value	CFI	TLI	RMSEA Estimate
1a. Configural	62	217.883	92	<.0000	0.949	0.972	0.052
2a. Metric	49	179.440	85	<.0000	0.962	0.977	0.047
3a. Scalar1	36	192.255	91	<.0000	0.959	0.977	0.047
3b. Scalar2 (v9)	37	187.398	91	<.0000	0.961	0.978	0.046
4a. Residuals Free	52	207.339	93	<.0000	0.953	0.974	0.049
4b. Residuals Fixed	38	190.319	91	<.0000	0.960	0.977	0.047

Table 21 Invariance hypothesis tests Argentina by gender

Model	Chi-Square DIFFTEST-Value	*df*	Chi-Square p-value
1a. Metric vs. Configural	10.248	10	0.5245
2a. Scalar1 vs. Metric	25.700	12	0.0118
2b. Scalar2 vs. Metric	13.209	11	0.2799
3a. Residual fixed vs. Residual Free	15.478	12	0.2163

Table 22 Model fit for configural and nested models Mexico by gender

Model	# Free Parm	Chi-Square Value	Chi-Square DF	Chi-Square p-value	CFI	TLI	RMSEA Estimate
1a. Configural	62	295.673	108	<.0000	0.969	0.982	0.043
2a. Metric	49	203.898	85	<.0000	0.980	0.986	0.039
3a. Scalar1	36	224.634	92	<.0000	0.978	0.985	0.039
3b. Scalar2 (v8)	37	217.584	92	<.0000	0.979	0.986	0.038
3c. Scalar3 (v8 & v6)	38	212.943	91	<.0000	0.980	0.986	0.038
4a. Residuals Free	53	278.663	109	<.0000	0.972	0.984	0.041
4b. Residuals Fixed	40	219.026	92	<.0000	0.979	0.986	0.039

Table 23 Invariance hypothesis tests Mexico by gender

Model	Chi-Square DIFFTEST-Value	*df*	Chi-Square p-value
1a. Metric vs. Configural	13.427	11	0.2663
2a. Scalar1 vs. Metric	37.123	12	0.0002
2b. Scalar2 vs. Metric	19.978	11	0.0456
2c. Scalar3 vs. Metric	12.809	10	0.2346
3a. Residual fixed vs. Residual Free	15.106	10	0.1282

Table 24 Model fit for configural and nested models Uruguay by gender

Model	# Free Parm	Chi-Square Value	Chi-Square DF	Chi-Square p-value	CFI	TLI	RMSEA Estimate
1a. Configural	62	262.916	113	<.0000	0.968	0.983	0.043
2a. Metric	49	197.551	94	<.0000	0.978	0.986	0.039
3a. Scalar1	36	210.742	101	<.0000	0.976	0.986	0.039
3b. Scalar2 (v15)	37	206.335	100	<.0000	0.977	0.987	0.038
4a. Residuals Free	52	257.916	113	<.0000	0.969	0.984	0.042
4b. Residuals Fixed	38	216.483	103	<.0000	0.976	0.986	0.039

Table 25 Invariance hypothesis tests Uruguay by gender

Model	Chi-Square DIFFTEST-Value	*df*	Chi-Square p-value
1a. Metric vs. Configural	14.469	11	0.2081
2a. Scalar1 vs. Metric	22.296	12	0.0343
2b. Scalar2 vs. Metric	15.647	11	0.1547
3a. Residual fixed vs. Residual Free	14.503	12	0.2698

Table 26 Model fit for configural and nested models Chile by Cuba

Model	# Free Parm	Chi-Square Value	Chi-Square DF	Chi-Square p-value	CFI	TLI	RMSEA Estimate
1. Configural	60	637.503	139	<.0000	0.963	0.983	0.049
2a. Metric1	46	415.797	102	<.0000	0.977	0.985	0.045
2b. Metric2 (Ecov V5&V7)	47	388.958	102	<.0000	0.979	0.987	0.043
3a. Scalar1	33	522.949	111	<.0000	0.970	0.982	0.050
3b. Scalar2 (V1 & V15)	35	446.245	109	<.0000	0.975	0.985	0.045
3c. Scalar3 (V7 & V5)	37	403.223	108	<.0000	0.978	0.987	0.043
3d. Scalar4 (V9)	38	399.612	107	<.0000	0.978	0.987	0.043
4a. Residuals Free	53	547.288	141	<.0000	0.970	0.986	0.044
4b. Residual Fixed1	43	422.023	115	<.0000	0.977	0.987	0.042
4c. Residual Fixed2 (V2 & V4)	45	425.809	118	<.0000	0.977	0.988	0.042

Table 27 Invariance hypothesis tests Chile by Cuba

Model	Chi-Square DIFFTEST Value	*df*	Chi-Square p-value
1a. Metric1 vs. Configural	31.317	12	0.0018
2a. Metric2 vs. Configural	17.493	11	0.0941
2b. Scalar1 vs. Metric2	232.881	13	0.0000
3a. Scalar2 vs. Metric2	98.760	11	0.0000
3b. Scalar3 vs. Metric2	19.051	9	0.0248
3c. Scalar4 vs. Metric2	15.153	8	0.0562
4a. Residual Fixed1 vs. Residual Free	22.926	9	0.0064
4b. Residual Fixed2 vs. Residual Free	12.222	7	0.0935

Table 28 Model fit for configural and nested models Mexico by Uruguay

Model	# Free Parm	Chi-Square Value	Chi-Square DF	Chi-Square p-value	CFI	TLI	RMSEA Estimate
1a. Configural	62	521.543	132	<.0000	0.966	0.983	0.044
2a. Metric1	49	383.299	106	<.0000	0.976	0.985	0.040
2b. Metric2 (Ecorr v1&v3)	50	365.427	106	<.0000	0.977	0.986	0.039
2c. Metric3 (Ecorr v1,v3, v10,v15)	51	353.656	105	<.0000	0.978	0.987	0.038
3a. Scalar1	38	414.987	114	<.0000	0.974	0.985	0.040
3b. Scalar2 (v9&v10)	40	385.680	112	<.0000	0.976	0.986	0.038
3c. Scalar3 (v2, v4,v9,v10)	42	367.836	111	<.0000	0.978	0.987	0.037
3d. Scalar4 (v2, v4,v9,v10,v15)	43	363.617	110	<.0000	0.978	0.987	0.037
4a. Residuals Free	58	485.083	134	<.0000	0.969	0.985	0.040
4b. Residuals Fixed1 (ResVar v2, v4, v9, v10)	48	413.295	121	<.0000	0.974	0.986	0.038
4c. Residuals Fixed2 (ResVar v2,v4,v9, v10, v3)	48	413.351	122	<.0000	0.975	0.986	0.038

Table 29 Invariance hypothesis tests Mexico by Uruguay

Model	Chi-Square DIFFTEST-Value	*df*	Chi-Square p-value
1a. Metric1 vs. Configural	29.882	12	0.0029
1b. Metric2 vs. Configural	19.873	11	0.0471
1c. Metric3 vs. Configural	15.168	10	0.1261
2a. Scalar1 vs. Metric3	93.703	12	0.0000
2b. Scalar2 vs. Metric3	48.608	10	0.0000
2c. Scalar3 vs. Metric3	18.469	9	0.0301
2d. Scalar4 vs. Metric3	12.886	8	0.1158
3a. Residual fixed1 vs. Residual Free	16.453	8	0.0363
3b. Residual fixed2 vs. Residual Free	11.487	7	0.1187

Table 30 Model fit for configural and nested models Mexico by Argentina

Model	# Free Parm	Chi-Square Value	Chi-Square DF	Chi-Square p-value	CFI	TLI	RMSEA Estimate
1. Configural	62	483.263	122	<.0000	0.961	0.979	0.046
2a. Metric1	49	353.048	100	<.0000	0.973	0.982	0.042
2b. Metric2 (Ecorr v14&v15)	50	341.644	100	<.0000	0.974	0.983	0.041
3a. Scalar1	37	420.268	108	<.0000	0.966	0.980	0.045
3b. Scalar2 (v10)	38	393.026	107	<.0000	0.969	0.981	0.043
3c. Scalar3 (v6, v10,v15)	40	361.707	106	<.0000	0.972	0.983	0.041
3d. Scalar4 (v5,v6,v8,v10,v15)	42	350.199	105	<.0000	0.974	0.984	0.040
4a. Residuals Free	57	462.993	124	<.0000	0.963	0.981	0.044
4b. Residuals Fixed1	42	350.199	105	<.0000	0.974	0.984	0.040
4c. Residuals Fixed2 (ResVarV13)	43	348.843	106	<.0000	0.974	0.984	0.040

Table 31 Invariance hypothesis tests Mexico by Argentina

Model	Chi-Square DIFFTEST-Value	*df*	Chi-Square p-value
1a. Metric1 vs. Configural	22.264	11	0.0224
1b. Metric2 vs. Configural	17.300	11	0.0993
2a. Scalar1 vs. Metric2	134.600	12	0.0000
2b. Scalar2 vs. Metric2	88.740	11	0.0000
2c. Scalar3 vs. Metric2	33.213	10	0.0003
2d. Scalar4 vs. Metric2	9.510	8	0.3011
3a. Residual fixed1 vs. Residual Free	22.357	12	0.0337
3b. Residual fixed2 vs. Residual Free	16.785	11	0.1144

Table 32 Model fit for configural and nested models Uruguay by Argentina

Model	# Free Parm	Chi-Square Value	Chi-Square DF	Chi-Square p-value	CFI	TLI	RMSEA Estimate
1a. Configural	62	434.286	123	<.0000	0.963	0.980	0.045
2a. Metric1	49	321.474	101	<.0000	0.974	0.983	0.042
2b. Metric2 (Ecorr v3&v4)	50	300.557	100	<.0000	0.976	0.984	0.040
3a. Scalar1	37	335.222	108	<.0000	0.973	0.983	0.041
3b. Scalar2 (v15)	38	322.393	108	<.0000	0.975	0.984	0.040
3c. Scalar3 (v12,v15)	39	314.153	107	<.0000	0.975	0.985	0.040
4a. Residuals Free	54	396.452	122	<.0000	0.967	0.982	0.043
4b. Residuals Fixed1	41	334.306	110	<.0000	0.973	0.984	0.041
4c. Residuals Fixed2 (ResVar v9,v12,v15)	42	334.422	112	<.0000	0.974	0.984	0.040

Table 33 Invariance hypothesis tests Uruguay by Argentina

Model	Chi-Square DIFFTEST-Value	*df*	Chi-Square p-value
1a. Metric1 vs. Configural	50.169	12	0.0000
1b. Metric2 vs. Configural	11.502	10	0.3198
2a. Scalar1 vs. Metric2	55.601	12	0.0000
2b. Scalar2 vs. Metric2	28.958	11	0.0023
2c. Scalar3 vs. Metric2	16.380	10	0.0893
3a. Residual fixed1 vs. Residual Free	20.943	11	0.0340
3b. Residual fixed2 vs. Residual Free	14.925	11	0.1859

Table 34 *Item parameters and standard errors for anchor items Chile by Gender*

Item	Content	a	b
1	Are you basically satisfied with your life	2.22 (.23)	1.15 (.08)
2	Have you dropped many of your activities and interests	1.18 (.12)	.59 (.09)
3	Do you feel that your life is empty	2.41 (.20)	.34 (.05)
4	Do you often get bored	2.30 (.19)	.33 (.05)
5	Are you in good spirits most of the time	2.18 (.20)	.97 (.07)
6	Are you afraid that something bad is going to happent to you	1.26 (.12)	.56 (.08)
7	Do you feel happy most of the time	2.59 (.23)	.81 (.06)
8	Do you often feel helpless	2.63 (.24)	.84 (.06)
9	Do you prefer to stay at home, rather than going out and doing things	.54 (.08)	-1.13 (.21)
10	Do you feel that you have more problems with memory than most	1.29 (.14)	1.19 (.13)
11	Do you think it is wonderful to be alive now	1.86 (.19)	1.56 (.13)
12	Do you feel worthless the way you are now	2.44 (.22)	.96 (.06)
14	Do you feel that your situation is hopeless	2.31 (.21)	.83 (.07)
15	Do you think that most people are better off than you are	.83 (.11)	-.05 (.11)

Table 35 Item parameters and standard errors for items exhibiting DIF Chile by Gender

					Tests for DIF: $\chi 2$ (P)	
Item	Content	Group	a	b	*a* Dif	*b* Dif
13	Do you feel full of energy	Men	1.68 (.26)	.74 (.15)	3.1 (.078)	1.5 (.220)
		Women	2.00 (.23)	.79 (.08)		

Item 13 did not exhibit DIF after using the B-H multiple comparisons adjustment

Table 36 Item parameters and standard errors for anchor items Cuba by Gender

Item	Content	a	b
1	Are you basically satisfied with your life	2.68 (.20)	.30 (.05)
2	Have you dropped many of your activities and interests	1.62 (.12)	.29 (.07)
3	Do you feel that your life is empty	2.91 (.21)	-.01(.04)
4	Do you often get bored	3.32 (.22)	-.02 (.03)
5	Are you in good spirits most of the time	2.54 (.19)	.33 (.05)
6	Are you afraid that something bad is going to happent to you	1.56 (.12)	.29 (.07)
8	Do you often feel helpless	2.88 (.23)	.50 (.05)
9	Do you prefer to stay at home, rather than going out and doing things	.92 (.08)	-.85 (.08)
10	Do you feel that you have more problems with memory than most	1.18 (.12)	1.06 (.16)
11	Do you think it is wonderful to be alive now	2.23 (.22)	1.26 (.12)
13	Do you feel full of energy	2.11 (.16)	.51 (.07)
14	Do you feel that your situation is hopeless	2.60 (.21)	.59 (.06)
15	Do you think that most people are better off than you are	1.43 (.11)	.21 (.08)

Table 37 Item parameters and standard errors for items exhibiting DIF Cuba by Gender

Item	Content	Group	a	b	Tests for DIF: $\chi 2$ (P)	
					a Dif	*b* Dif
7	Do you feel happy most of the time	Men	3.10 (.20)	.13 (.07)	2.3 (.129)	4.1 (.042)
		Women	3.10 (.20)	.00 (.00)		
12	Do you feel worthless the way you are now	Men	2.46 (.17)	.46 (.09)	2.8 (.094)	3.5 (.061)
		Women	2.46 (.17)	.00 (.00)		

Items 7 and 12 did not exhibit DIF After using the B-H multiple comparisons adjustment

Table 38 Item parameters and standard errors for anchor items Argentina by Gender

Item	Content	a	b
1	Are you basically satisfied with your life	2.94 (.31)	.44 (.06)
2	Have you dropped many of your activities and interests	1.67 (.18)	.49 (.10)
3	Do you feel that your life is empty	2.59 (.28)	.43 (.07)
4	Do you often get bored	1.97 (.22)	.49 (.09)
5	Are you in good spirits most of the time	2.19 (.25)	.71 (.09)
6	Are you afraid that something bad is going to happent to you	1.13 (.16)	.97 (.19)
7	Do you feel happy most of the time	2.82 (.30)	.44 (.07)
8	Do you often feel helpless	2.79 (.35)	.80 (.09)
10	Do you feel that you have more problems with memory than most	1.03 (.21)	2.00 (.43)
13	Do you feel full of energy	1.47 (.17)	.54 (.11)
14	Do you feel that your situation is hopeless	2.18 (.27)	.84 (.12)

Table 39 Item parameters and standard errors for items exhibiting DIF Argentina by Gender

					Tests for DIF: $\chi 2$ (P)	
Item	Content	Group	a	b	*a* Dif	*b* Dif
9	Do you prefer to stay home rather than going out and doing things	Men	.87 (.10)	.02 (.19)	0.7 (.402)	3.6 (.057)
		Women	.87 (.10)	.00 (.00)		
11	Do you think its wonderful to be alive now	Men	2.23 (.60)	.97 (.27)	6.0 (.014)	0.0 (1.000)
		Women	1.82 (.34)	1.17 (.20)		
12	Do you feel worthless the way you are now	Men	2.22 (.20)	.72 (.12)	1.0 (.317)	4.0 (.045)
		Women	2.22 (.20)	.00 (.00)		
15	Do you think that most people are better off than you are	Men	1.00 (.13)	-.26 (.18)	0.0 (1.000)	6.1 (.013)
		Women	1.00 (.13)	.00 (.00)		

No items exhibited DIF after using the B-H multiple comparisons adjustment

Table 40 Item parameters and standard errors for anchor items Mexico by Gender

Item	Content	a	b
1	Are you basically satisfied with your life	2.14 (.19)	1.00 (.08)
2	Have you dropped many of your activities and interests	1.48 (.12)	.50 (.07)
3	Do you feel that your life is empty	2.76 (.19)	.29 (.04)
4	Do you often get bored	2.40 (.17)	.36 (.05)
5	Are you in good spirits most of the time	1.77 (.15)	.99 (.09)
7	Do you feel happy most of the time	2.13 (.18)	.85 (.07)
9	Do you prefer to stay at home, rather than going out and doing things	1.22 (.10)	-.31 (.06)
10	Do you feel that you have more problems with memory than most	1.26 (.11)	.79 (.10)
11	Do you think it is wonderful to be alive now	1.68 (.19)	1.78 (.18)
12	Do you feel worthless the way you are now	2.21 (.18)	.78 (.07)
13	Do you feel full of energy	1.63 (.13)	.83 (.09)
14	Do you feel that your situation is hopeless	2.38 (.18)	.62 (.06)
15	Do you think that most people are better off than you are	1.31 (.11)	.30 (.07)

Table 41 Item parameters and standard errors for items exhibiting DIF Mexico by Gender

Item	Content	Group	a	b	Tests for DIF: $\chi 2$ (P)	
					a Dif	*b* Dif
6	**Are you afraid that something bad is going to happent to you**	**Men**	**1.54 (.10)**	**.39 (.10)**	**2.8 (.094)**	**7.2 (.007)**
		Women	**1.54 (.10)**	**.00 (.00)**		
8	**Do you often feel helpless**	**Men**	**3.16 (.20)**	**.72 (..08)**	**0.6 (.438)**	**24.3 (.000)**
		Women	**3.16 (.20)**	**.00 (.00)**		

Items bolded were significant after BH-Adjustment

Table 42 Item parameters and standard errors for anchor items Uruguay by Gender

Item	Content	a	b
1	Are you basically satisfied with your life	2.12 (.19)	1.05 (.07)
2	Have you dropped many of your activities and interests	1.56 (.14)	.72 (.07)
3	Do you feel that your life is empty	2.04 (.17)	.90 (.07)
4	Do you often get bored	1.85 (.16)	.74 (.07)
7	Do you feel happy most of the time	3.33 (.30)	.79 (.05)
8	Do you often feel helpless	2.86 (.26)	.99 (.06)
9	Do you prefer to stay at home, rather than going out and doing things	.89 (.09)	-.16 (.09)
10	Do you feel that you have more problems with memory than most	1.08 (.15)	1.93 (.24)
11	Do you think it is wonderful to be alive now	2.42 (.27)	1.42 (.10)
13	Do you feel full of energy	1.67 (.15)	.92 (.08)
15	Do you think that most people are better off than you are	.96 (.11)	.80 (.13)

Table 43 Item parameters and standard errors for items exhibiting DIF Uruguay by Gender

Item	Content	Group	a	b	Tests for DIF: $\chi 2$ (P) *a* Dif	Tests for DIF: $\chi 2$ (P) *b* Dif
5	**Are you in good spirits most of the time**	**Men**	**1.76 (.30)**	**1.07 (.17)**	**5.9 (.015)**	**3.3 (.069)**
		Women	**2.81 (.31)**	**1.02 (.07)**		
6	Are you afraid that something bad is going to happent to you	Men	1.16 (.22)	1.27 (.27)	2.3 (.129)	1.8 (.179)
		Women	1.66 (.17)	.81 (.09)		
12	Do you feel worthless the way you are now	Men	2.93 (.59)	1.25 (.14)	4.0 (.045)	0.1 (.751)
		Women	2.21 (.30)	1.44 (.12)		
14	**Do you feel that your situation is hopeless**	**Men**	**2.34 (.43)**	**1.17 (.16)**	**11.4 (.000)**	**1.1 (.294)**
		Women	**2.90 (.35)**	**1.14 (.07)**		

Items 6 and 12 did not exhibit DIF After using the B-H multiple comparisons adjustment
Items bolded were significant after BH-Adjustment

Table 44 Summary of DIF analyses of the GDS-15 gender anchor items

Item Content	Argentina	Chile	Cuba	Mexico	Uruguay
v1. Are you basically satisfied with your life	√	√	√	√	√
v2. Have you dropped many of your activities and interests	√	√	√	√	√
v3. Do you feel that your life is empty	√	√	√	√	√
v4. Do you often get bored	√	√	√	√	√
v5. Are you in good spirits most of the time	√	√	√	√	√
v6. Are you afraid that something bad is going to happen to you	√	√	√		
v7. Do you feel happy most of the time	√	√		√	
v8. Do you often feel helpless	√	√	√		√
v9. Do you prefer to stay at home, rather than going out and doing things		√	√	√	√
v10. Do you feel that you have more problems with memory than most	√	√	√	√	√
v11. Do you think it is wonderful to be alive now		√	√	√	√
v12. Do you feel worthless the way you are now		√		√	
v13. Do you feel full of energy			√	√	√
v14. Do you feel that your situation is hopeless	√	√	√	√	
v15. Do you think that most people are better off than you are		√	√	√	√

√=*Anchor Items by Country*

Table 45 Summary of DIF analyses of the GDS-15: gender type of DIF

Item Content	Argentina Type of DIF, if Present	Chile Type of DIF, if Present	Cuba Type of DIF, if Present	Mexico Type of DIF, if Present	Uruguay Type of DIF, if Present
v1. Are you basically satisfied with your life					
v2. Have you dropped many of your activities and interests					
v3. Do you feel that your life is empty					
v4. Do you often get bored					
v5. Are you in good spirits most of the time					*NU*
v6. Are you afraid that something bad is going to happen to you				*U*	*NU*
v7. Do you feel happy most of the time			*U*		
v8. Do you often feel helpless				*U*	
v9. Do you prefer to stay at home, rather than going out and doing things	*U*				
v10. Do you feel that you have more problems with memory than most					
v11. Do you think it is wonderful to be alive now	*NU*				
v12. Do you feel worthless the way you are now	*U*		*U*		*NU*
v13. Do you feel full of energy		*NU*			
v14. Do you feel that your situation is hopeless					*NU*
v15. Do you think that most people are better off than you are	*U*				

NU=Non-Uniform DIF, U=Uniform DIF

Table 46 Summary of DIF analyses of the GDS-15 gender BH-Adjustment

Item Content	Argentina DIF After BH Adjustment	Chile DIF After BH Adjustment	Cuba DIF After BH Adjustment	Mexico DIF After BH Adjustment	Uruguay DIF After BH Adjustment
v1. Are you basically satisfied with your life					
v2. Have you dropped many of your activities and interests					
v3. Do you feel that your life is empty					
v4. Do you often get bored					
v5. Are you in good spirits most of the time					*yes*
v6. Are you afraid that something bad is going to happent to you				*yes*	*no*
v7. Do you feel happy most of the time			*no*		
v8. Do you often feel helpless				*yes*	
v9. Do you prefer to stay at home, rather than going out and doing things	*no*				
v10. Do you feel that you have more problems with memory than most					
v11. Do you think it is wonderful to be alive now	*no*				
v12. Do you feel worthless the way you are now	*no*		*no*		*no*
v13. Do you feel full of energy		*no*			
v14. Do you feel that your situation is hopeless					*yes*
v15. Do you think that most people are better off than you are	*no*				

** *Yes=significant after Benjimini-Hochberg Adjustment, No=non-significant after adjustment***

Table 47 Item parameters and standard errors for anchor items Chile by Cuba

Item	Content	*a*	*b*
1	Are you basically satisfied with your life	2.43 (.15)	.62 (.04)
2	Have you dropped many of your activities and interests	1.29 (.13)	.34 (.09)
3	Do you feel that your life is empty	2.89 (.15)	.10 (.03)
5	Are you in good spirits most of the time	2.45 (.15)	.55 (.04)
6	Are you afraid that something bad is going to happen to you	1.38 (.13)	.32 (.08)
7	Do you feel happy most of the time	2.88 (.26)	.53 (.05)
9	Do you prefer to stay at home, rather than going out and doing things	.83 (.06)	-.82 (.07)
10	Do you feel that you have more problems with memory than most	1.34 (.10)	.99 (.09)
11	Do you think it is wonderful to be alive now	2.06 (.23)	1.22 (.11)
12	Do you feel worthless the way you are now	2.56 (.16)	.69 (.05)
13	Do you feel full of energy	2.15 (.13)	.55 (.05)
14	Do you feel that your situation is hopeless	2.61 (.16)	.62 (.04
15	Do you think that most people are better off than you are	1.28 (.08)	.11 (.06)

Table 48 Item parameters and standard errors for items exhibiting DIF for Chile by Cuba

Item	Content	Group	*a*	*b*	Tests for DIF: $\chi 2$ (P)	
					a Dif	*b Dif*
4	Do you often get bored	Chile	3.19 (.16)	.08 (.03)	**-0.0 (1.000)**	**23 (.0000)**
		Cuba	3.19 (.16)	.00 (.00)		
8	Do you often feel helpless	Chile	2.87 (.14)	.56 (.06)	**0.3 (.584)**	**16.4 (.0001)**
		Cuba	2.87 (.14)	.00 (.00)		

Items bolded were significant after BH-Adjustment

Table 49 Item parameters and standard errors for anchor items Mexico by Uruguay

Item	Content	*a*	*b*
1	Are you basically satisfied with your life	2.41 (.13)	.55 (.05)
2	Have you dropped many of your activities and interests	1.69 (.09)	.18 (.05)
4	Do you often get bored	2.40 (.12)	.11 (.03)
5	Are you in good spirits most of the time	2.29 (.15)	.51 (.05)
8	Do you often feel helpless	3.21 (.19)	.25 (.03)
10	Do you feel that you have more problems with memory than most	1.27 (.09)	.78 (.09)
11	Do you think it is wonderful to be alive now	2.25 (.16)	1.03 (.08)
13	Do you feel full of energy	1.88 (.10)	.40 (.05)
14	Do you feel that your situation is hopeless	2.71 (.17)	.41 (.04)
15	Do you think that most people are better off than you are	1.27 (.08)	.10 (.06)

Table 50 Item parameters and standard errors for items exhibiting DIF for Mexico by Uruguay

					Tests for DIF: $\chi 2$ (P)	
Item	Content	Group	*a*	*b*	*a* Dif	*b* Dif
3	Do you feel that your life is empty	Uruguay	2.20 (.17)	.30 (.06)	5.4(.020)	1.0 (.317)
		Mexico	3.26 (.23)	.03 (.03)		
6	Are you afraid that something bad is going to happen to you	Uruguay	1.66 (.08)	.29 (.06)	**1.1 (.294)**	**36.6 (.000)**
		Mexico	1.66 (.08)	.00 (.00)		
7	Do you feel happy most of the time	Uruguay	3.39 (.17)	.21 (.04)	**2.9 (.089)**	**11.1 (.001)**
		Mexico	3.39 (.17)	.00 (.00)		
9	Do you prefer to stay at home, rather than going out and doing things	Uruguay	.85 (.06)	-.70 (.09)	**2.9 (.089)**	**7.6 (.006)**
		Mexico	.85 (.06)	.00 (.00)		
12	Do you feel worthless the way you are now	Uruguay	3.11 (.16)	.66 (.05)	**1.4 (.237)**	**42.1 (.000)**
		Mexico	3.11 (.16)	.00 (.00)		

Item 3 did not exhibit DIF using the B-H multiple comparisons procedure

Items bolded were significant after BH-Adjustment

Table 51 Item parameters and standard errors for anchor items Argentina by Mexico

Item	Content	*a*	*b*
1	Are you basically satisfied with your life	2.30 (0.16)	0.70 (0.05)
3	Do you feel that your life is empty	2.86 (0.17)	0.23 (0.03)
4	Do you often get bored	2.37 (0.14)	0.28 (0.04)
5	Are you in good spirits most of the time	1.91 (0.14)	0.79 (0.07)
6	Are you afraid that something bad is going to happent to you	1.37 (0.09)	0.37 (0.06)
8	Do you often feel helpless	3.01 (0.19)	0.44 (0.04)
10	Do you feel that you have more problems with memory than most	1.25 (0.10)	0.93 (0.10)
13	Do you feel full of energy	1.55 (0.10)	0.64 (0.07)
14	Do you feel that your situation is hopeless	2.41 (0.16)	0.57 (0.05)
15	Do you think that most people are better off than you are	1.16 (0.08)	0.16 (0.07)

Table 52 Item parameters and standard errors for items exhibiting DIF for Argentina by Mexico

Item	Content	Group	*a*	*b*	Tests for DIF: $\chi 2$ (P) *a* Dif	Tests for DIF: $\chi 2$ (P) *b* Dif
2	Do you feel that your life is empty	Argentina	1.88 (0.10)	0.29 (0.07)	**1.3 (.254)**	**35.4 (.000)**
		Mexico	1.88 (0.10)	.00 (.00)		
7	Do you feel happy most of the time	Argentina	2.60 (0.29)	0.33 (0.07)	**5.1 (.024)**	**8.6 (.003)**
		Mexico	2.27 (0.19)	0.72 (0.07)		
9	Do you prefer to stay at home, rather than going out and doing things	Argentina	0.98 (0.07)	-0.46 (0.10)	**0.1 (.752)**	**7.5 (.006)**
		Mexico	0.98 (0.07)	0.00 (0.00)		
11	Do you think it is wonderful to be alive now	Argentina	1.83 (0.25)	1.01 (0.17)	**6.9 (.009)**	**0.4 (.527)**
		Mexico	1.78 (0.20)	1.62 (0.17)		
12	Do you feel worthless the way you are now	Argentina	2.54 (0.13)	0.69 (0.07)	**1.4 (.237)**	**26.9 (.000)**
		Mexico	2.54 (0.13)	.00 (.00)		

Items bolded were significant after BH-Adjustment

Table 53 Item parameters and standard errors for anchor items Argentina by Uruguay

Item	Content	*a*	*b*
1	Are you basically satisfied with your life	2.35 (.16)	.83 (.05)
3	Do you feel that your life is empty	2.20 (.15)	.73 (.05)
4	Do you often get bored	1.90 (.13)	.65 (.05)
5	Are you in good spirits most of the time	2.25 (.16)	.92 (.05)
8	Do you often feel helpless	2.81 (.21)	.92 (.05)
9	Do you prefer to stay at home, rather than going out and doing things	.98 (.08)	-.20 (.06)
10	Do you feel that you have more problems with memory than most	1.06 (.12)	1.94 (.21)
12	Do you feel worthless the way you are now	2.32 (.20)	1.21 (.07)
13	Do you feel full of energy	1.57 (.11)	.78 (.07)
14	Do you feel that your situation is hopeless	2.43 (.18)	1.04 (.06)
15	Do you think that most people are better off than you are	.91 (.08)	.55 (.10)

Table 54 Item parameters and standard errors for Items exhibiting DIF for Argentina by Uruguay

Item	Content	Group	*a*	*b*	Tests for DIF: $\chi 2$ (P)	
					a Dif	*b* Dif
2	Do you feel that your life is empty	Argentina	1.67 (.10)	.70 (.07)	**2.5 (.113)**	**8.1 (.004)**
		Uruguay	1.67 (.10)	.00 (.00)		
6	Are you afraid that something bad is going to happent to you	Argentina	1.07 (.15)	1.23 (.20)	**9.4 (.002)**	**2.8 (.094)**
		Uruguay	1.54 (.14)	.77 (.08)		
7	Do you feel happy most of the time	Argentina	2.66 (.29)	.68 (.07)	**7.2 (.007)**	**0.1 (.751)**
		Uruguay	3.37 (.30)	.66 (.05)		
11	Do you think it is wonderful to be alive now	Argentina	1.86 (.28)	1.36 (.16)	**10.7 (.001)**	**0.1 (.751)**
		Uruguay	2.45 (.27)	1.28 (.10)		

Items bolded were significant after BH-Adjustment

Table 55 Item parameters and standard errors for anchor items Argentina by Chile

Item	Content	*a*	*b*
1	Are you basically satisfied with your life	2.06 (.14)	.90 (.06)
3	Do you feel that your life is empty	2.49 (.16)	.32 (.04)
4	Do you often get bored	2.17 (.14)	.33 (.04)
5	Are you in good spirits most of the time	2.01 (.13)	.87 (.06)
6	Are you afraid that something bad is going to happen to you	1.23 (.09)	.63 (.08)
8	Do you often feel helpless	2.66 (.19)	.79 (.05)
9	Do you prefer to stay at home, rather than going out and doing things	.83 (.07)	-.65 (.08)
10	Do you feel that you have more problems with memory than most	1.25 (.11)	1.31 (.12)
11	Do you think that it is wonderful to be alive now	1.70 (.15)	1.45 (.11)
12	Do you feel worthless the way you are now	2.35 (.17)	.89 (.06)
13	Do you feel full of energy	1.53 (.10)	.68 (.07)
14	Do you feel that your situation is hopeless	2.22 (.16)	.79 (.06)
15	Do you think that most people are better off than you are	.86 (.08)	-.04 (.09)

Table 56 Item parameters and standard errors for items exhibiting DIF for Argentina by Chile

Item	Content	Group	*a*	*b*	Tests for DIF: $\chi 2$ (P)	
					a Dif	*b* Dif
2	Do you feel that your life is empty	Argentina	1.39 (.09)	.56 (.09)	**0.1 (.751)**	**60.6 (.000)**
		Chile	1.39 (.09)	.00 (.00)		
7	Do you feel happy most of the time	Argentina	2.46 (.25)	.46 (.08)	5.2 (.023)	0.0 (1.000)
		Chile	2.60 (.23)	.74 (.06)		

Item 7 did not exhibit DIF using the B-H multiple comparisons procedure

Items bolded were significant after BH-Adjustment

Table 57 Item parameters and standard errors for anchor items Argentina by Cuba

Item	Content	*a*	*b*
1	Are you basically satisfied with your life	2.64 (.17)	.68 (.04)
3	Do you feel that your life is empty	2.76 (.16)	.45 (.03)
4	Do you often get bored	2.72 (.16)	.44 (.03)
6	Are you afraid that something bad is going to happen to you	1.39 (.10)	.80 (.07)
8	Do you often feel helpless	2.80 (.20)	.93 (.04)
9	Do you prefer to stay at home, rather than going out and doing things	.99 (.07)	-.35 (.06)
10	Do you feel that you have more problems with memory than most	1.14 (.11)	1.64 (.15)
11	Do you think that it is wonderful to be alive now	1.89 (.17)	1.62 (.11)
12	Do you feel worthless the way you are now	2.30 (.17)	1.04 (.06)
13	Do you feel full of energy	1.68 (.12)	.88 (.07)
14	Do you feel that your situation is hopeless	2.35 (.17)	1.01 (.06)
15	Do you think that most people are better off than you are	1.17 (.09)	.52 (.07)

Table 58 Item parameters and standard errors for items exhibiting DIF for Argentina by Cuba

Item	Content	Group	*a*	*b*	Tests for DIF: $\chi 2$ (P) *a* Dif	Tests for DIF: $\chi 2$ (P) *b* Dif
2	Have you dropped many of your activities and interests	Argentina	1.66 (.09)	.65 (.08)	**1.1 (.294)**	**10.6 (.001)**
		Cuba	1.66 (.09)	.00 (.00)		
5	Are you in good spirits most of the time	Argentina	2.09 (.24)	.89 (.10)	**5.5 (.019)**	**1.5 (.220)**
		Cuba	2.47 (.19)	.75 (.05)		
7	Do you feel happy most of the time	Argentina	2.63 (.29)	.62 (.07)	**13 (.000)**	**0.0 (1.000)**
		Cuba	2.90 (.22)	.49 (.04)		

Items bolded were significant after BH-Adjustment

Table 59 Item parameters and standard errors for anchor items Mexico by Chile

Item	Content	*a*	*b*
1	Are you basically satisfied with your life	2.46 (.16)	.67 (.05)
2	Have you dropped many of your activities and interests	1.51 (.10)	.22 (.05)
3	Do you feel that your life is empty	2.98 (.16)	.01 (.03)
4	Do you often get bored	2.71 (.15)	.04 (.03)
5	Are you in good spirits most of the time	2.22 (.14)	.59 (.05)
6	Are you afraid that something bad is going to happen to you	1.48 (.09)	.11 (.05)
9	Do you prefer to stay at home, rather than going out and doing things	1.01 (.07)	-.69 (.05)
10	Do you feel that you have more problems with memory than most	1.35 (.10)	.63 (.07)
13	Do you feel full of energy	2.01 (.12)	.42 (.05)
14	Do you feel that your situation is hopeless	2.60 (.15)	.37 (.04)

Table 60 Item parameters and standard errors for items exhibiting DIF for Mexico by Chile

Item	Content	Group	*a*	*b*	Tests for DIF: $\chi 2$ (P) *a* Dif	*b* Dif
7	Do you feel happy most of the time	Chile	**3.15 (.16)**	**.40 (.04)**	**0.4 (527)**	**10.8 (001)**
		Mexico	**3.15 (.16)**	**.00 (.00)**		
8	Do you often feel helpless	Chile	**2.97 (.27)**	**.44 (.05)**	**4.8 (.028)**	**3.3 (.069)**
		Mexico	**3.46 (.26)**	**.14 (.04)**		
11	Do you think that it is wonderful to be alive now	Chile	2.08 (.23)	1.10 (.11)	4.4 (.035)	0.4 (.527)
		Mexico	1.94 (.22)	1.31 (.15)		
12	Do you feel worthless the way you are now	Chile	**3.09 (.15)**	**.51 (.04)**	**3.4 (.065)**	**4.9 (.026)**
		Mexico	**3.09 (.15)**	**.00 (.00)**		
15	Do you think that most people are better off than you are	Chile	1.30 (.08)	-.36 (.07)	0.2 (.654)	4.9 (.026)
		Mexico	1.30 (.08)	.00 (.00)		

Items 11 and 15 did not exhibit DIF using the B-H multiple comparisons procedure

Items bolded were significant after BH-Adjustment

Table 61 Item parameters and standard errors for anchor items Mexico by Cuba

Item	Content	*a*	*b*
1	Are you basically satisfied with your life	2.64 (.16)	.39 (.04)
3	Do you feel that your life is empty	3.14 (.16)	-.04 (.02)
5	Are you in good spirits most of the time	2.36 (.13)	.39 (.04)
6	Are you afraid that something bad is going to happen to you	1.60 (.09)	.10 (.04)
8	Do you often feel helpless	3.21 (.17)	.24 (.03)
9	Do you prefer to stay at home, rather than going out and doing things	1.16 (.07)	-.68 (.04)
10	Do you feel that you have more problems with memory than most	1.35 (.09)	.64 (.08)
13	Do you feel full of energy	2.09 (.12)	.40 (.05)
15	Do you think that most people are better off than you are	1.53 (.09)	.04 (.05)

Table 62 Item parameters and standard errors for items exhibiting DIF for Mexico by Cuba

Item	Content	Group	*a*	*b*	Tests for DIF: $\chi 2$ (P) *a* Dif	Tests for DIF: $\chi 2$ (P) *b* Dif
2	Have you dropped many of your activities and interests	Mexico	1.91 (.08)	.10 (.04)	**3.3 (.069)**	**12.9 (000)**
		Cuba	1.91 (.08)	.00 (.00)		
4	Do you often get bored	Mexico	2.98 (.14)	.03 (.03)	**1.1 (.294)**	**6.4 (.011)**
		Cuba	2.98 (.14)	.00 (.00)		
7	Do you feel happy most of the time	Mexico	2.95 (.13)	.39 (.04)	**0.0 (1.000)**	**12 (.000)**
		Cuba	2.95 (.13)	.00 (.00)		
11	Do you think that it is wonderful to be alive now	Mexico	1.96 (.21)	1.25 (.15)	3.6 (.057)	0.4 (.527)
		Cuba	2.45 (.25)	1.07 (.11)		
12	Do you feel worthless the way you are now	Mexico	2.99 (.14)	.35 (.04)	**1.6 (.205)**	**27.9 (.000)**
		Cuba	2.99 (.14)	.00 (.00)		
14	Do you feel that your situation is hopeless	Mexico	3.15 (.15)	.23 (.03)	**0.0 (1.000)**	**22.8 (.000)**
		Cuba	3.15 (.15)	.00 (.00)		

Item 11 did not exhibit DIF using the B-H multiple comparisons procedure

Items bolded were significant after BH-Adjustment

Table 63 Item parameters and standard errors for anchor items Uruguay by Chile

Item	Content	*a*	*b*
1	Are you basically satisfied with your life	2.22 (0.15)	0.77 (0.05)
4	Do you often get bored	2.26 (0.13)	0.21 (0.04)
5	Are you in good spirits most of the time	2.36 (0.15)	0.66 (0.05)
6	Are you afraid that something bad is going to happen to you	1.51 (0.10)	0.41 (0.05)
7	Do you feel happy most of the time	3.05 (0.19)	0.47 (0.04)
8	Do you often feel helpless	2.94 (0.19)	0.57 (0.04)
9	Do you prefer to stay at home, rather than going out and doing things	0.81 (0.06)	-0.73 (0.07)
11	Do you think that it is wonderful to be alive now	2.21 (0.18)	1.13 (0.08)
12	Do you feel worthless the way you are now	2.61 (0.18)	0.80 (0.05)
13	Do you feel full of energy	1.89 (0.12)	0.51 (0.05)
15	Do you think that most people are better off than you are	1.02 (0.08)	0.09 (0.07)

Table 64 Item parameters and standard errors for items exhibiting DIF for Uruguay by Chile

Item	Content	Group	*a*	*b*	Tests for DIF: $\chi 2$ (P) *a* Dif	Tests for DIF: $\chi 2$ (P) *b* Dif
2	Have you dropped many of your activities and interests	Chile	**1.27 (0.13)**	**0.36 (0.09)**	**4.6 (.031)**	**41.3 (.000)**
		Uruguay	**1.71 (0.15)**	**0.30 (0.07)**		
3	Do you feel that your life is empty	Chile	2.60 (0.22)	0.13 (0.04)	2.9 (.089)	2.3 (.129)
		Uruguay	2.24 (0.19)	0.46 (0.06)		
10	Do you feel that you have more problems with memory than most	Chile	**1.39 (0.15)**	**0.92 (0.12)**	**4.6 (.032)**	**18.5 (.000)**
		Uruguay	**1.15 (0.16)**	**1.45 (0.23)**		
14	Do you feel that your situation is hopeless	Chile	**2.61 (0.14)**	**0.57 (0.05)**	**1.3 (.254)**	**26.7 (.000)**
		Uruguay	**2.61 (0.14)**	**0.00 (0.00)**		

Item 3 did not exhibit DIF using the B-H multiple comparisons procedure

Items bolded were significant after BH-Adjustment

Table 65 Item parameters and standard errors for anchor items Uruguay by Cuba

Item	Content	*a*	*b*
1	Are you basically satisfied with your life	2.40 (0.14)	0.73 (0.04)
2	Have you dropped many of your activities and interests	1.58 (0.10)	0.58 (0.05)
3	Do you feel that your life is empty	2.47 (0.13)	0.47 (0.03)
5	Are you in good spirits most of the time	2.44 (0.15)	0.73 (0.04)
7	Do you feel happy most of the time	3.08 (0.18)	0.50 (0.03)
9	Do you prefer to stay at home, rather than going out and doing things	0.92 (0.06)	-0.44 (0.06)
10	Do you feel that you have more memory problems than most	1.15 (0.10)	1.52 (0.13)
11	Do you think that it is wonderful to be alive now	2.21 (0.18)	1.45 (0.08)
12	Do you feel worthless the way you are now	2.36 (0.16)	1.06 (0.06)
13	Do you feel full of energy	1.82 (0.11)	0.81 (0.06)
14	Do you feel that your situation is hopeless	2.61 (0.17)	0.94 (0.05)
15	Do you think that most people are better off than you are	1.21 (0.08)	0.56 (0.07)

Table 66 Item parameters and standard errors for items exhibiting DIF for Uruguay by Cuba

Item	Content	Group	*a*	*b*	Tests for DIF: $\chi 2$ (P) *a* Dif	Tests for DIF: $\chi 2$ (P) *b* Dif
4	Do you often get bored	Uruguay	2.51 (0.12)	0.39 (0.04)	**0.0 (1.000)**	**12.7 (.000)**
		Cuba	2.51 (0.12)	0.00 (0.00)		
6	Are you afraid that something bad is going to happen to you	Uruguay	1.54 (0.14)	0.65 (0.08)	**4.6 (.031)**	**8.9 (.002)**
		Cuba	1.52 (0.13)	0.67 (0.07)		
8	Do you often feel helpless	Uruguay	2.96 (0.27)	0.72 (0.05)	**20.4 (.000)**	**0.1 (.752)**
		Cuba	2.77 (0.25)	0.90 (0.05)		

Items bolded were significant after BH-Adjustment

Table 67 Summary of DIF analyses of the GDS-15: country by country anchor items

Item Content	Chile/Cuba Anchor Items	Mexico/ Uruguay Anchor Items	Argentina/ Mexico Anchor Items	Argentina/ Uruguay Anchor Items	Argentina/Chile Anchor Items
v1. Are you basically satisfied with your life	√	√	√	√	√
v2. Have you dropped many of your activities and interests	√	√			
v3. Do you feel that your life is empty	√		√	√	√
v4. Do you often get bored		√	√	√	√
v5. Are you in good spirits most of the time	√	√	√	√	√
v6. Are you afraid that something bad is going to happen to you	√		√		√
v7. Do you feel happy most of the time	√				
v8. Do you often feel helpless		√	√	√	√
v9. Do you prefer to stay at home, rather than going out and doing things	√			√	√
v10. Do you feel that you have more problems with memory than most	√	√	√	√	√
v11. Do you think it is wonderful to be alive now	√	√			√
v12. Do you feel worthless the way you are now	√			√	√
v13. Do you feel full of energy	√	√	√	√	√
v14. Do you feel that your situation is hopeless	√	√	√	√	√
v15. Do you think that most people are better off than you are	√	√	√	√	√

*** √ =Anchor Items by Country***

Table 68 Summary of DIF analyses of the GDS-15: country by country anchor items

Item Content	Argentina/ Cuba Anchor Items	Mexico/ Chile Anchor Items	Mexico/ Cuba Anchor Items	Uruguay / Chile Anchor Items	Uruguay/ Cuba Anchor Items
v1. Are you basically satisfied with your life	√	√	√	√	√
v2. Have you dropped many of your activities and interests		√			√
v3. Do you feel that your life is empty	√	√	√		√
v4. Do you often get bored	√	√		√	
v5. Are you in good spirits most of the time		√	√	√	√
v6. Are you afraid that something bad is going to happent to you	√	√	√	√	
v7. Do you feel happy most of the time				√	√
v8. Do you often feel helpless	√		√	√	
v9. Do you prefer to stay at home, rather than going out and doing things	√	√	√	√	√
v10. Do you feel that you have more problems with memory than most	√	√	√		√
v11. Do you think it is wonderful to be alive now	√			√	√
v12. Do you feel worthless the way you are now	√			√	√
v13. Do you feel full of energy	√	√	√	√	√
v14. Do you feel that your situation is hopeless	√	√			√
v15. Do you think that most people are better off than you are	√		√	√	√

** √ =*Anchor Items by Country***

Table 69 Summary of DIF analyses of the GDS-15: country by country type of DIF

Item Content	Chile/Cuba Type of DIF, if Present	Mexico/ Uruguay Type of DIF, if Present	Argentina/Mexico Type of DIF, if Present	Argentina/ Uruguay Type of DIF, if Present	Argentina/ Chile Type of DIF, if Present
v1. Are you basically satisfied with your life					
v2. Have you dropped many of your activities and interests			*U*	*U*	*U*
v3. Do you feel that your life is empty		*NU*			
v4. Do you often get bored	*U*				
v5. Are you in good spirits most of the time					
v6. Are you afraid that something bad is going to happen to you		*U*		*NU*	
v7. Do you feel happy most of the time		*U*	*NU*	*NU*	*NU*
v8. Do you often feel helpless	*U*				
v9. Do you prefer to stay at home, rather than going out and doing things		*U*	*U*		
v10. Do you feel that you have more problems with memory than most					
v11. Do you think it is wonderful to be alive now			*NU*	*NU*	
v12. Do you feel worthless the way you are now		*U*	*U*		
v13. Do you feel full of energy					
v14. Do you feel that your situation is hopeless					
v15. Do you think that most people are better off than you are					

NU=Non-Uniform DIF, U=Uniform DIF

Table 70 Summary of DIF analyses of the GDS-15: country by country type of DIF

Item Content	Argentina/Cuba Type of DIF, if Present	Mexico/ Chile Type of DIF, if Present	Mexico/ Cuba Type of DIF, if Present	Uruguay / Chile Type of DIF, if Present	Uruguay/ Cuba Type of DIF, if Present
v1. Are you basically satisfied with your life					
v2. Have you dropped many of your activities and interests	*U*		*U*	*NU*	
v3. Do you feel that your life is empty				*NU*	
v4. Do you often get bored			*U*		*U*
v5. Are you in good spirits most of the time	*NU*				
v6. Are you afraid that something bad is going to happent to you					*NU*
v7. Do you feel happy most of the time	*NU*	*U*	*U*		
v8. Do you often feel helpless		*NU*			*NU*
v9. Do you prefer to stay at home, rather than going out and doing things					
v10. Do you feel that you have more problems with memory than most				*NU*	
v11. Do you think it is wonderful to be alive now		*NU*	*NU*		
v12. Do you feel worthless the way you are now		*U*	*U*		
v13. Do you feel full of energy					
v14. Do you feel that your situation is hopeless			*U*	*U*	
v15. Do you think that most people are better off than you are		*U*			

NU=Non-Uniform DIF, U=Uniform DIF

Table 71 Summary of DIF analyses of the GDS-15: country by country BH-Adjustment

Item Content	Chile/Cuba DIF After BH Adjustment	Mexico/ Uruguay DIF After BH Adjustment	Argentina/ Mexico DIF After BH Adjustment	Argentina/ Uruguay DIF After BH Adjustment	Argentina/ Chile DIF After BH Adjustment
v1. Are you basically satisfied with your life					
v2. Have you dropped many of your activities and interests			*yes*	*yes*	*no*
v3. Do you feel that your life is empty		*no*			
v4. Do you often get bored	*yes*				
v5. Are you in good spirits most of the time					
v6. Are you afraid that something bad is going to happen to you		*yes*		*yes*	
v7. Do you feel happy most of the time		*yes*	*yes*	*yes*	*no*
v8. Do you often feel helpless	*yes*				
v9. Do you prefer to stay at home, rather than going out and doing things		*yes*	*yes*		
v10. Do you feel that you have more problems with memory than most					
v11. Do you think it is wonderful to be alive now			*yes*	*yes*	
v12. Do you feel worthless the way you are now		*yes*	*yes*		
v13. Do you feel full of energy					
v14. Do you feel that your situation is hopeless					
v15. Do you think that most people are better off than you are					

*** Yes=significant after Benjimini-Hochberg Adjustment, No=non-significant after adjustment****

Table 72 Summary of DIF analyses of the GDS-15: country by country BH-Adjustment

Item Content	Argentina/Cuba DIF After BH Adjustment	Mexico/ Chile DIF After BH Adjustment	Mexico Cuba DIF After BH Adjustment	Uruguay/ Chile DIF After BH Adjustment	Uruguay/ Cuba DIF After BH Adjustment
v1. Are you basically satisfied with your life					
v2. Have you dropped many of your activities and interests	*yes*		*yes*	*yes*	
v3. Do you feel that your life is empty				*no*	
v4. Do you often get bored			*yes*		*yes*
v5. Are you in good spirits most of the time	*yes*				
v6. Are you afraid that something bad is going to happent to you					*yes*
v7. Do you feel happy most of the time	*yes*	*yes*	*yes*		
v8. Do you often feel helpless		*yes*			*yes*
v9. Do you prefer to stay at home, rather than going out and doing things					
v10. Do you feel that you have more problems with memory than most				*yes*	
v11. Do you think it is wonderful to be alive now		*no*	*no*		
v12. Do you feel worthless the way you are now		*yes*	*yes*		
v13. Do you feel full of energy					
v14. Do you feel that your situation is hopeless			*yes*	*yes*	
v15. Do you think that most people are better off than you are		*no*			

*** Yes=significant after Benjimini-Hochberg Adjustment, No=non-significant after adjustment****

Table 73 *Descriptive statistics and correlations for the GDS-15 Argentina*

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
V1															
V2	0.60														
V3	0.68	0.54													
V4	0.47	0.51	0.78												
V5	0.65	0.36	0.50	0.46											
V6	0.45	0.50	0.46	0.37	0.24										
V7	0.76	0.46	0.63	0.50	0.83	0.26									
V8	0.64	0.47	0.75	0.61	0.54	0.45	0.70								
V9	0.40	0.55	0.45	0.48	0.27	0.34	0.26	0.46							
V10	0.26	0.32	0.33	0.37	0.28	0.32	0.27	0.39	0.32						
V11	0.64	0.16	0.39	0.39	0.73	0.23	0.79	0.55	0.12	0.14					
V12	0.61	0.64	0.65	0.60	0.46	0.43	0.50	0.64	0.52	0.42	0.35				
V13	0.56	0.32	0.28	0.35	0.64	0.20	0.72	0.43	0.29	0.34	0.68	0.40			
V14	0.67	0.55	0.61	0.48	0.47	0.42	0.53	0.67	0.41	0.42	0.49	0.69	0.37		
V15	0.41	0.28	0.34	0.40	0.27	0.29	0.18	0.39	0.36	0.32	0.05	0.43	0.23	0.37	
M	0.16	0.21	0.18	0.19	0.14	0.19	0.17	0.10	0.44	0.09	0.09	0.10	0.22	0.11	0.37
SD	0.37	0.41	0.38	0.40	0.34	0.39	0.38	0.30	0.50	0.29	0.28	0.31	0.41	0.32	0.48

Table 74 *Descriptive statistics and correlations for the GDS-15 Chile*

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
V1															
V2	0.52														
V3	0.65	0.37													
V4	0.63	0.45	0.76												
V5	0.60	0.46	0.62	0.69											
V6	0.40	0.37	0.51	0.43	0.48										
V7	0.76	0.46	0.65	0.67	0.72	0.55									
V8	0.62	0.41	0.75	0.66	0.61	0.59	0.71								
V9	0.16	0.23	0.27	0.26	0.17	0.21	0.20	0.27							
V10	0.41	0.39	0.44	0.47	0.50	0.34	0.39	0.44	0.25						
V11	0.67	0.42	0.52	0.57	0.58	0.21	0.63	0.53	0.24	0.51					
V12	0.58	0.54	0.62	0.64	0.59	0.41	0.60	0.72	0.30	0.57	0.63				
V13	0.58	0.53	0.44	0.51	0.62	0.43	0.67	0.57	0.27	0.52	0.63	0.64			
V14	0.62	0.50	0.62	0.57	0.60	0.47	0.62	0.70	0.34	0.51	0.62	0.78	0.62		
V15	0.39	0.21	0.38	0.31	0.39	0.30	0.43	0.40	0.11	0.40	0.40	0.42	0.33	0.47	
M	0.14	0.33	0.33	0.34	0.18	0.32	0.20	0.19	0.61	0.20	0.10	0.17	0.24	0.20	0.48
SD	0.35	0.47	0.47	0.47	0.38	0.46	0.40	0.39	0.48	0.40	0.30	0.38	0.42	0.40	0.49

Table 75 *Descriptive statistics and correlations for the GDS-15 Mexico*

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
V1															
V2	0.52														
V3	0.68	0.51													
V4	0.59	0.52	0.70												
V5	0.64	0.40	0.54	0.63											
V6	0.35	0.36	0.46	0.53	0.37										
V7	0.72	0.41	0.56	0.59	0.78	0.50									
V8	0.57	0.48	0.73	0.67	0.54	0.60	0.62								
V9	0.32	0.43	0.49	0.46	0.24	0.40	0.28	0.51							
V10	0.30	0.36	0.45	0.46	0.33	0.40	0.27	0.52	0.43						
V11	0.56	0.28	0.53	0.41	0.50	0.27	0.61	0.52	0.19	0.34					
V12	0.53	0.50	0.65	0.54	0.39	0.44	0.47	0.66	0.40	0.50	0.58				
V13	0.58	0.48	0.52	0.46	0.58	0.31	0.70	0.48	0.31	0.29	0.57	0.58			
V14	0.55	0.46	0.67	0.59	0.42	0.50	0.51	0.71	0.43	0.51	0.51	0.70	0.46		
V15	0.39	0.39	0.42	0.39	0.33	0.36	0.38	0.50	0.40	0.47	0.35	0.52	0.37	0.60	
M	0.11	0.25	0.24	0.24	0.14	0.30	0.14	0.19	0.46	0.22	0.05	0.15	0.18	0.17	0.31
SD	0.32	0.43	0.43	0.42	0.35	0.46	0.35	0.40	0.50	0.41	0.22	0.35	0.38	0.38	0.46

Table 76 *Descriptive statistics and correlations for the GDS-15 Cuba*

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
V1															
V2	0.56														
V3	0.70	0.54													
V4	0.74	0.59	0.81												
V5	0.65	0.48	0.63	0.71											
V6	0.49	0.48	0.50	0.57	0.49										
V7	0.77	0.47	0.70	0.71	0.83	0.54									
V8	0.67	0.47	0.75	0.70	0.62	0.59	0.70								
V9	0.29	0.43	0.41	0.42	0.34	0.34	0.32	0.32							
V10	0.29	0.37	0.41	0.41	0.44	0.32	0.39	0.45	0.33						
V11	0.66	0.44	0.61	0.60	0.63	0.30	0.60	0.60	0.14	0.40					
V12	0.58	0.56	0.58	0.65	0.57	0.46	0.54	0.62	0.36	0.52	0.68				
V13	0.56	0.51	0.48	0.63	0.63	0.43	0.65	0.55	0.36	0.51	0.62	0.73			
V14	0.58	0.51	0.68	0.65	0.59	0.52	0.62	0.71	0.35	0.41	0.56	0.72	0.64		
V15	0.53	0.48	0.49	0.49	0.47	0.44	0.57	0.53	0.32	0.43	0.36	0.52	0.47	0.58	
M	0.18	0.24	0.26	0.26	0.18	0.24	0.23	0.14	0.52	0.15	0.05	0.13	0.16	0.13	0.26
SD	0.39	0.42	0.44	0.44	0.38	0.43	0.42	0.34	0.50	0.36	0.22	0.33	0.37	0.33	0.44

Table 77 *Descriptive statistics and correlations for the GDS-15 Uruguay*

	V1	V2	V3	V4	V5	V6	V7	V8	V9	V10	V11	V12	V13	V14	V15
V1															
V2	0.55														
V3	0.58	0.49													
V4	0.48	0.55	0.72												
V5	0.66	0.41	0.53	0.57											
V6	0.48	0.50	0.37	0.45	0.53										
V7	0.75	0.50	0.63	0.58	0.84	0.57									
V8	0.66	0.57	0.70	0.60	0.61	0.63	0.74								
V9	0.26	0.42	0.34	0.41	0.28	0.31	0.26	0.39							
V10	0.31	0.36	0.40	0.35	0.29	0.42	0.30	0.34	0.30						
V11	0.69	0.40	0.54	0.51	0.70	0.37	0.72	0.65	0.26	0.47					
V12	0.45	0.55	0.62	0.57	0.48	0.49	0.61	0.73	0.43	0.48	0.59				
V13	0.48	0.47	0.38	0.45	0.60	0.37	0.69	0.45	0.39	0.33	0.59	0.55			
V14	0.62	0.49	0.63	0.54	0.55	0.56	0.70	0.69	0.38	0.43	0.67	0.71	0.59		
V15	0.34	0.38	0.33	0.30	0.30	0.24	0.40	0.35	0.34	0.37	0.38	0.44	0.43	0.54	
M	0.16	0.26	0.19	0.24	0.16	0.23	0.18	0.15	0.49	0.12	0.08	0.09	0.21	0.12	0.30
SD	0.37	0.44	0.39	0.43	0.36	0.42	0.38	0.36	0.50	0.33	0.28	0.29	0.41	0.33	0.46

Table 78 *GDS-15 cutoff scores by country*

Country	GDS<6	GDS>6
Argentina	85%	15%
Chile	73%	27%
Cuba	80%	20%
Mexico	80%	20%
Uruguay	85%	15%
Across all countries	81%	19%

***Scores below 6 indicate no depression, scores greater than 6 indicate depression

Table 79 *GDS-15 cutoff scores by gender and country*

Country	GDS <6		GDS ≥6	
	Women	Men	Women	Men
Argentina	84%	87%	16%	13%
Chile	70%	79%	30%	21%
Cuba	75%	88%	25%	12%
Mexico	79%	85%	21%	15%
Uruguay	82%	90%	18%	9%
Across all countries	77%	86%	23%	14%

****Scores below 6 indicate no depression, scores greater than 6 indicate depression*

Table 80 *Countries with the most difficulty endorsing items*

Country	Item	Content
Mexico	2	have you dropped many of your activities and interests
	4	do you often get bored
	6	are you afraid something bad is going to happen to you
	7	do you not feel happy most of the time
	12	do you feel worthless the way you are now
Argentina	2	have you dropped many of your activities and interests
	6	are you afraid something bad is going to happen to you
	7	do you not feel happy most of the time
	9	do you prefer to stay at home, rather than going and doing things
	11	do you not think it is wonderful to be alive now

Figure 1 *Chile scree-plot*

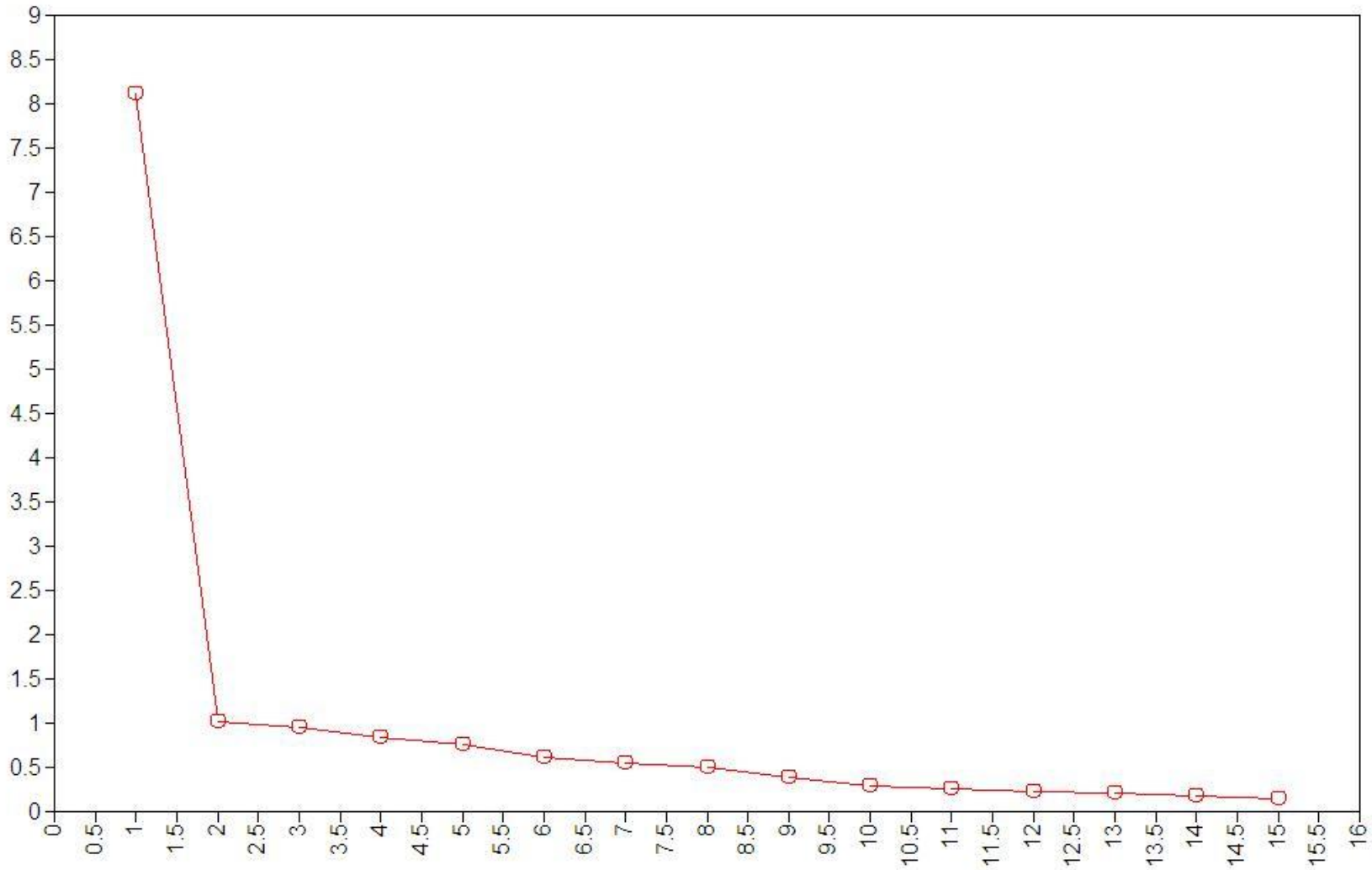


Figure 2 *Cuba scree-plot*

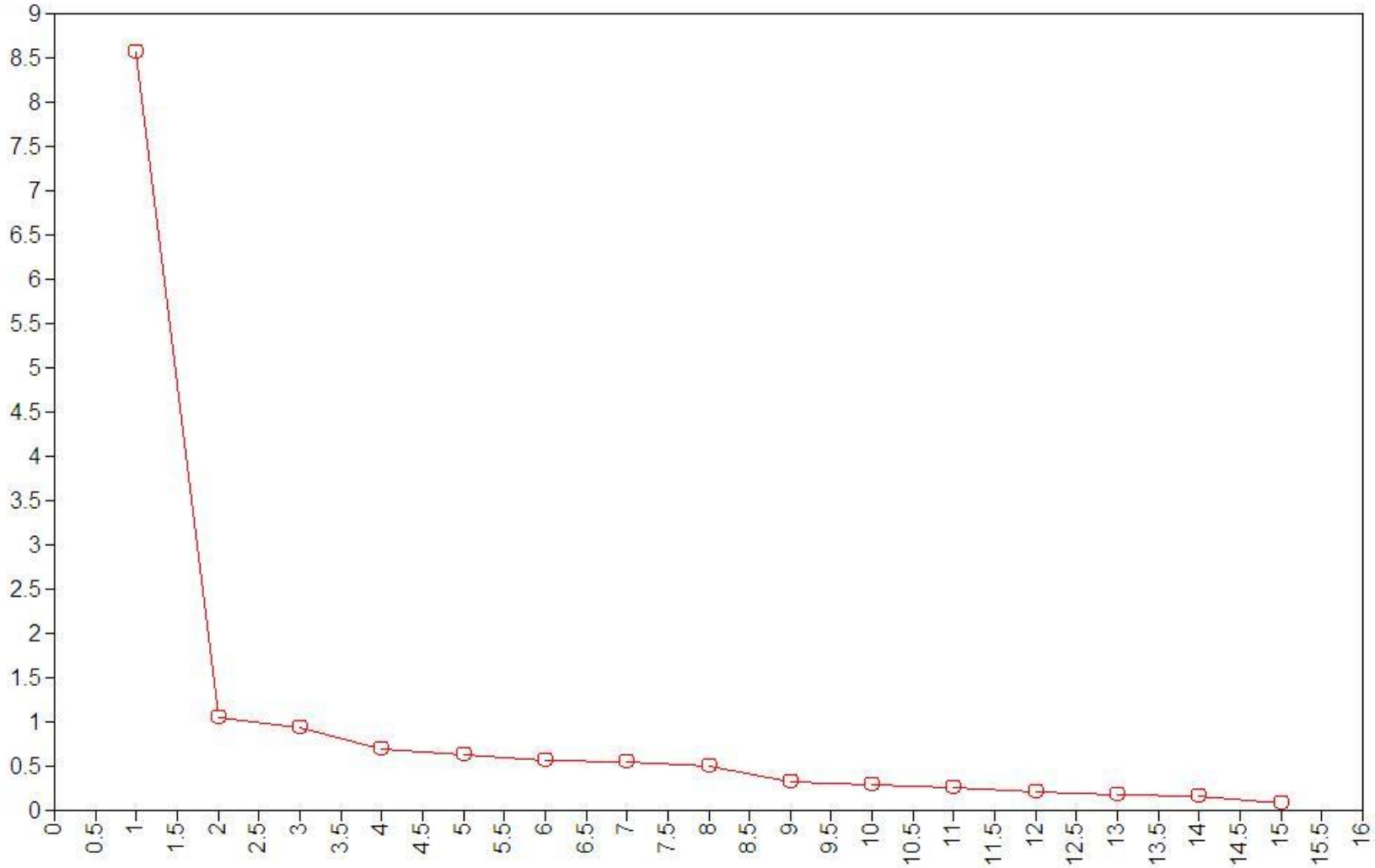


Figure 3 *Argentina scree-plot*

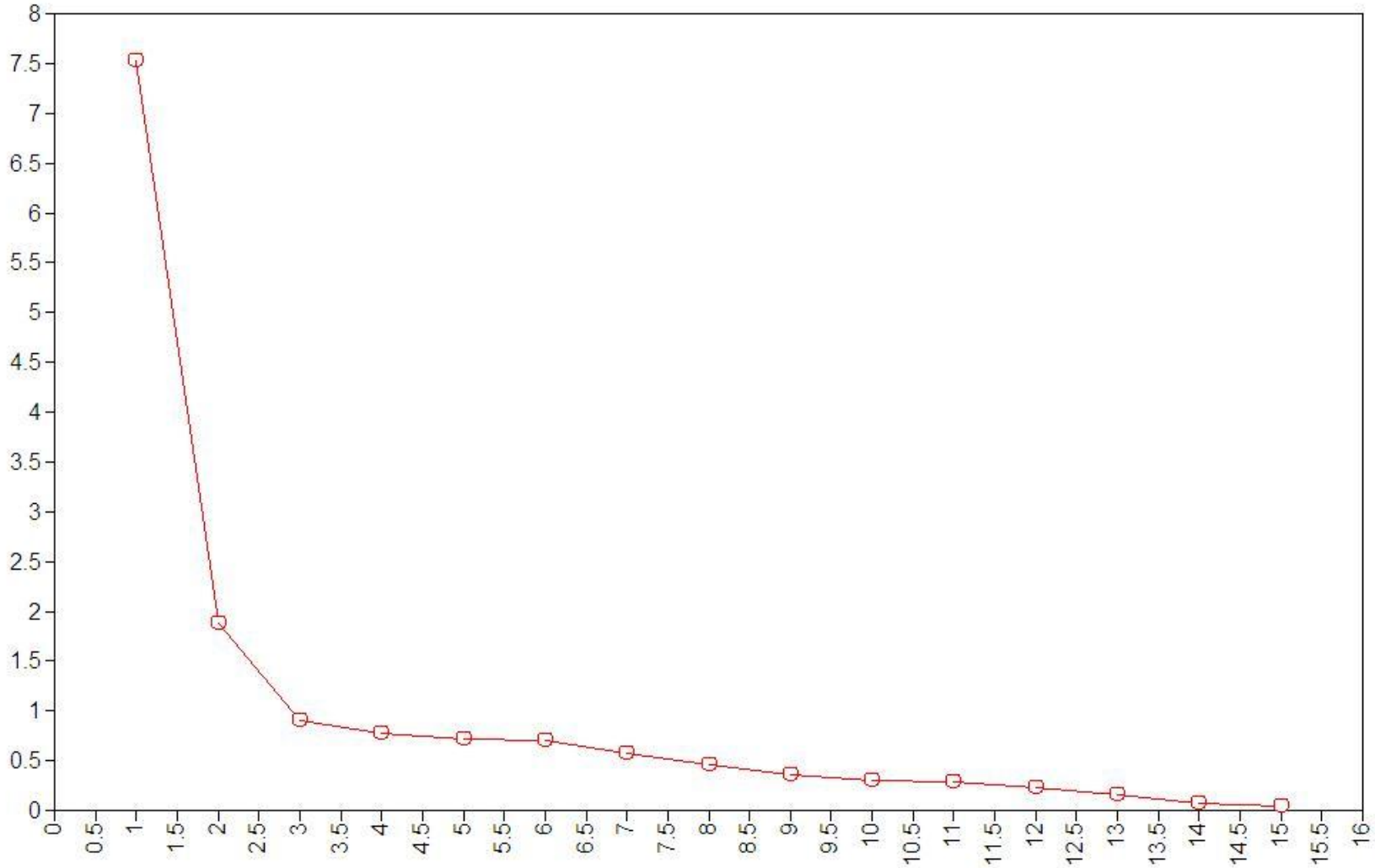


Figure 4 *Mexico scree-plot*

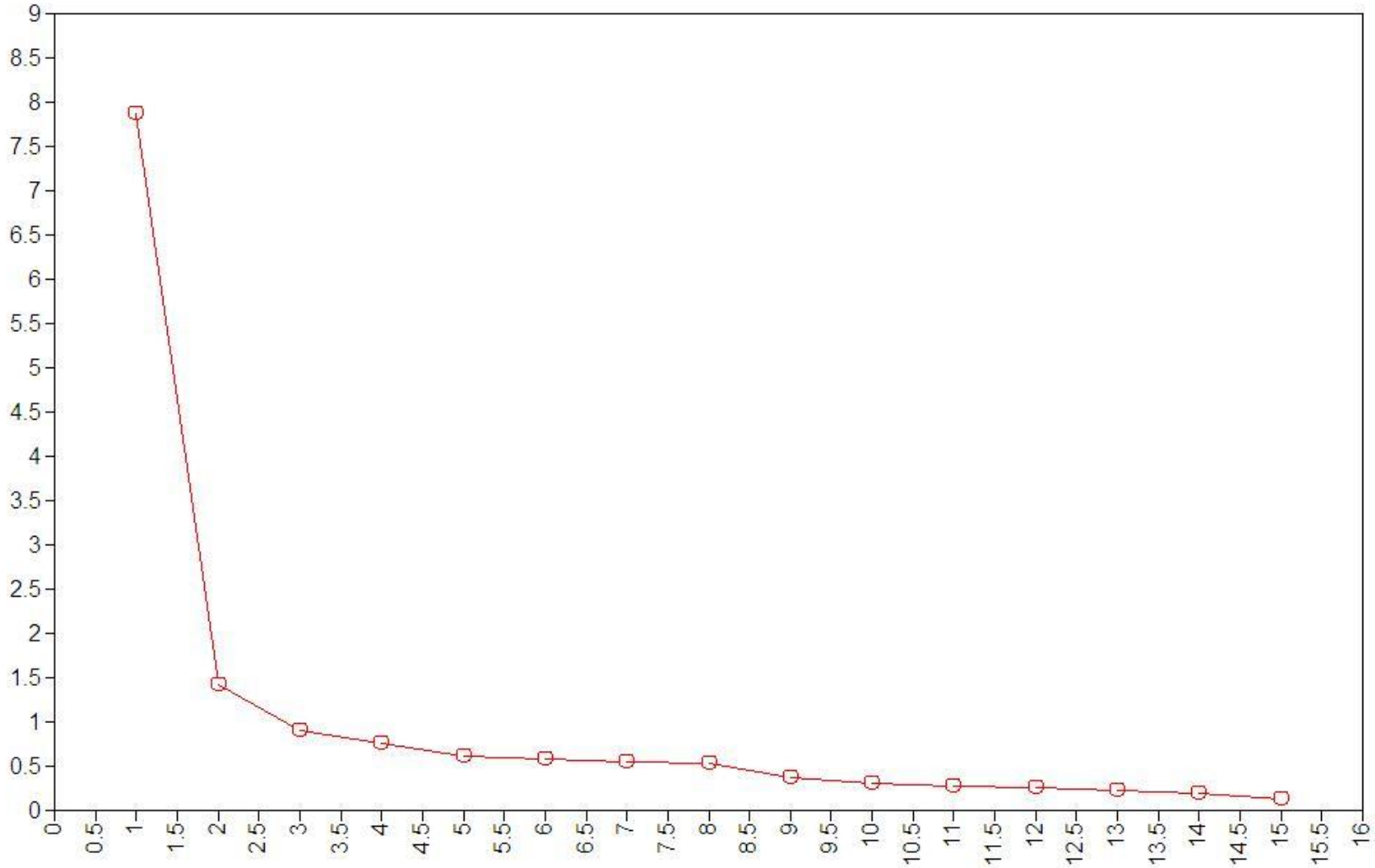


Figure 5 *Uruguay scree-plot*

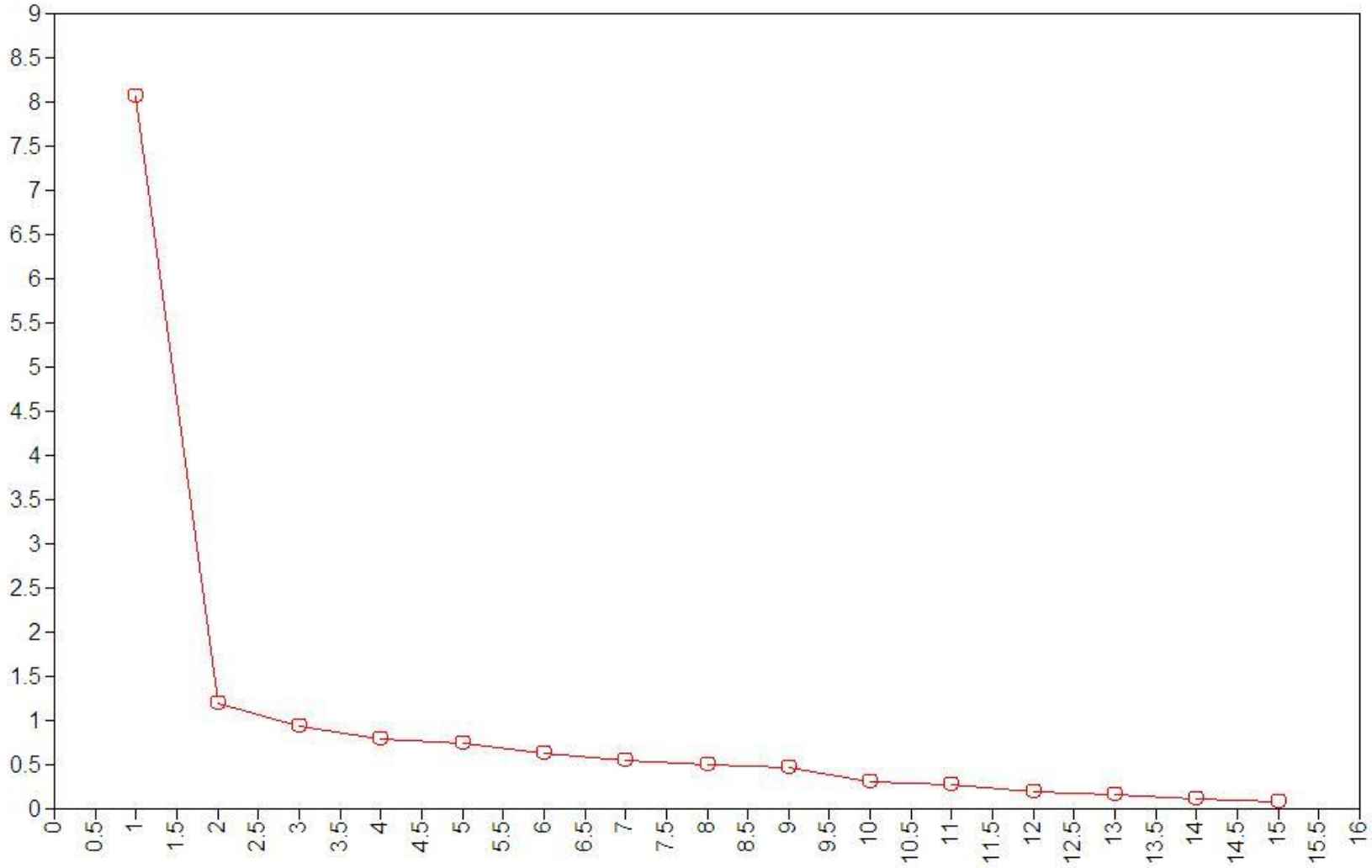


Figure 6 Chile by gender

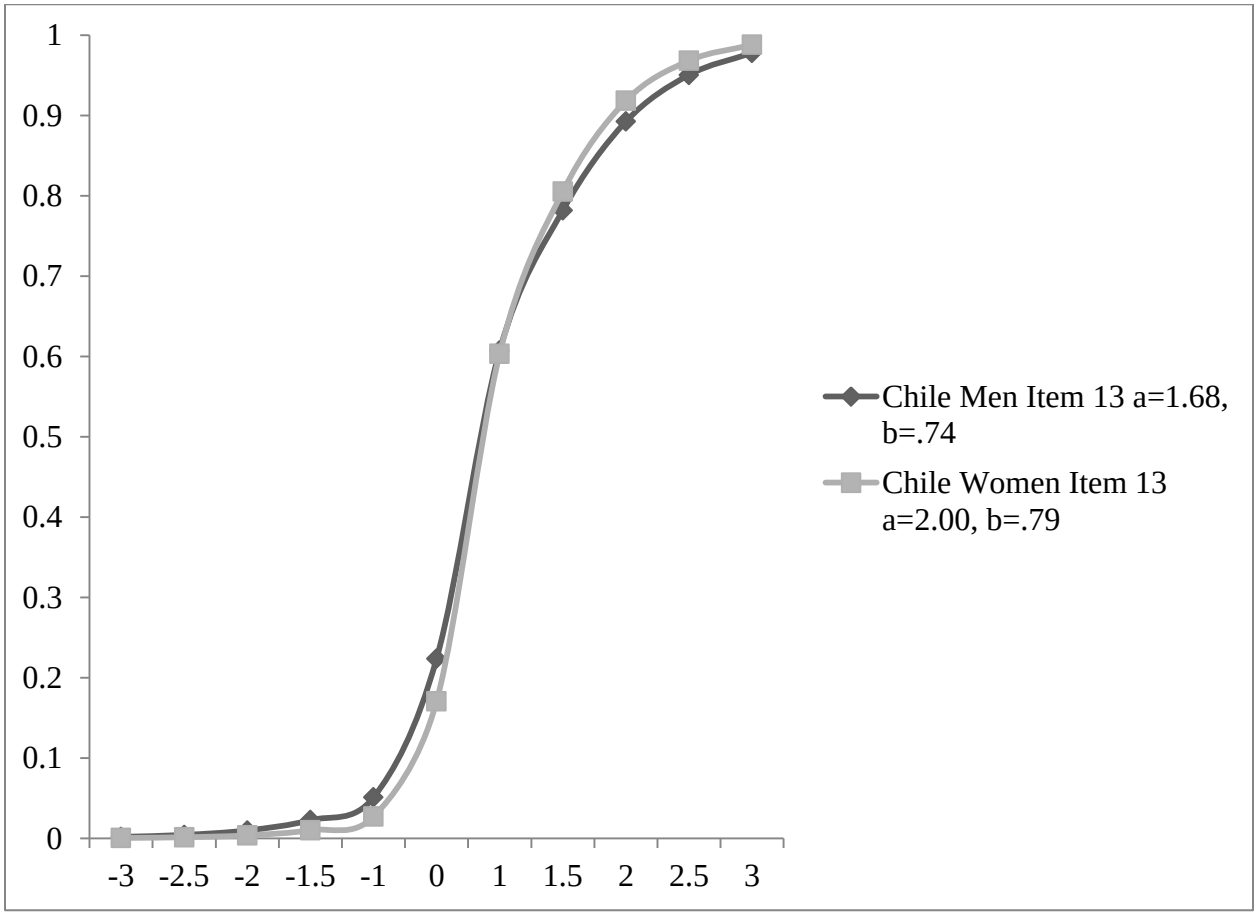


Figure 7 *Test information curve Chile by gender:* ***for interpretation of the references to color in this and all other figures, the reader is referred to the electronic version of this dissertation***

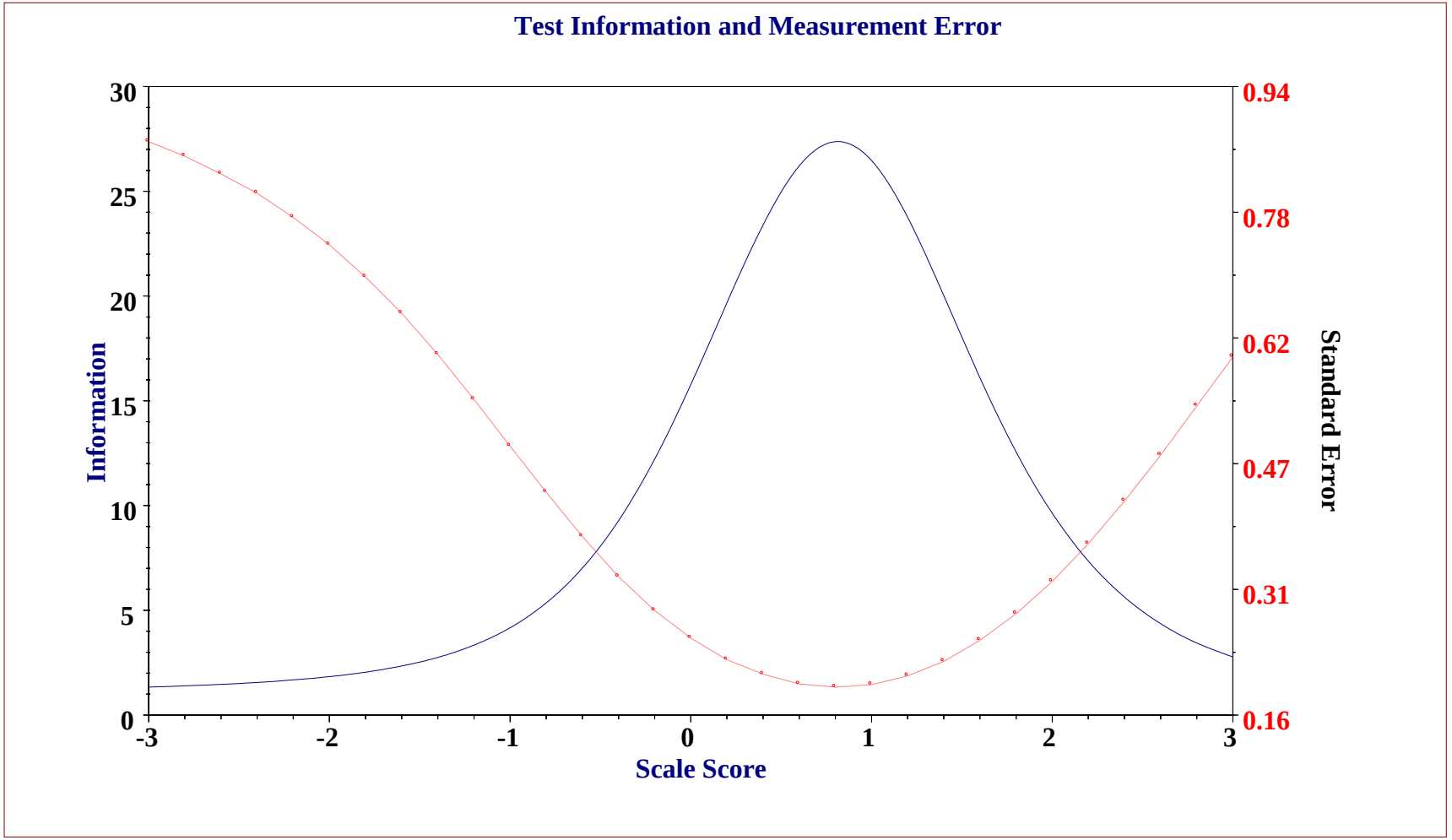


Figure 8 Cuba by gender item 7

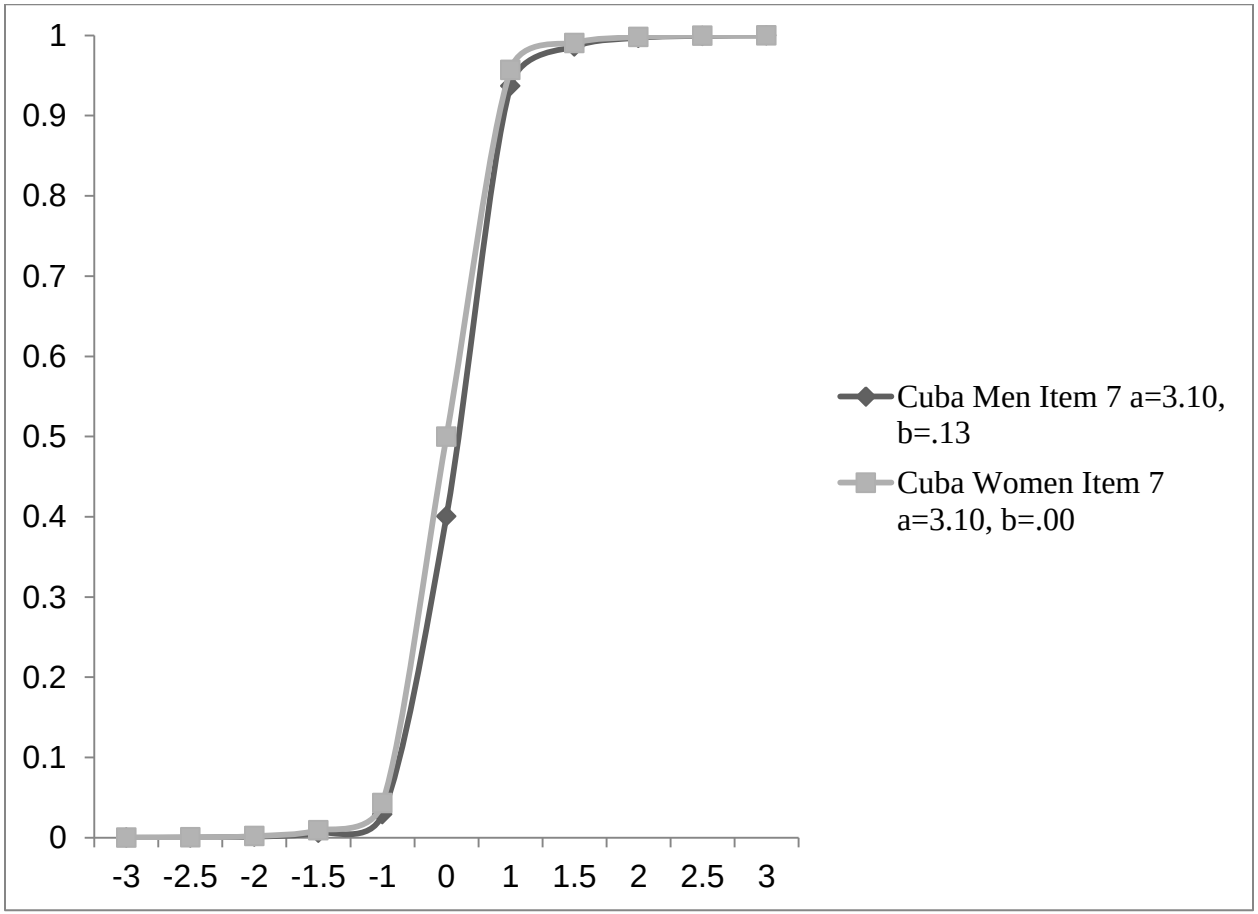


Figure 9 Cuba by gender item 12

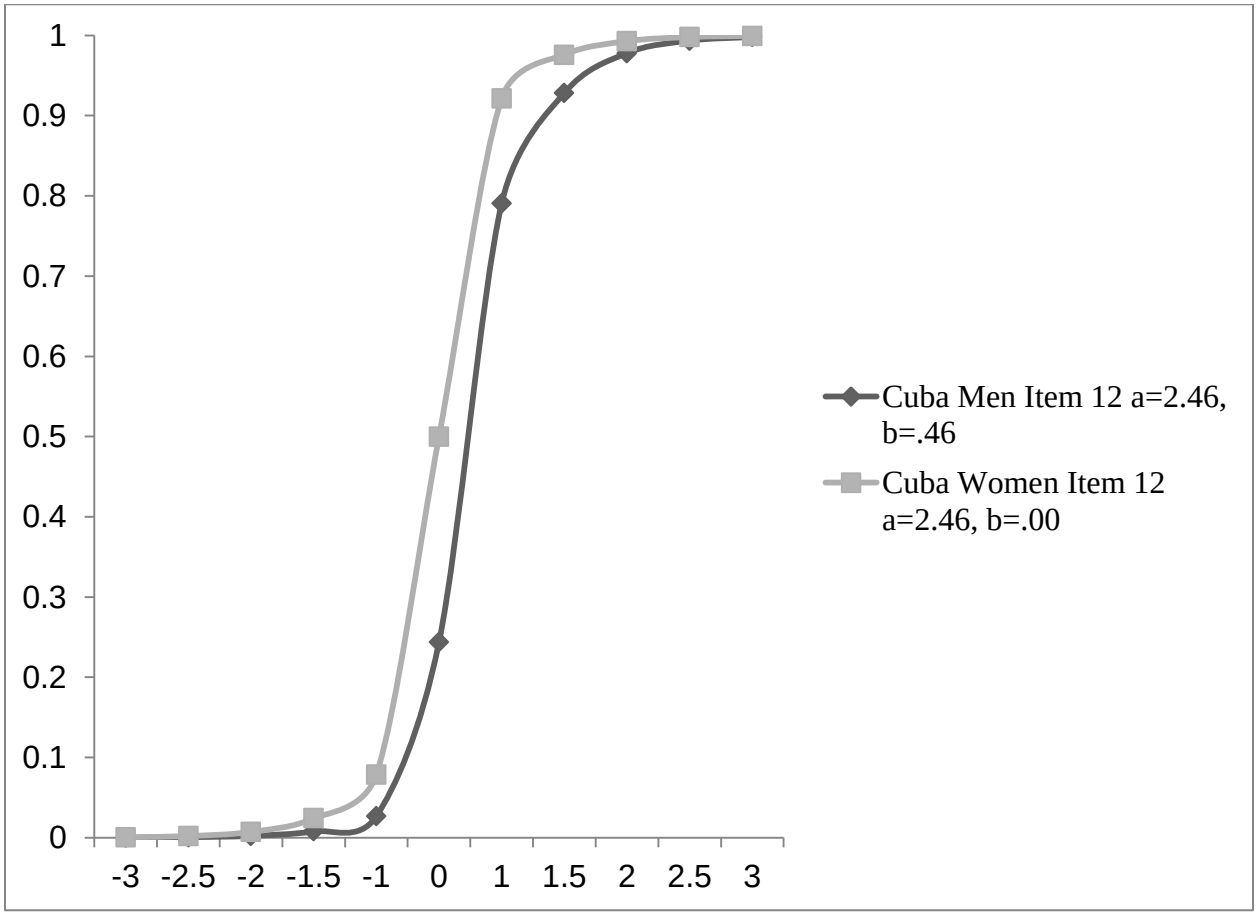


Figure 10 Test information curve Cuba by gender

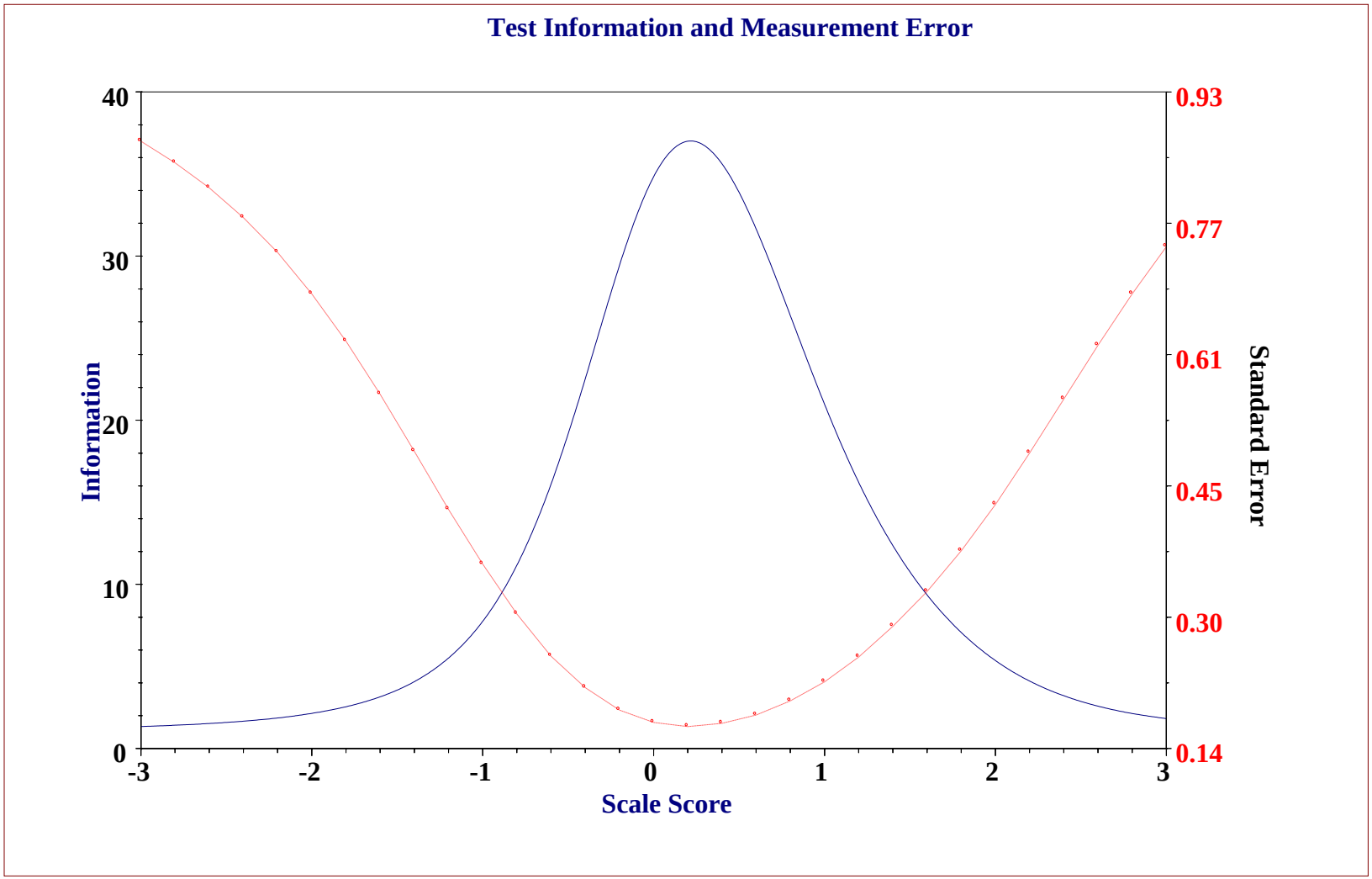


Figure 11 Argentina by gender item 9

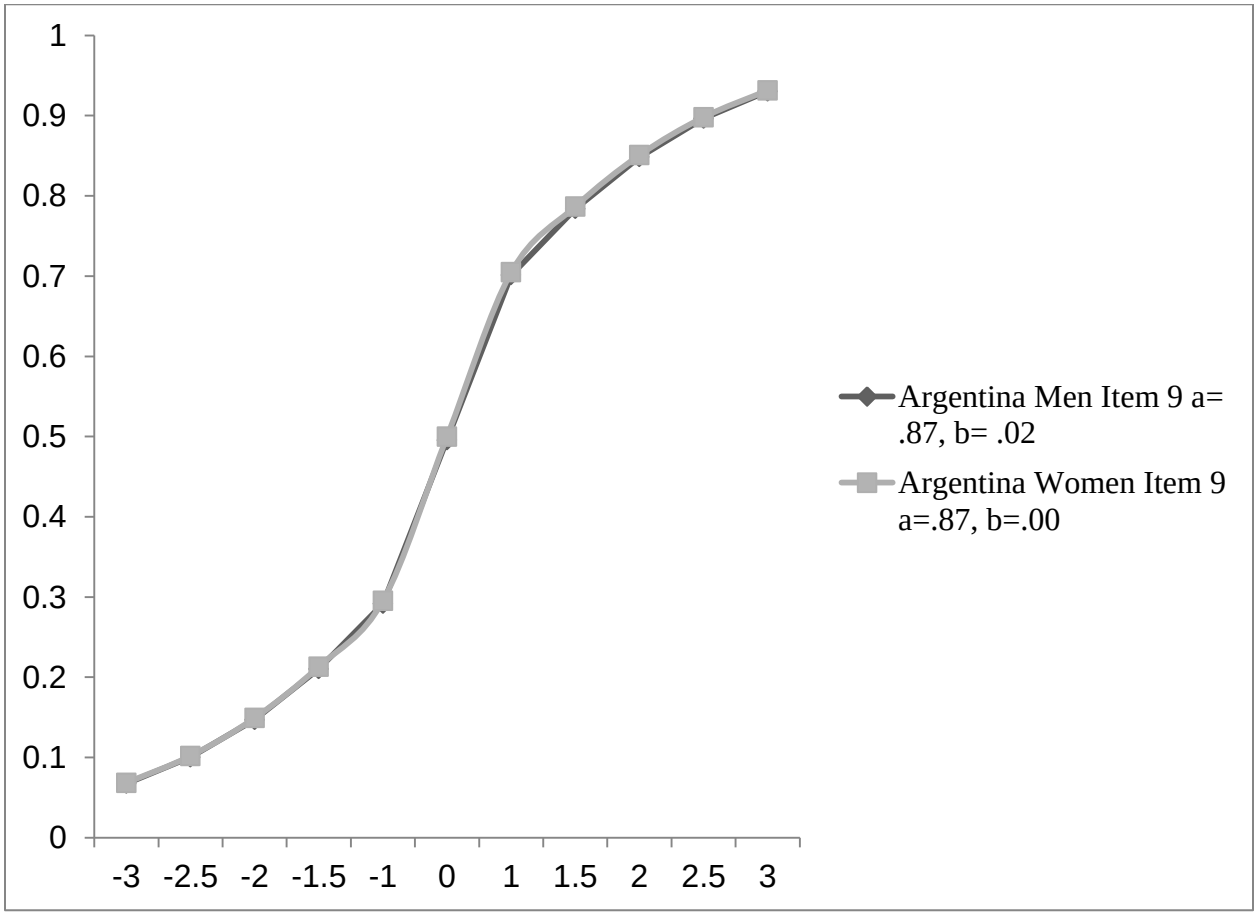


Figure 12 Argentina by gender item 11

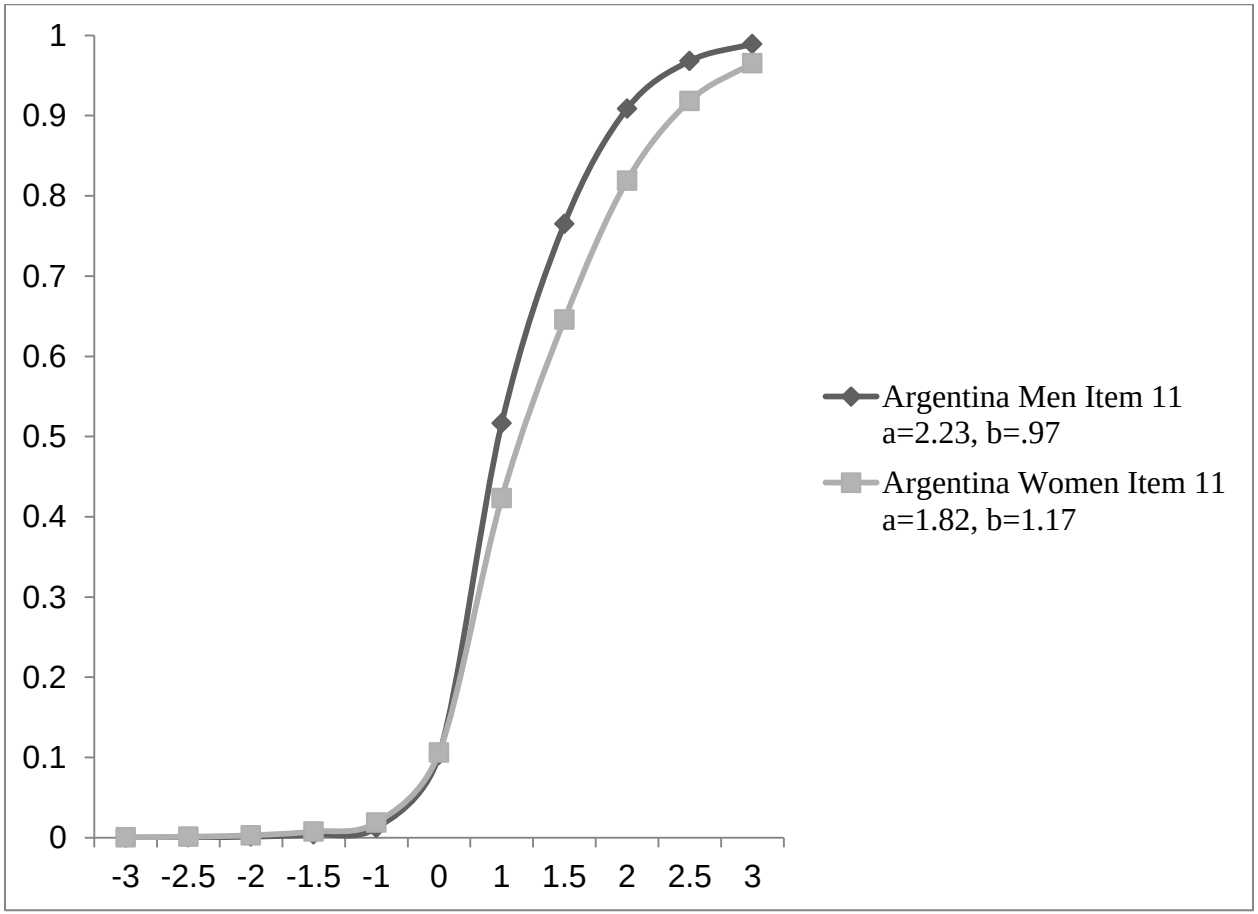


Figure 13 Argentina by gender item 12

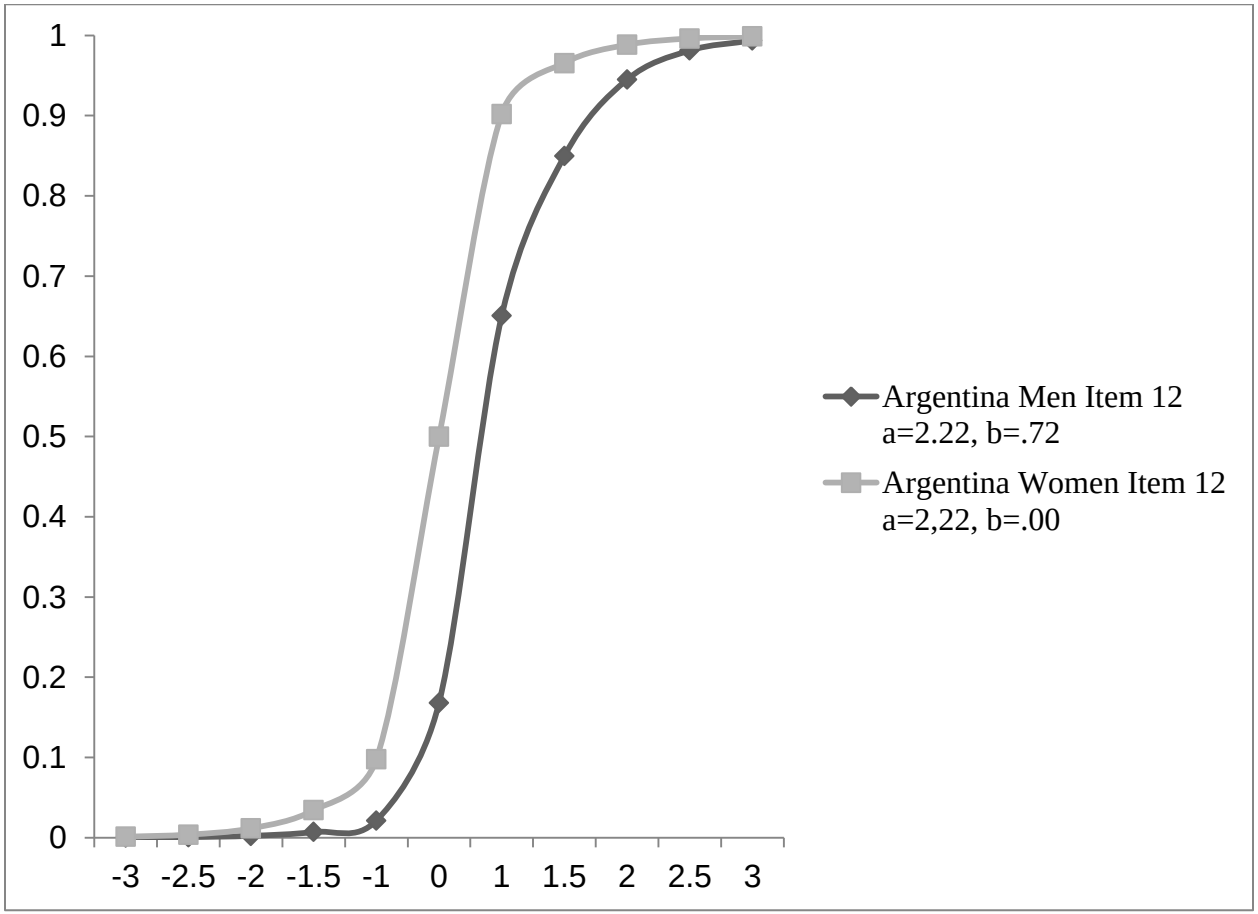


Figure 14 Argentina by gender item 15

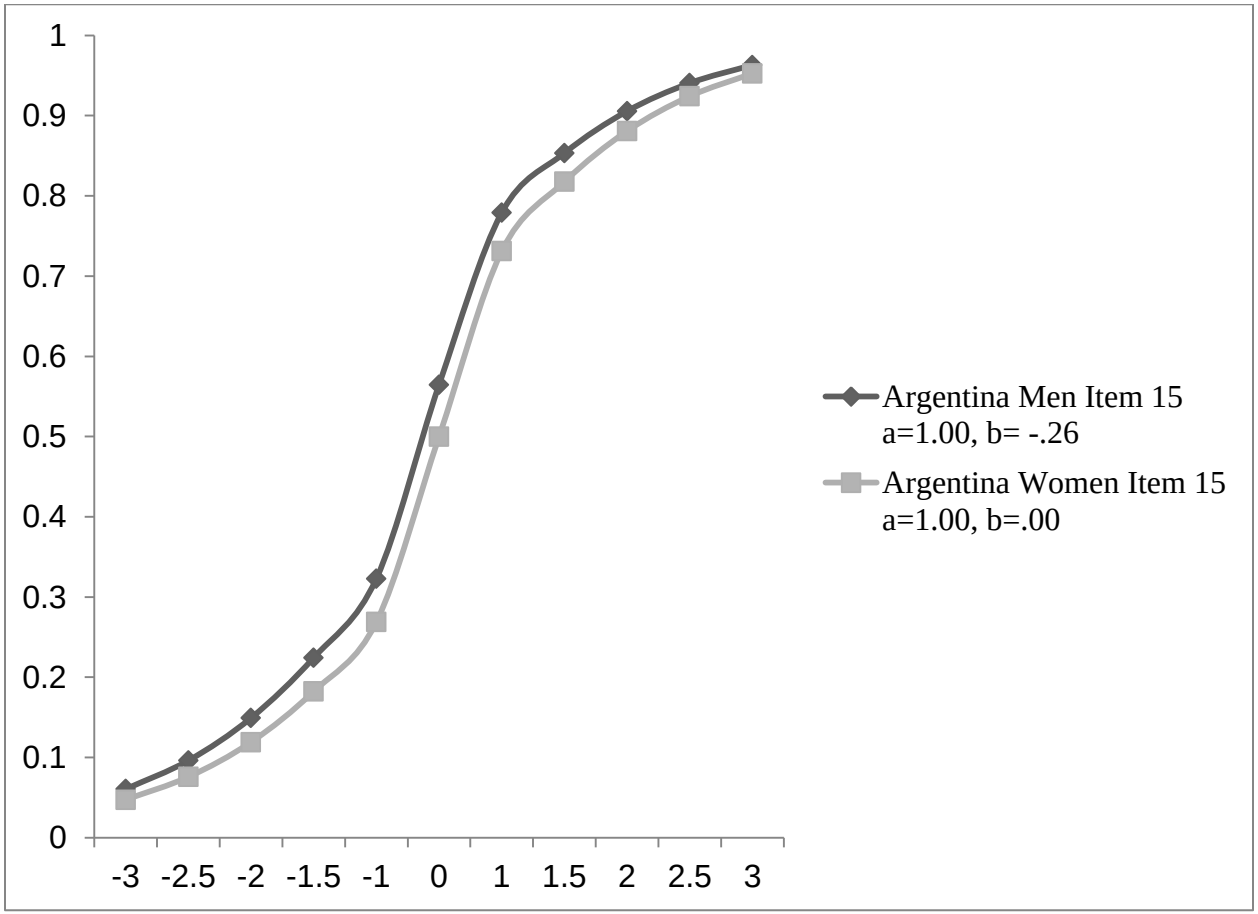


Figure 15 Test information curve Argentina by gender

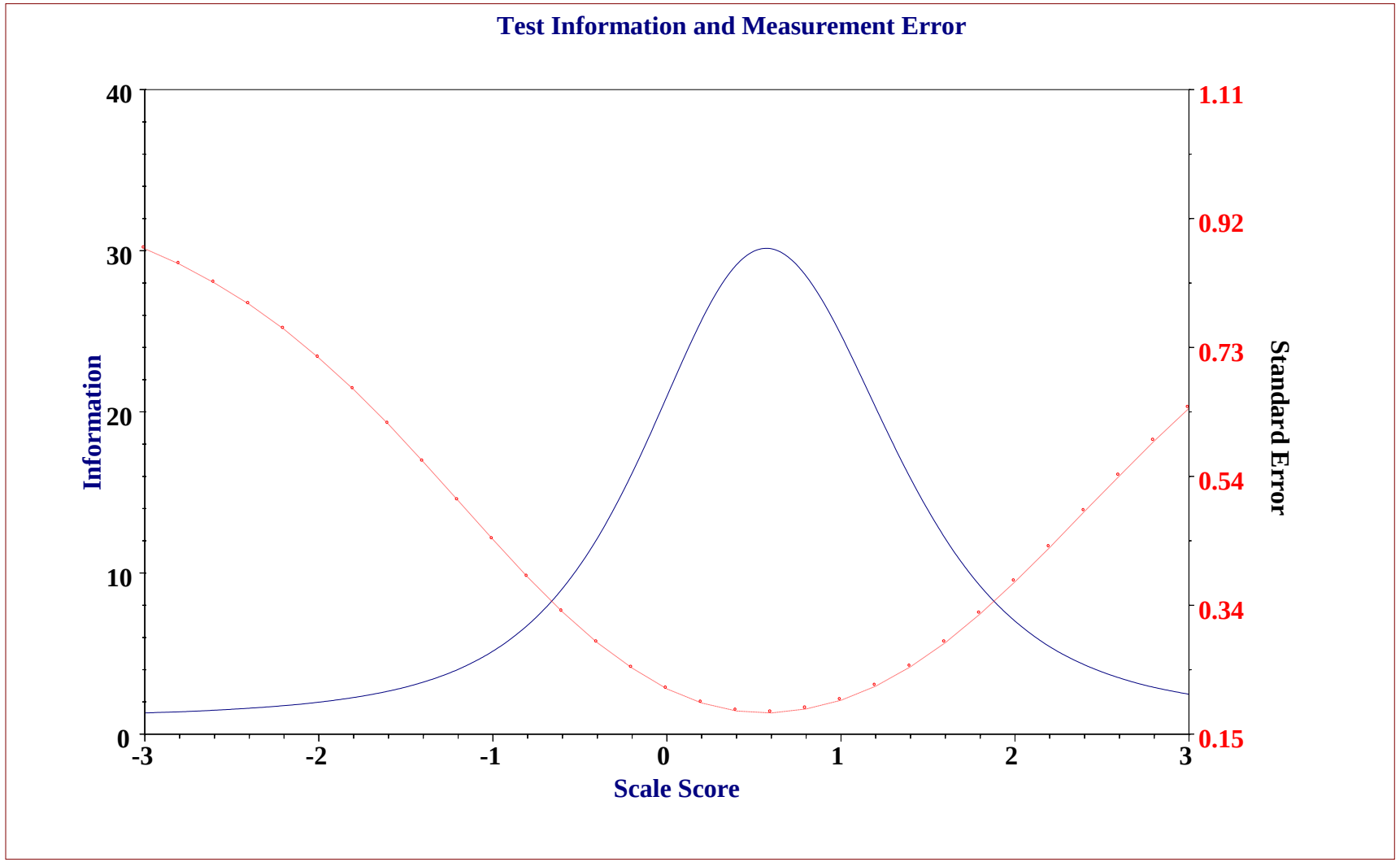


Figure 16 Mexico by gender item 6

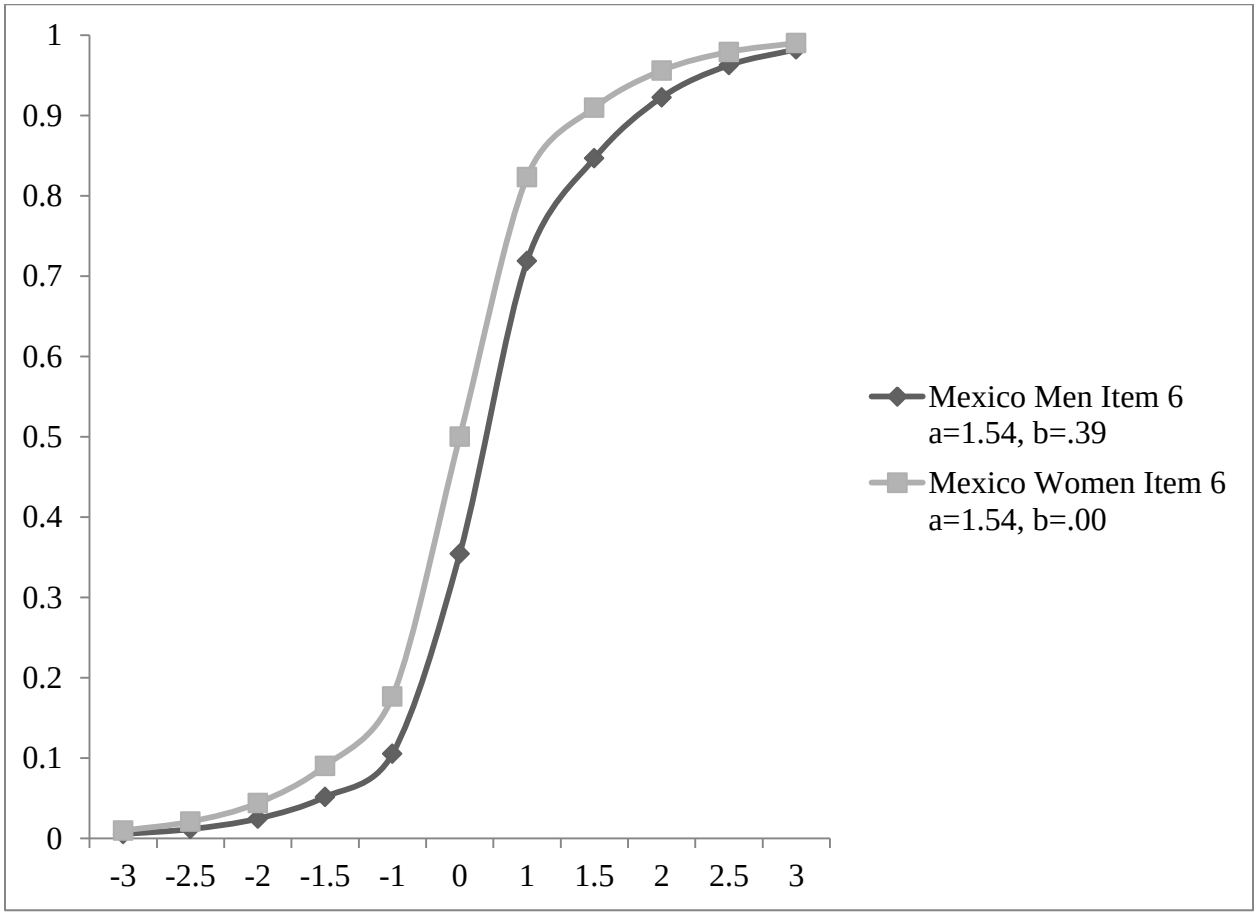


Figure 17 Mexico by gender item 8

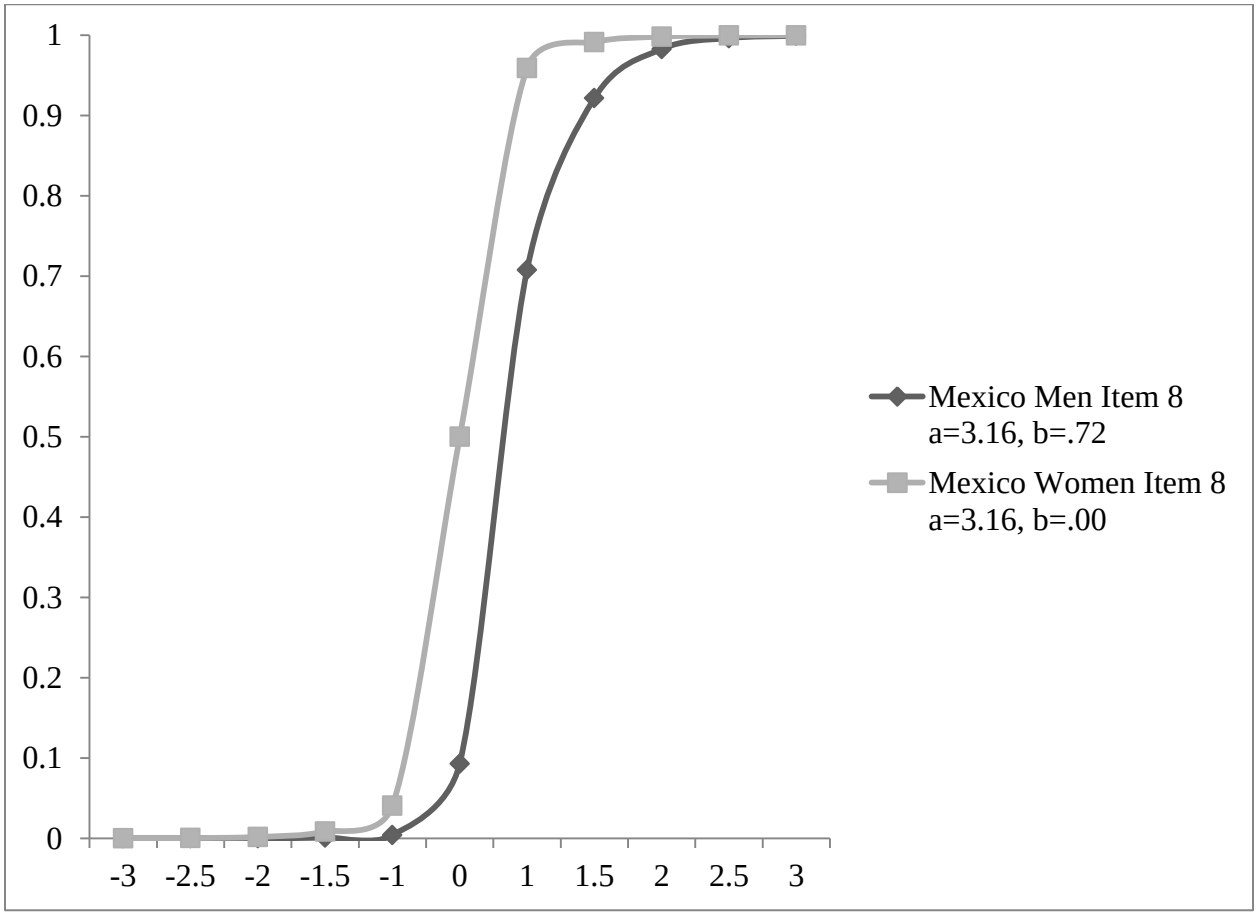


Figure 18 Test information curve Mexico by gender

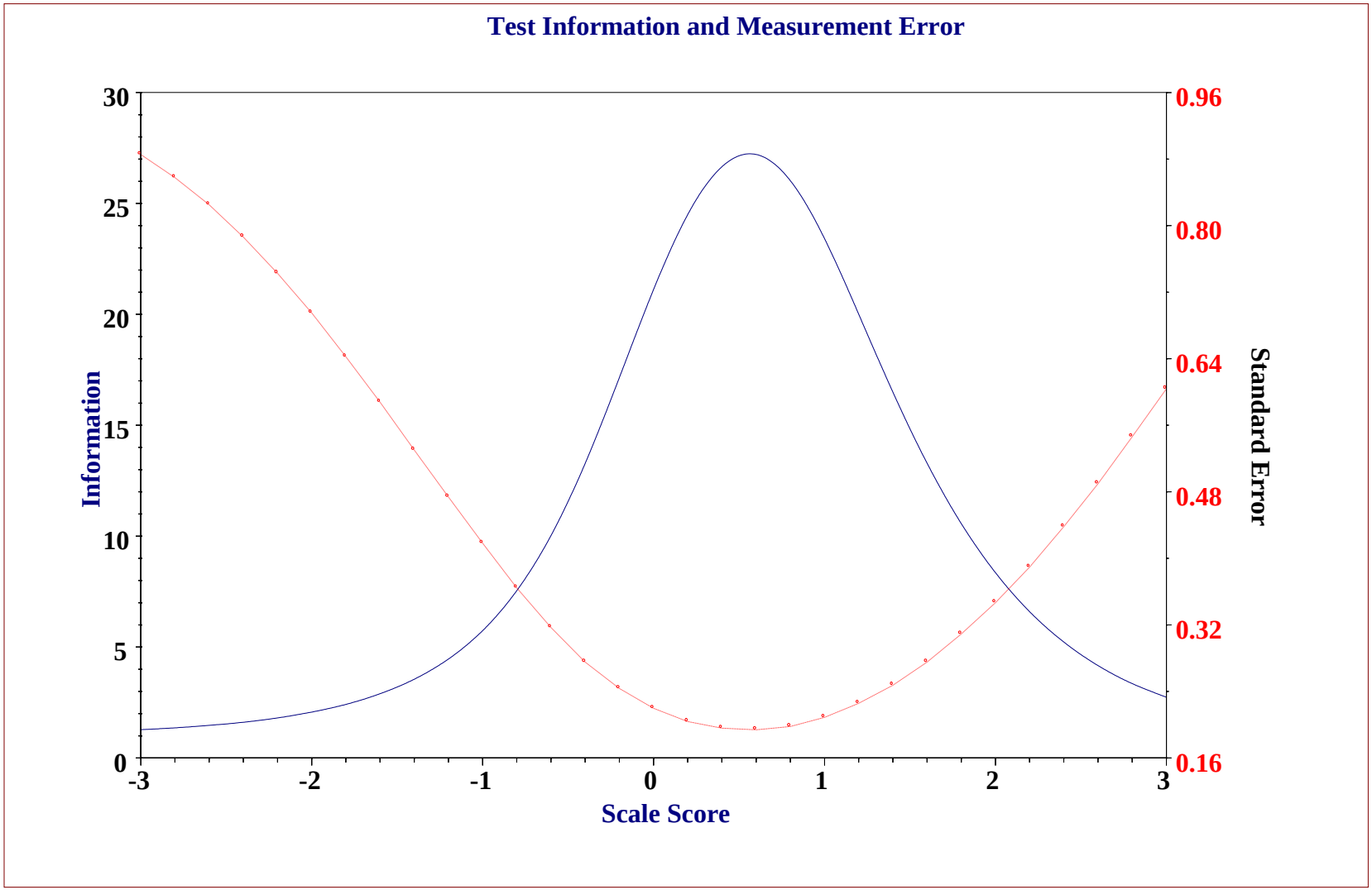


Figure 19 Uruguay by gender item 5

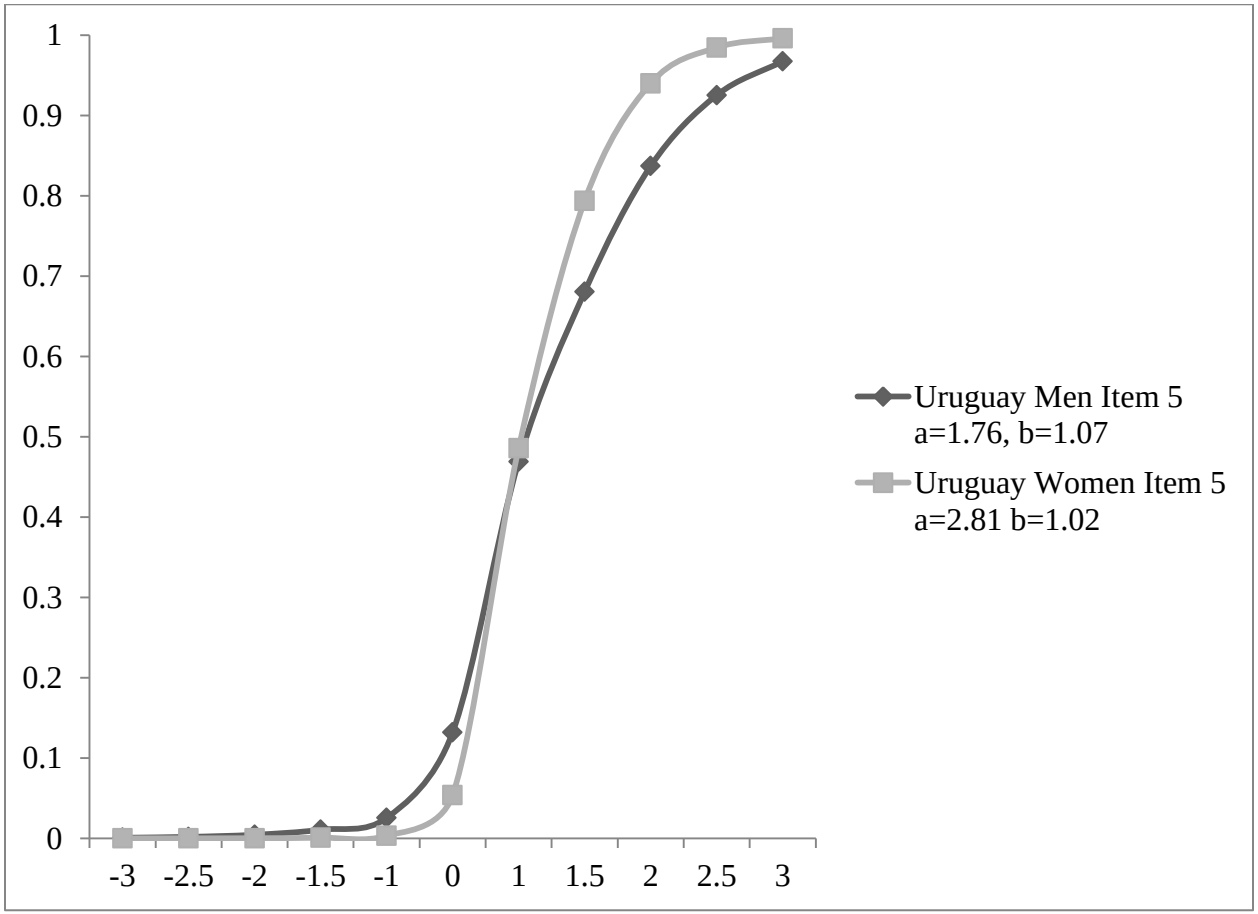


Figure 20 Uruguay by gender item 6

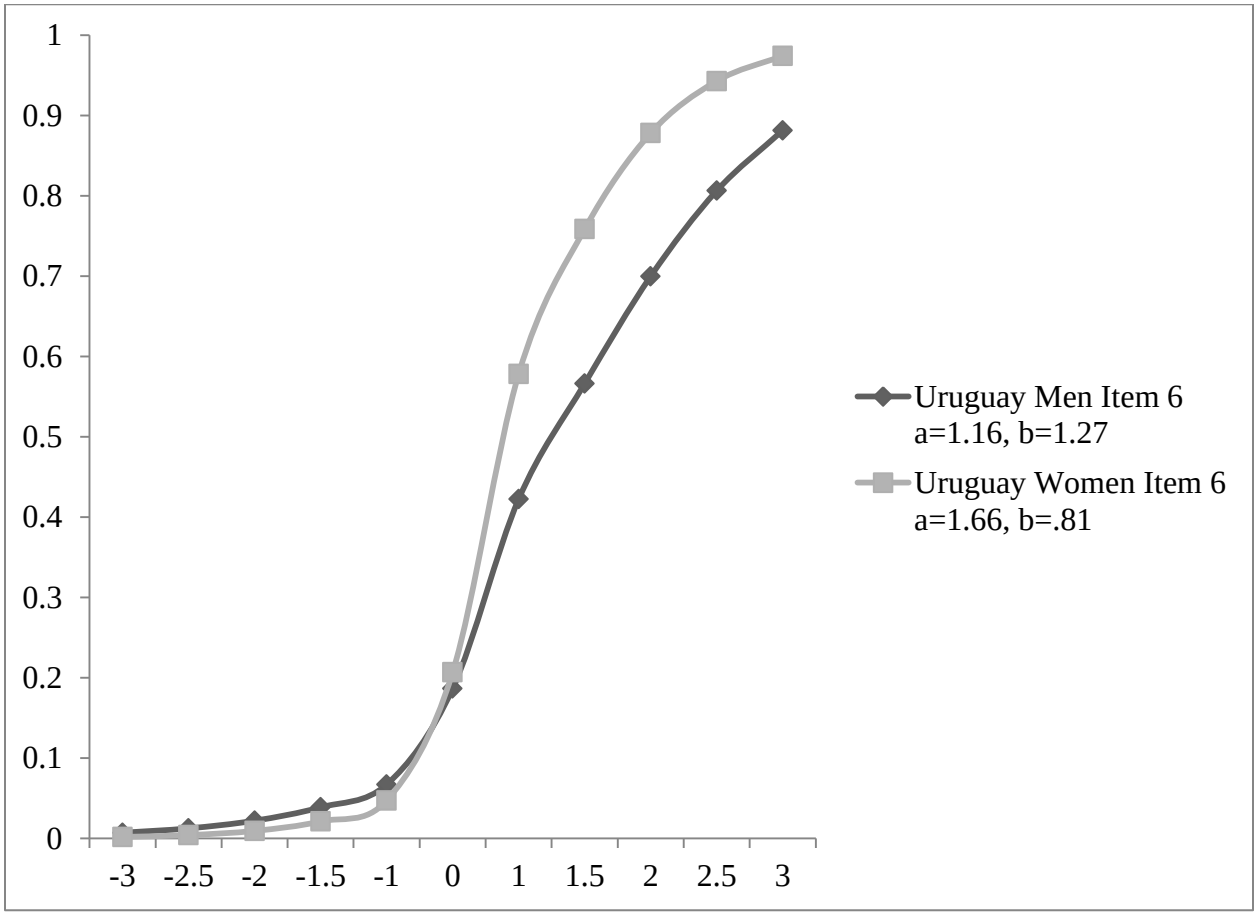


Figure 21 Uruguay by gender item 12

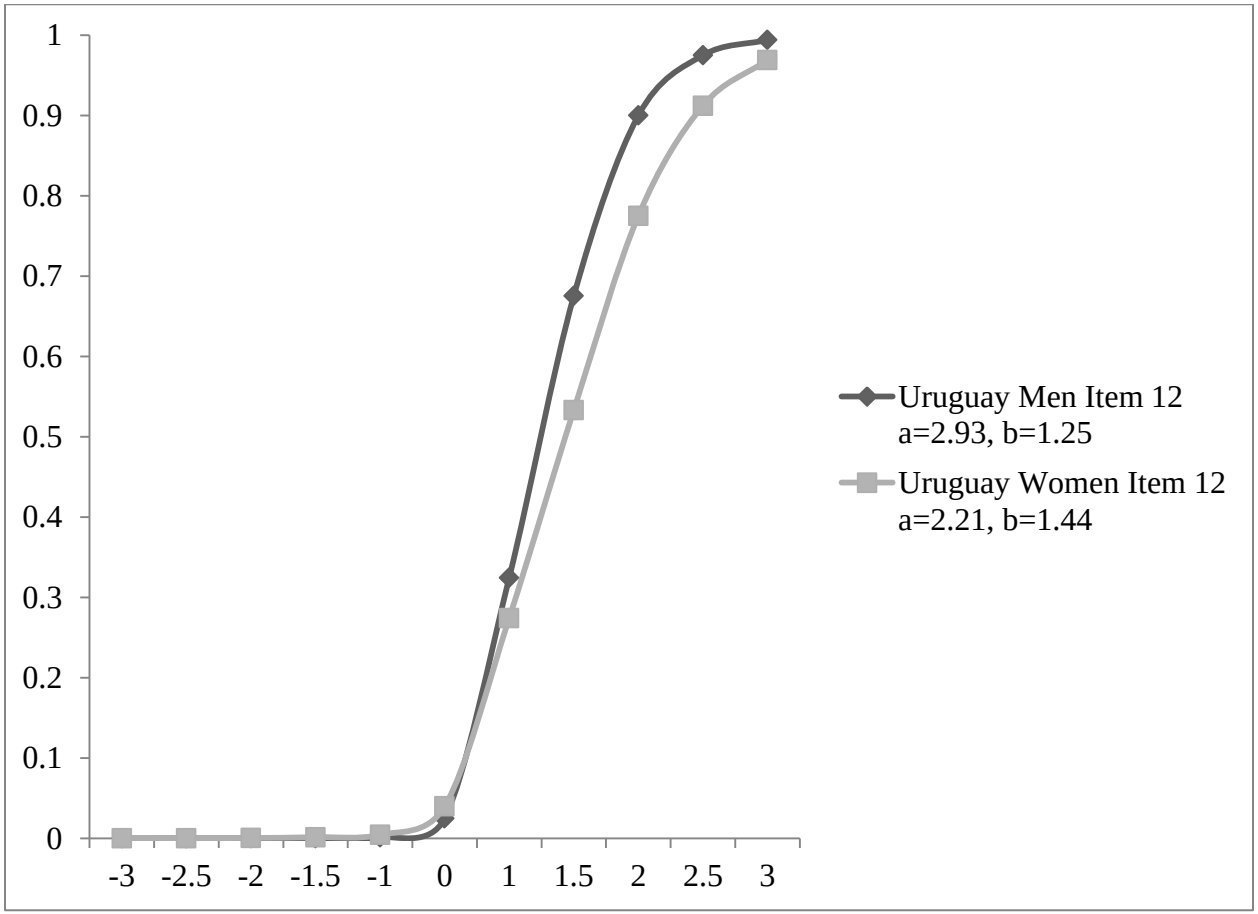


Figure 22 Uruguay by gender item 14

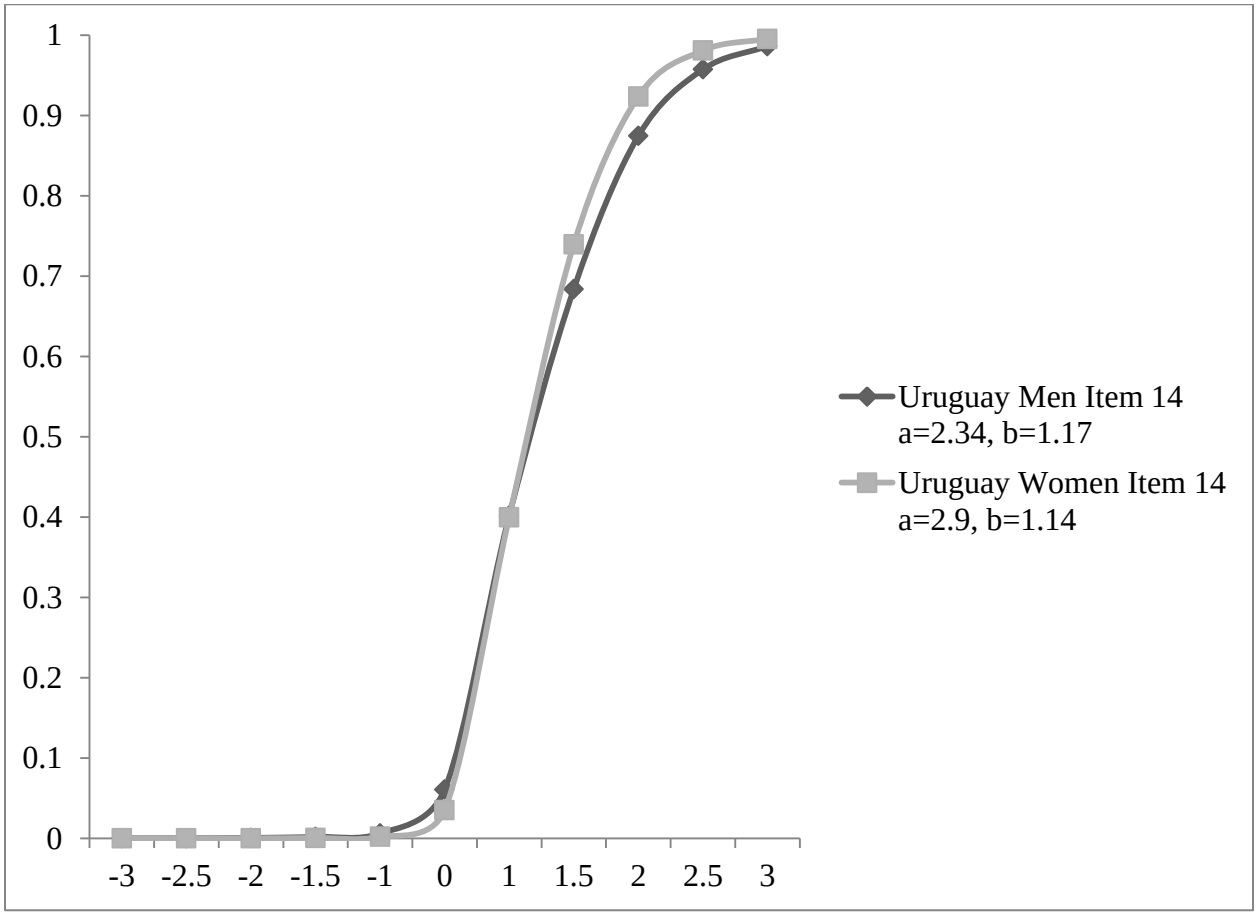


Figure 23 Test information curve Uruguay by gender

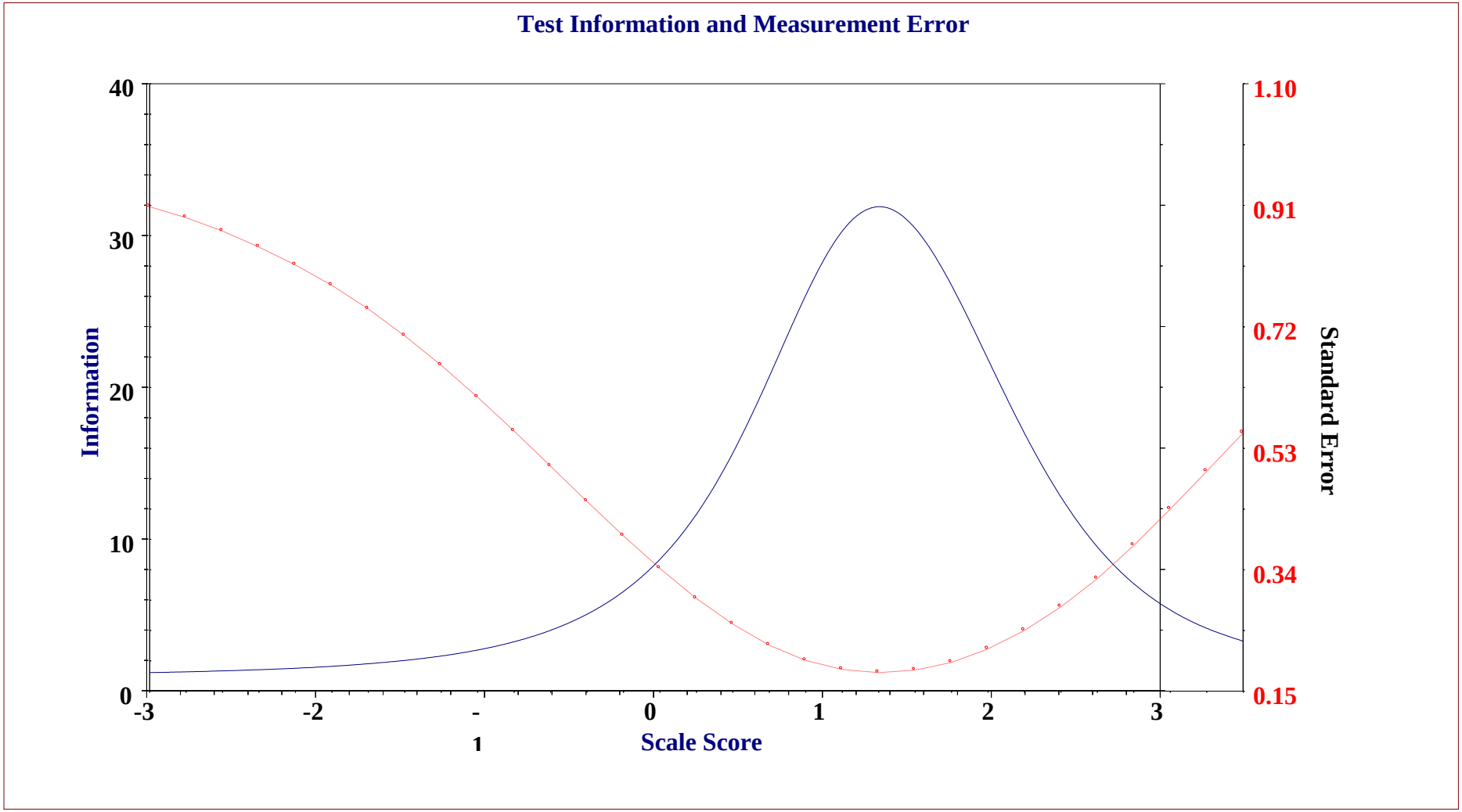


Figure 24 Chile by Cuba item 4

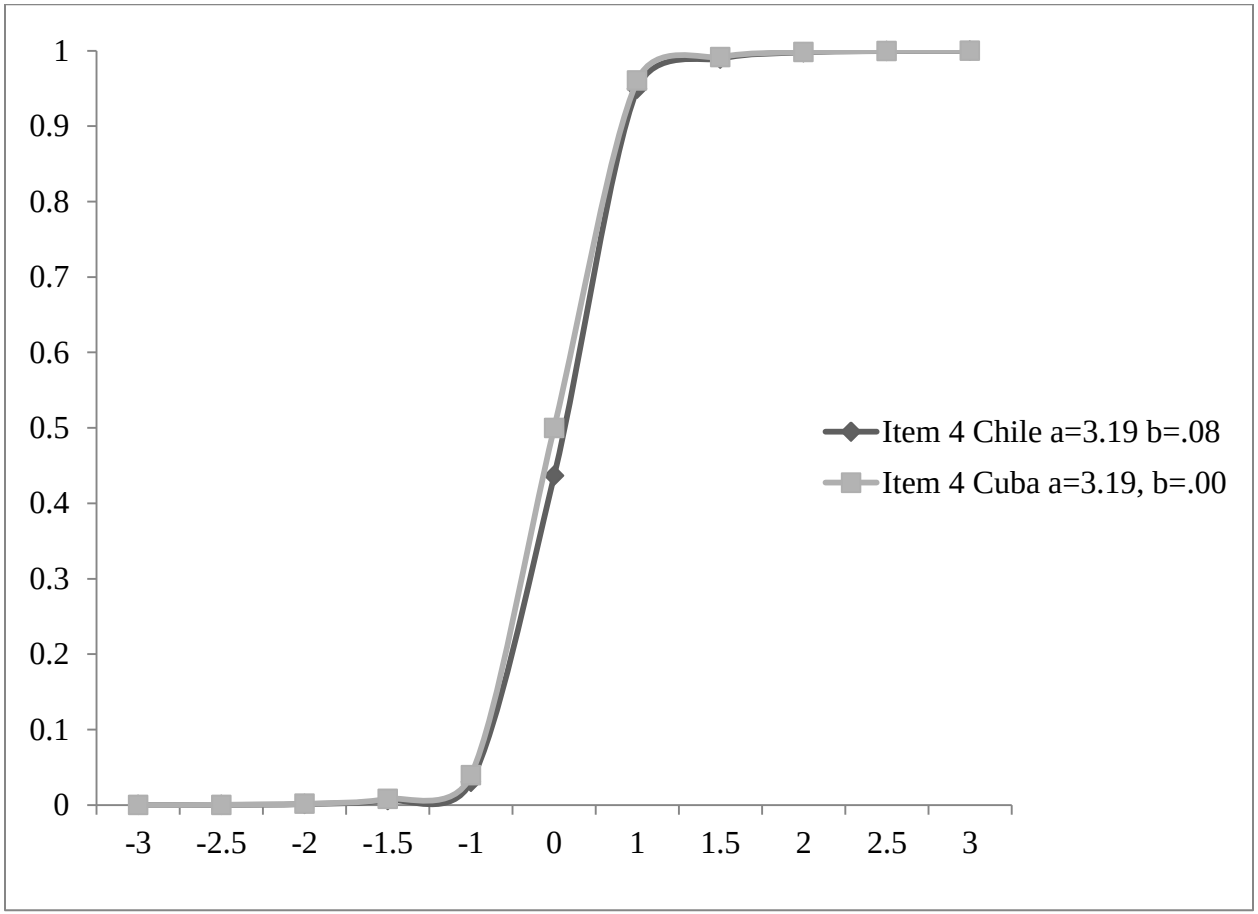


Figure 25 Chile by Cuba item 8

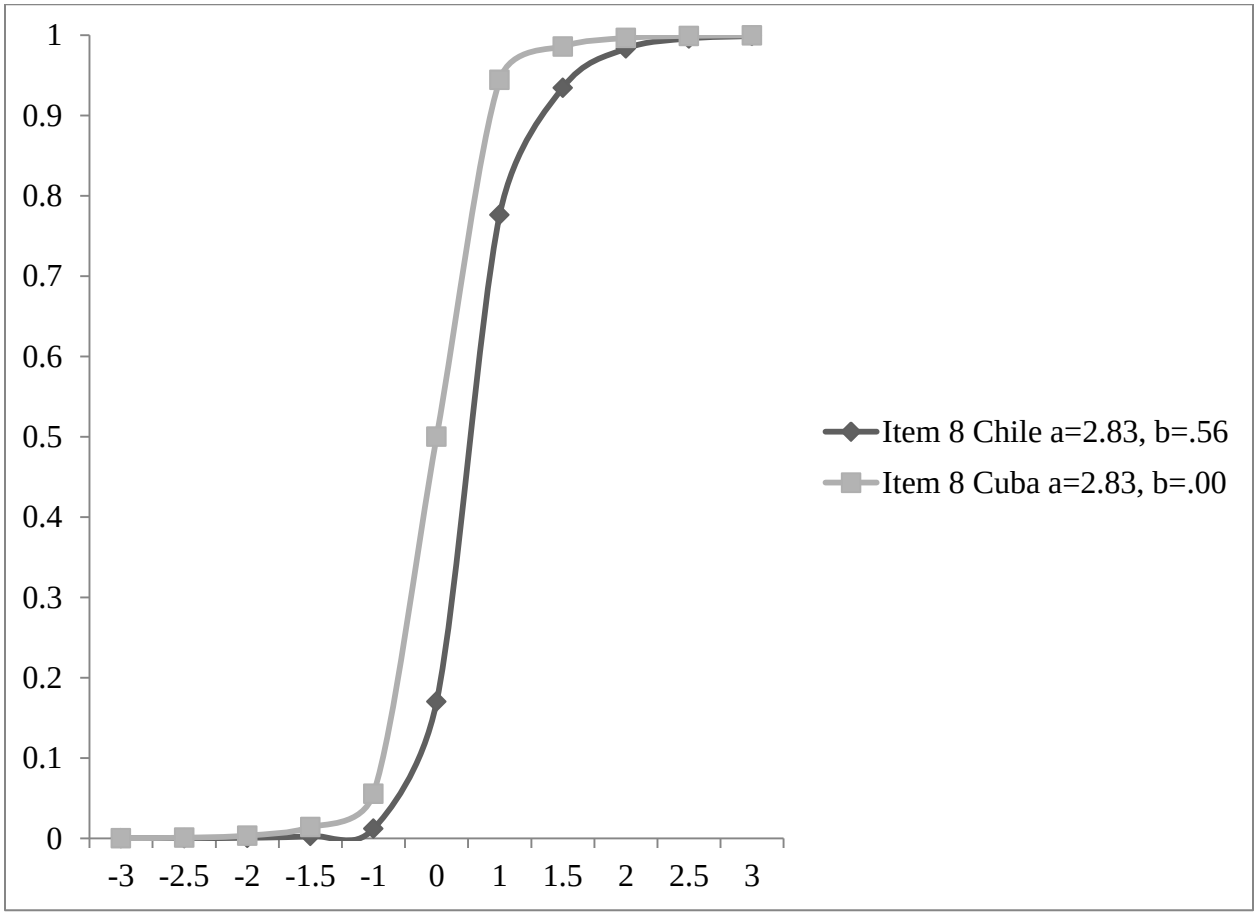


Figure 26 Chile by Cuba test information curve

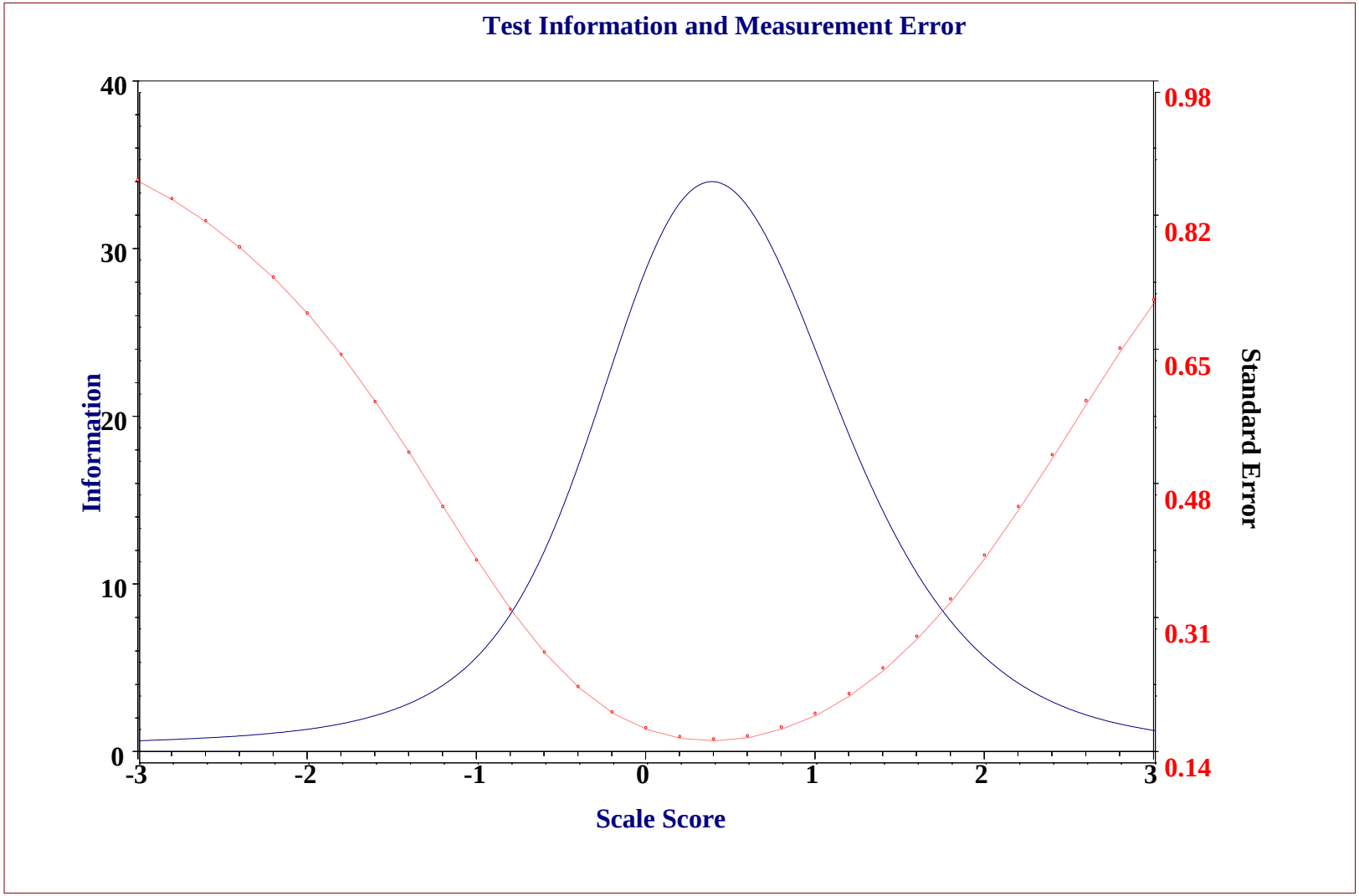


Figure 27 Mexico by Uruguay item 6

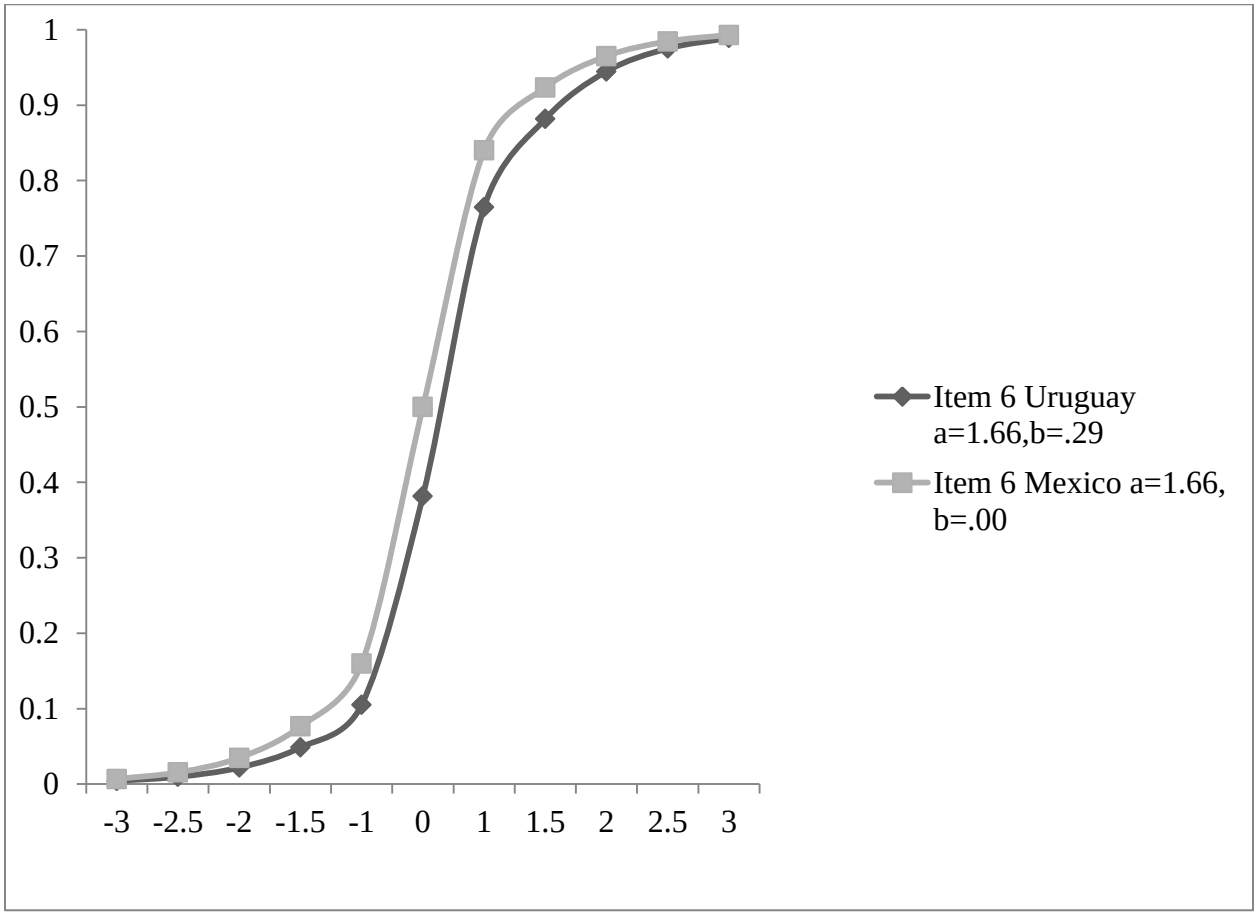


Figure 28 Mexico by Uruguay item 7

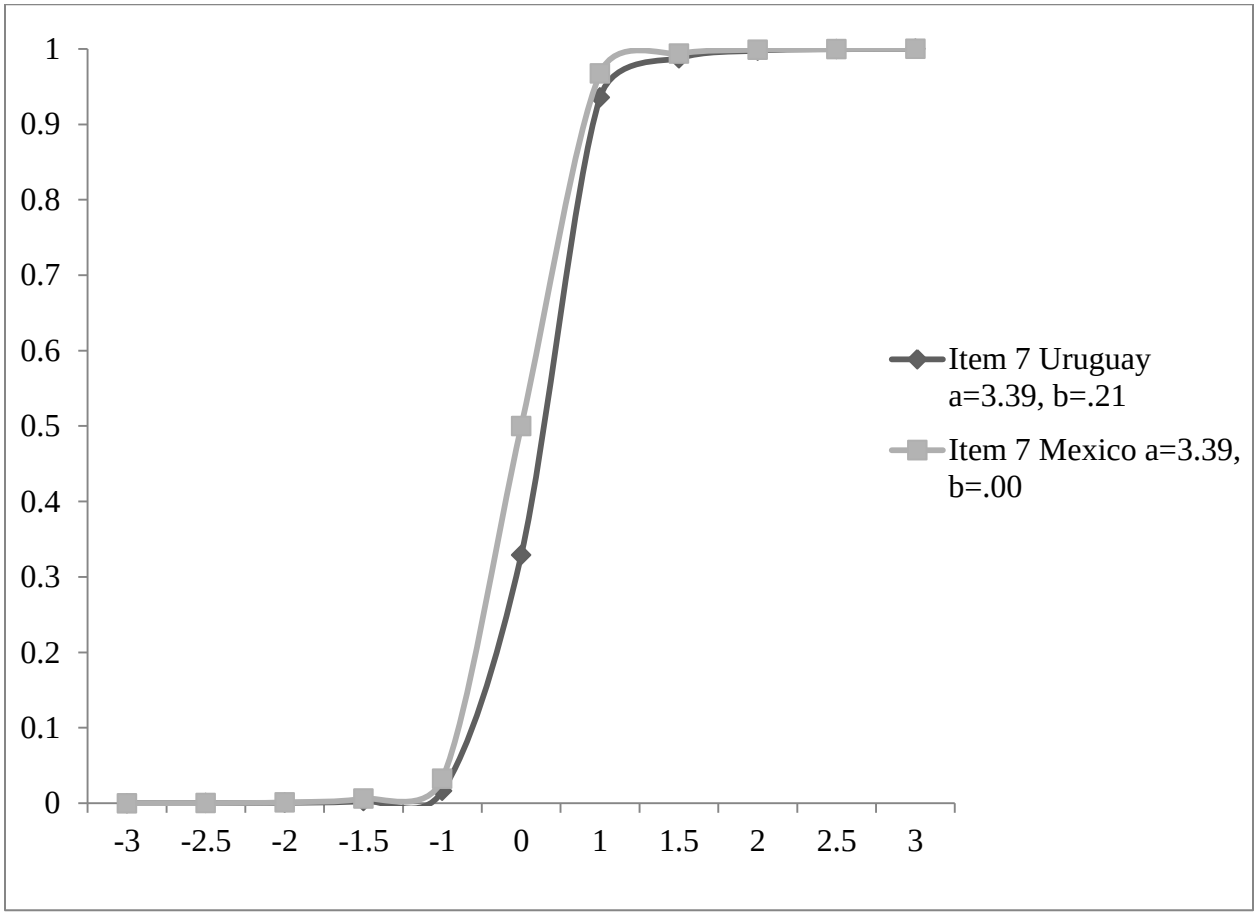


Figure 29 Mexico by Uruguay item 9

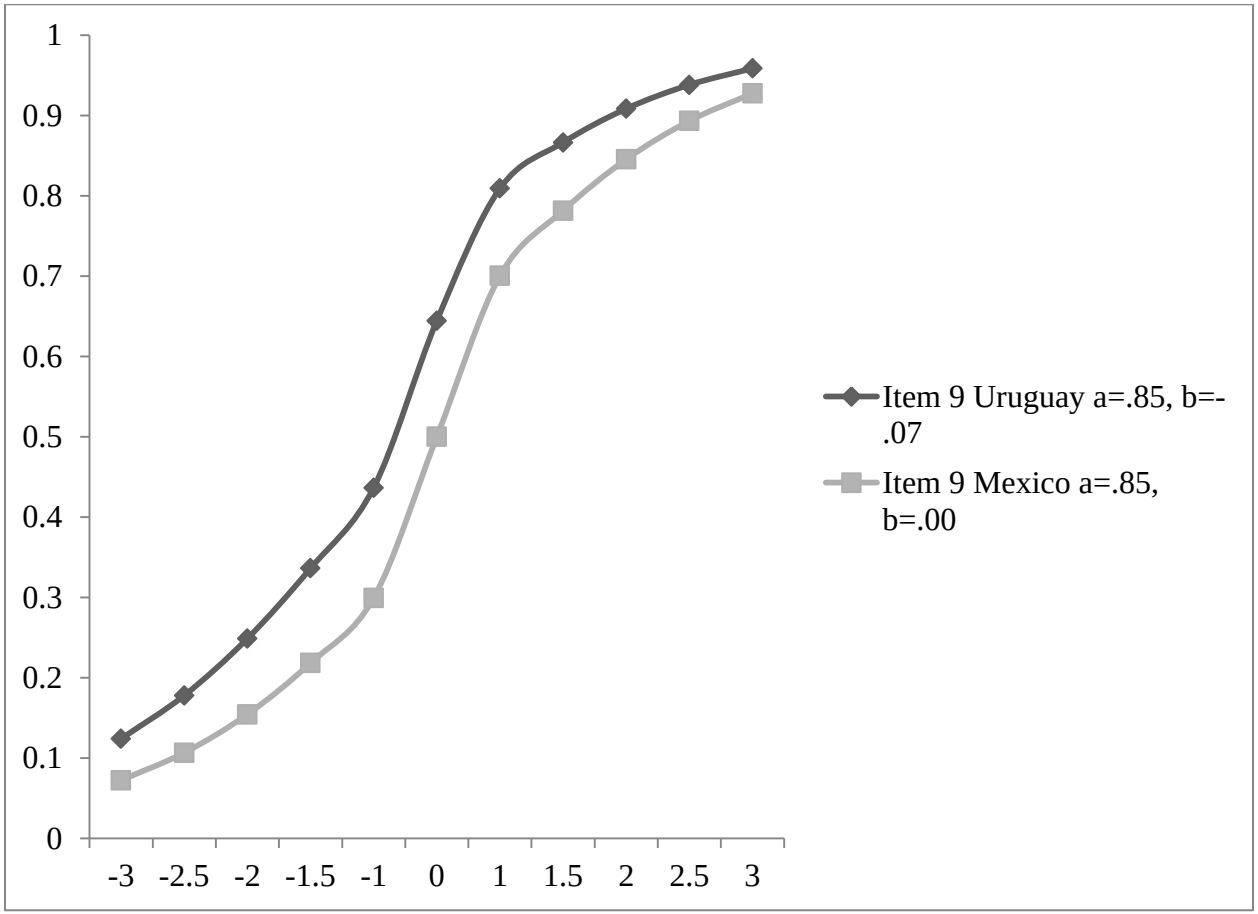


Figure 30 Mexico by Uruguay item 12

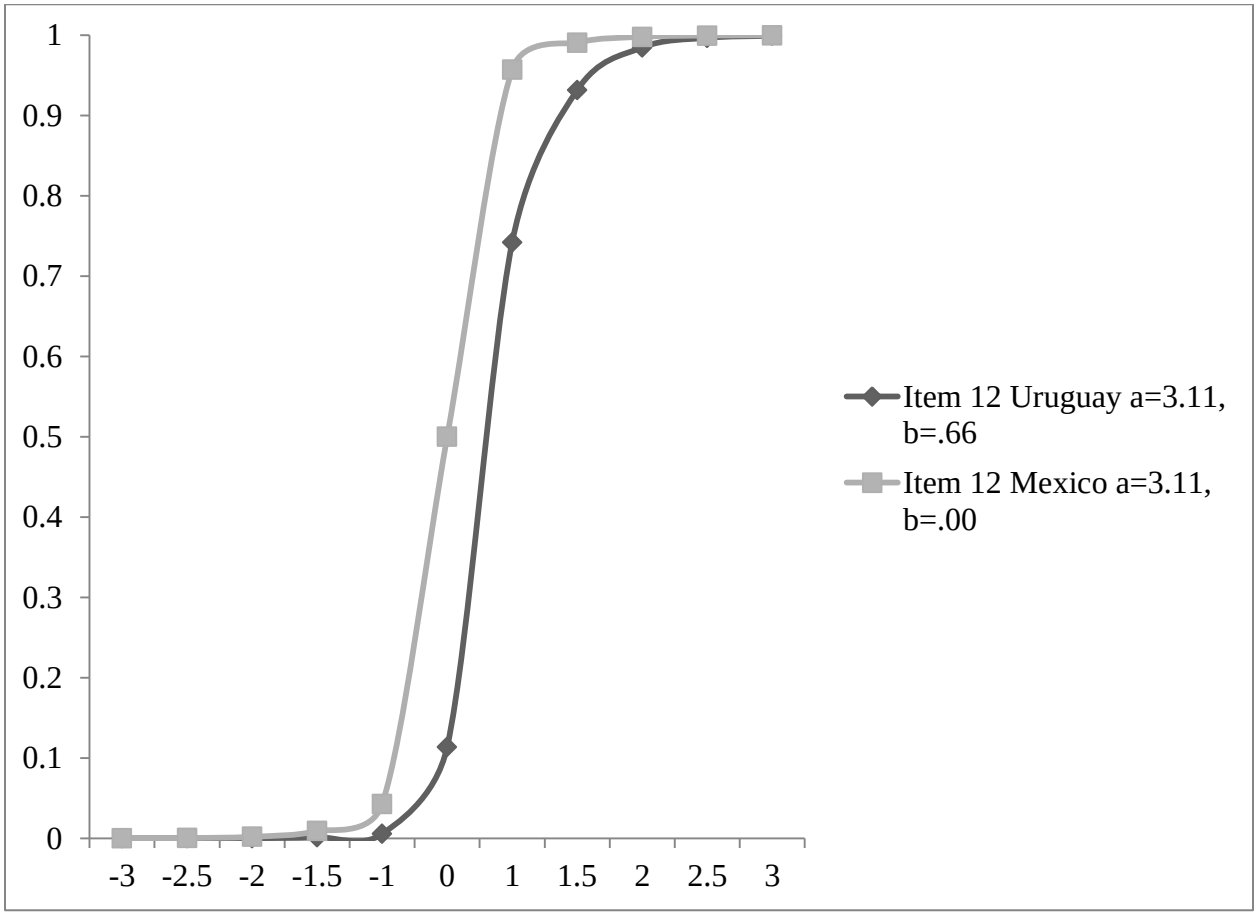


Figure 31 Mexico by Uruguay test information curve

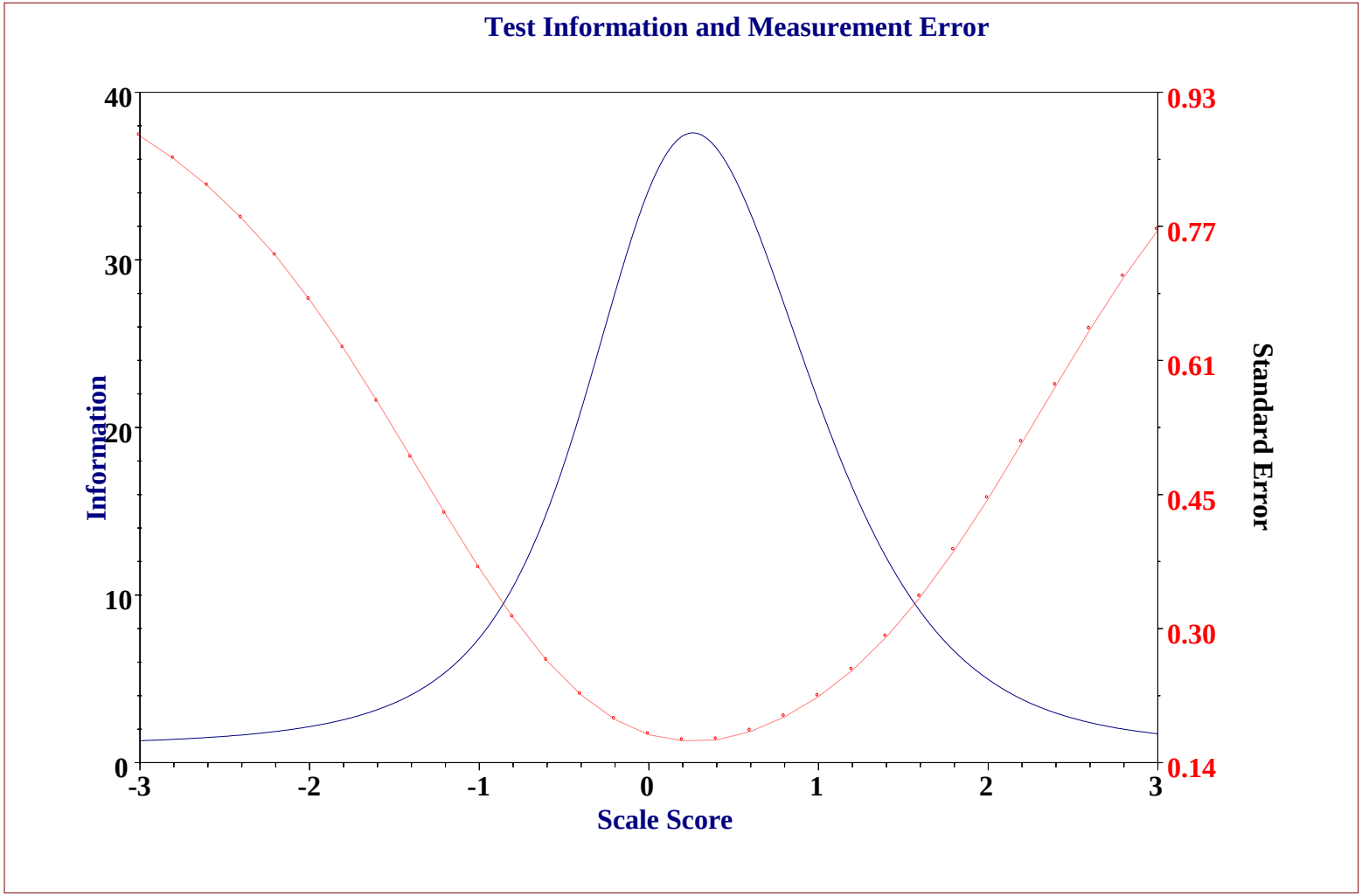


Figure 32 Mexico by Argentina item 2

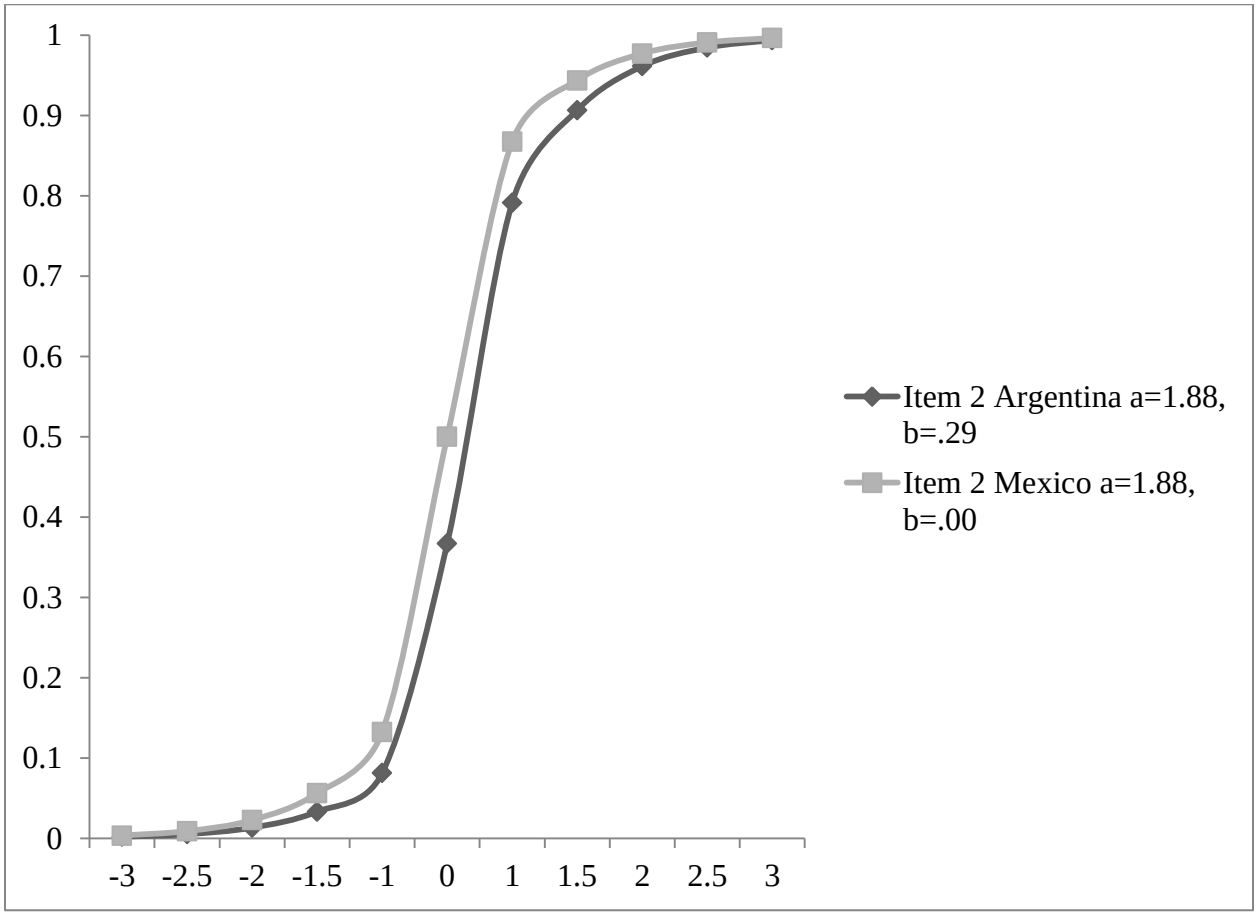


Figure 33 Mexico by Argentina item 7

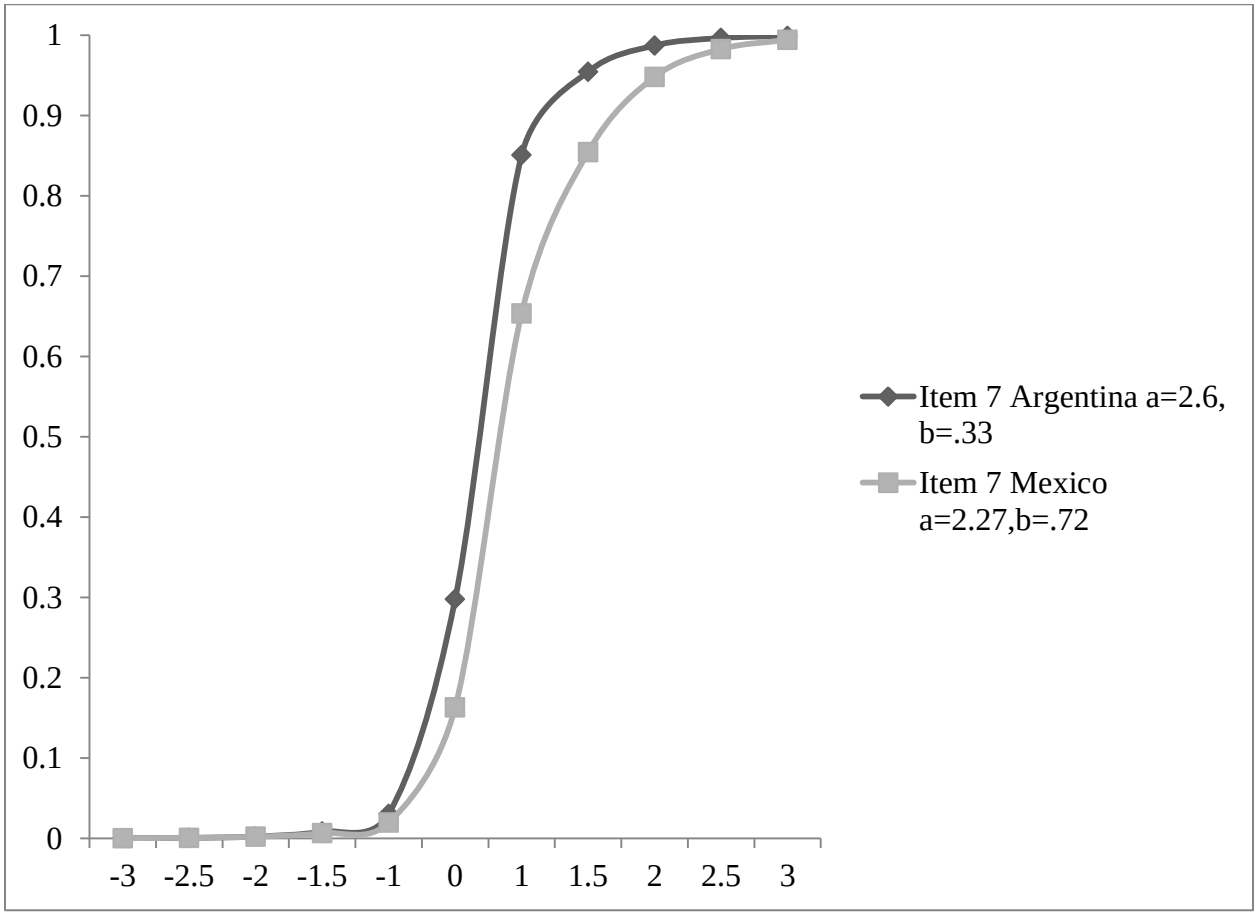


Figure 34 Mexico by Argentina item 9

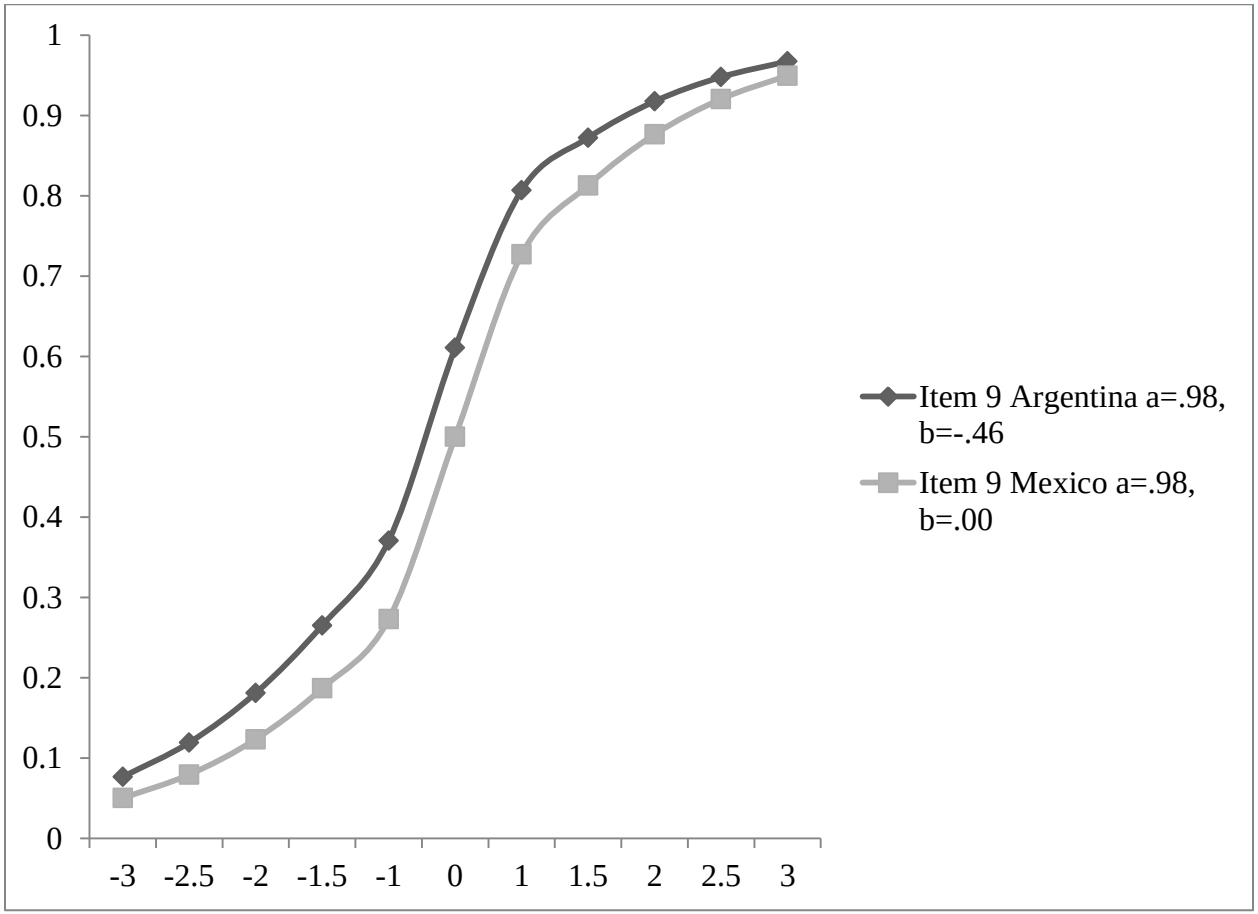


Figure 35 Mexico by Argentina item 11

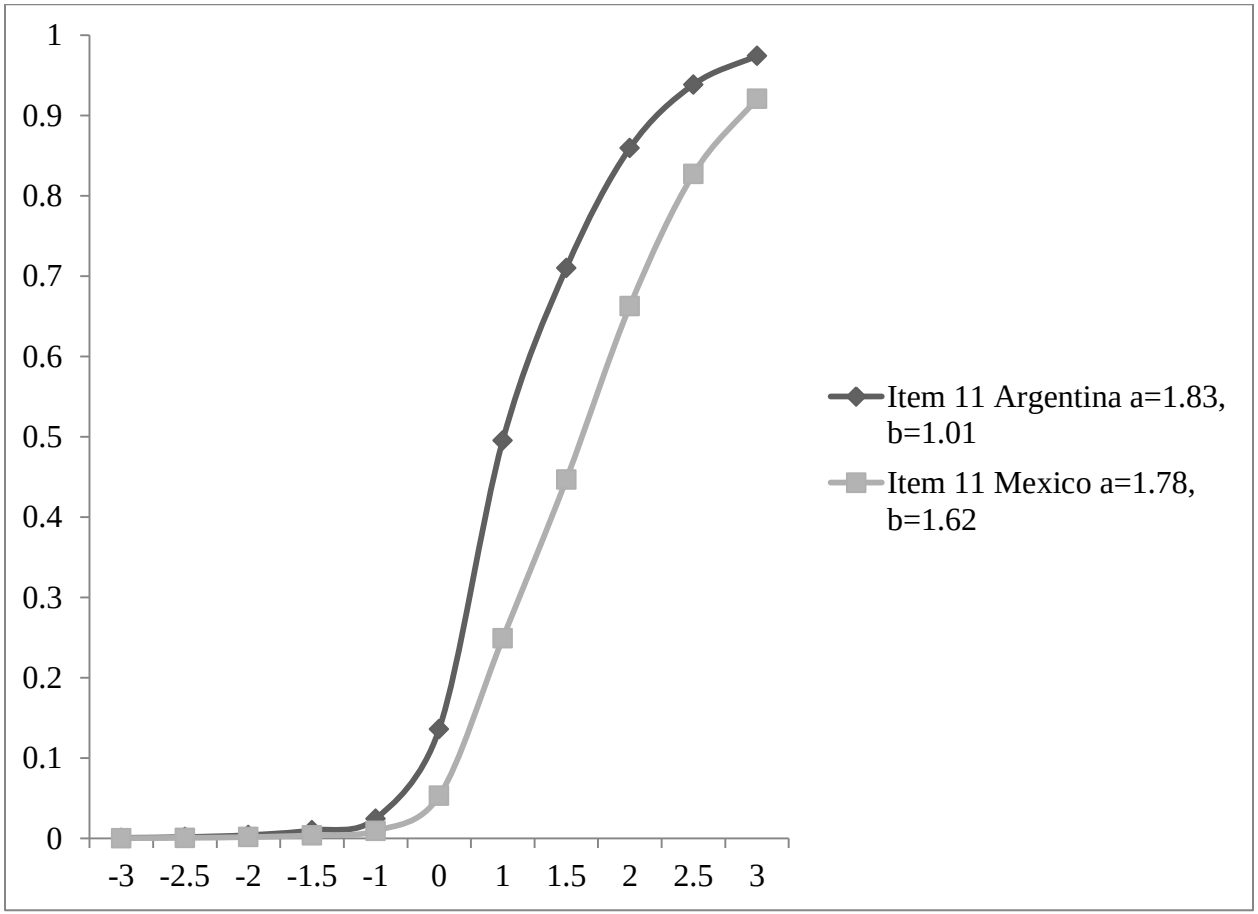


Figure 36 Mexico by Argentina item 12

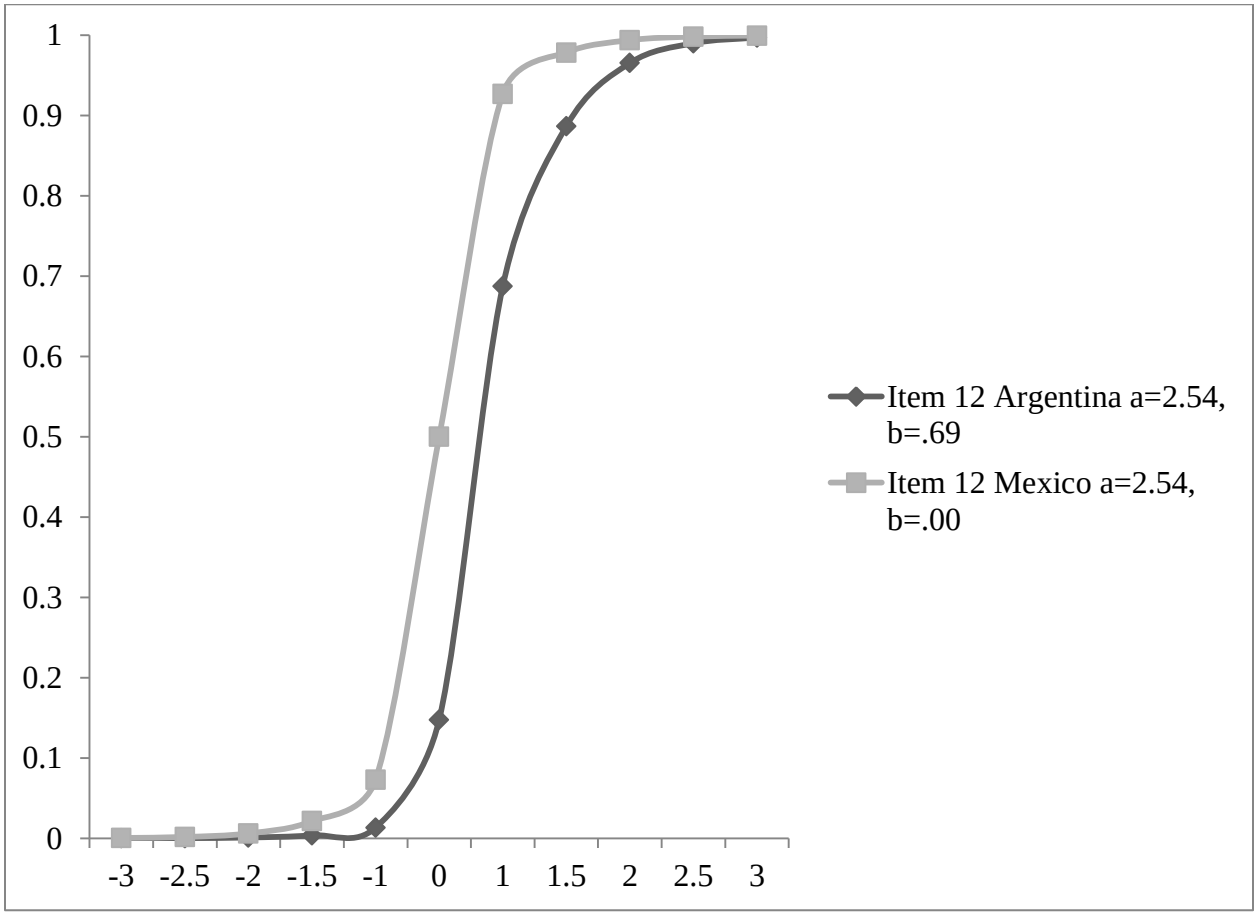


Figure 37 Mexico by Argentina test information curve

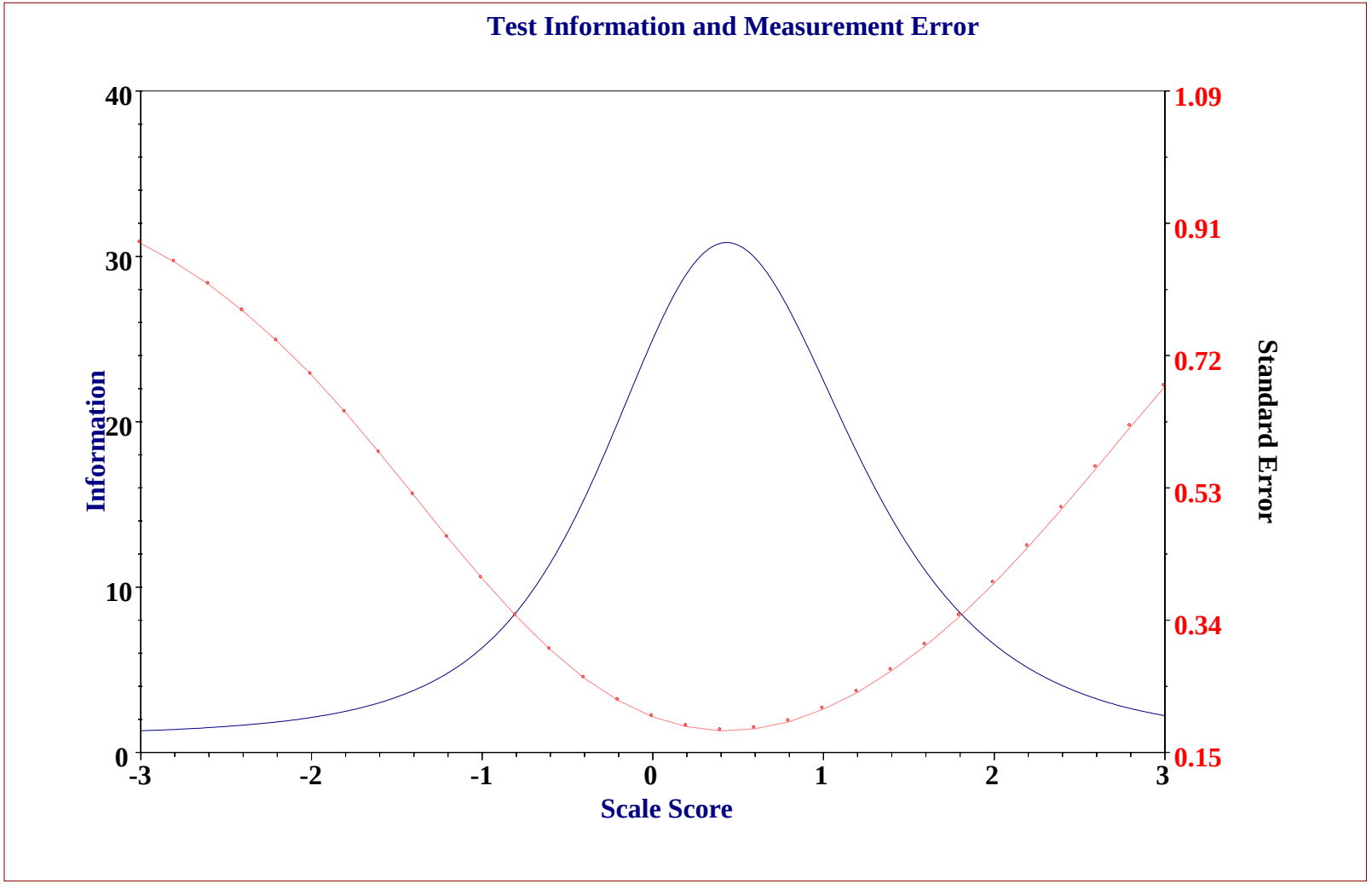


Figure 38 Argentina by Uruguay item 2

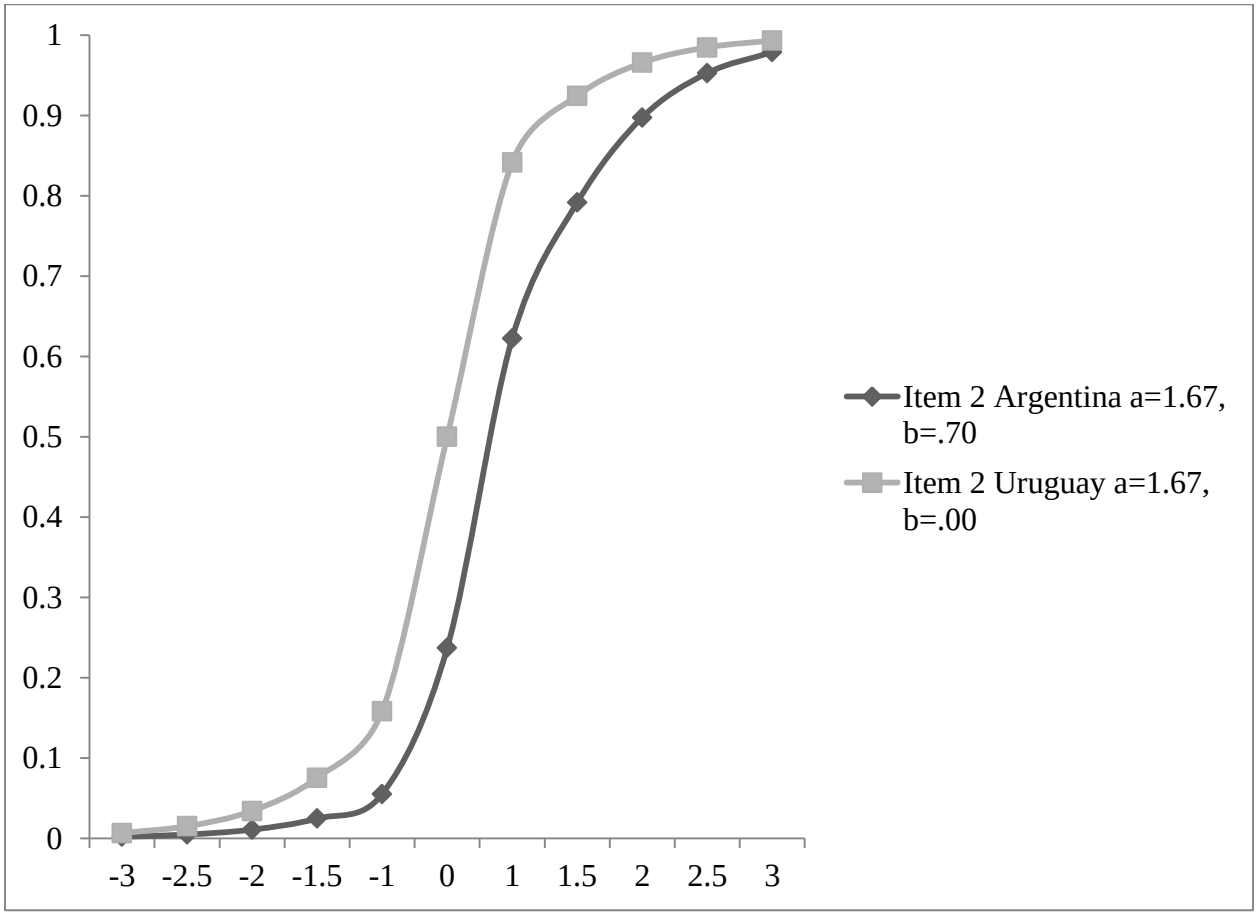


Figure 39 Argentina by Uruguay item 6

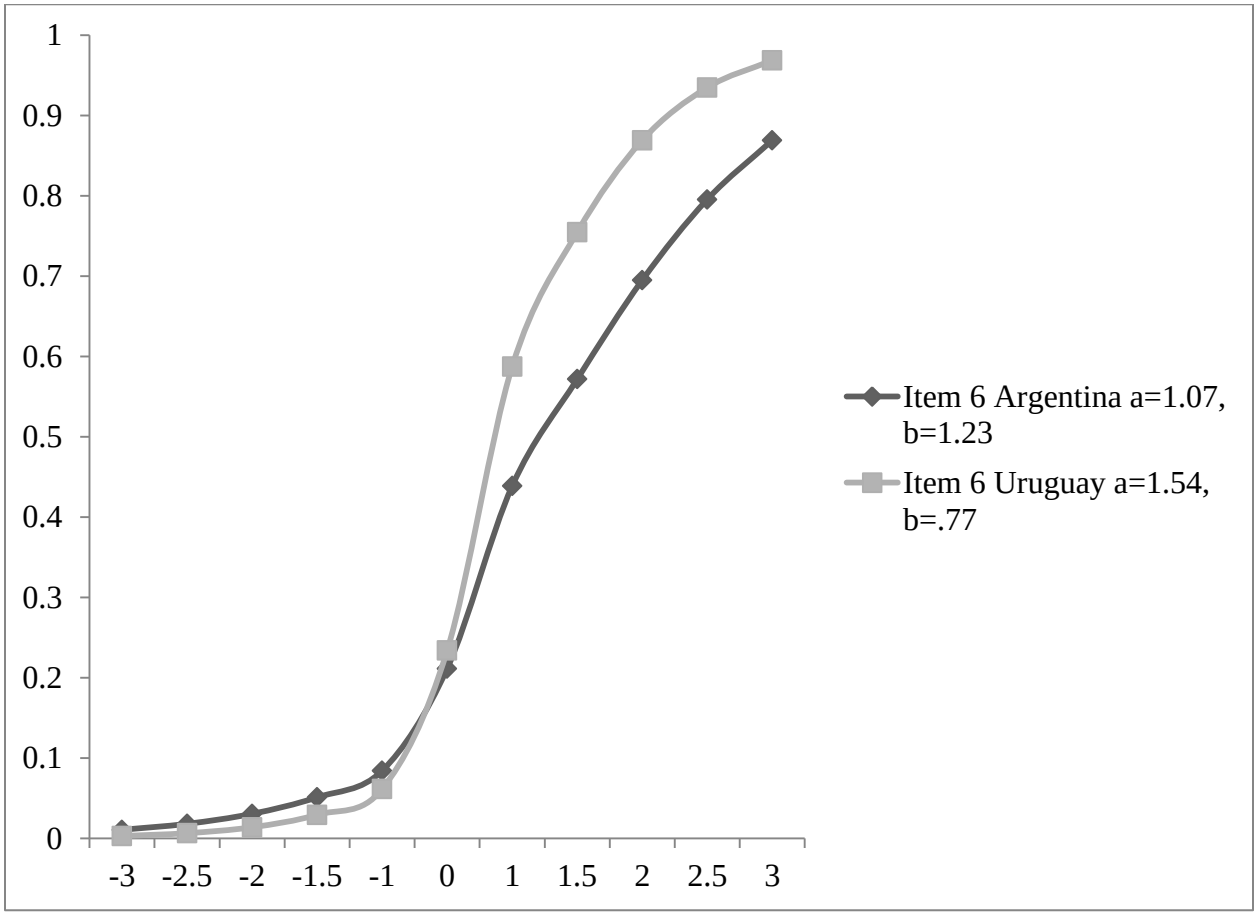


Figure 40 Argentina by Uruguay item 7

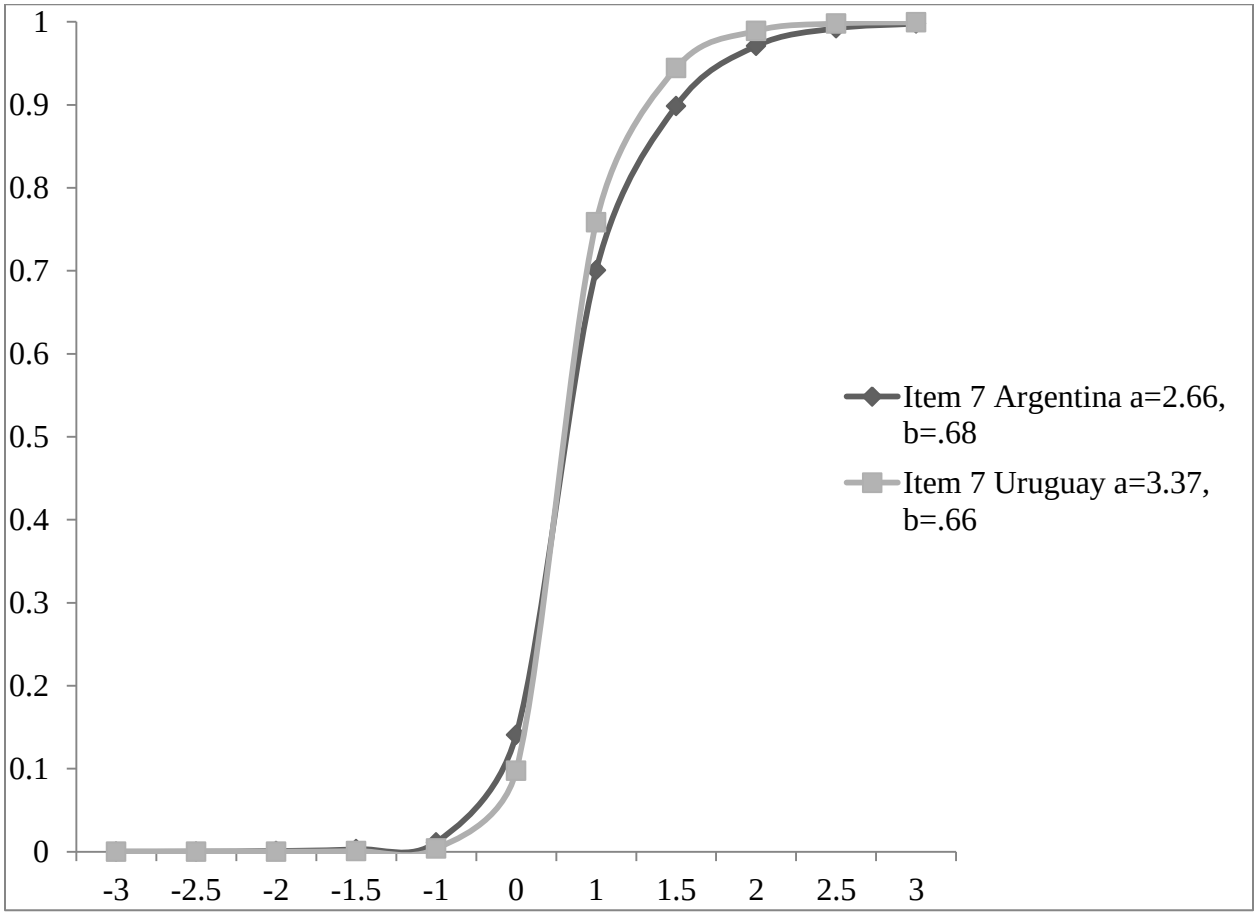


Figure 41 Argentina by Uruguay item 11

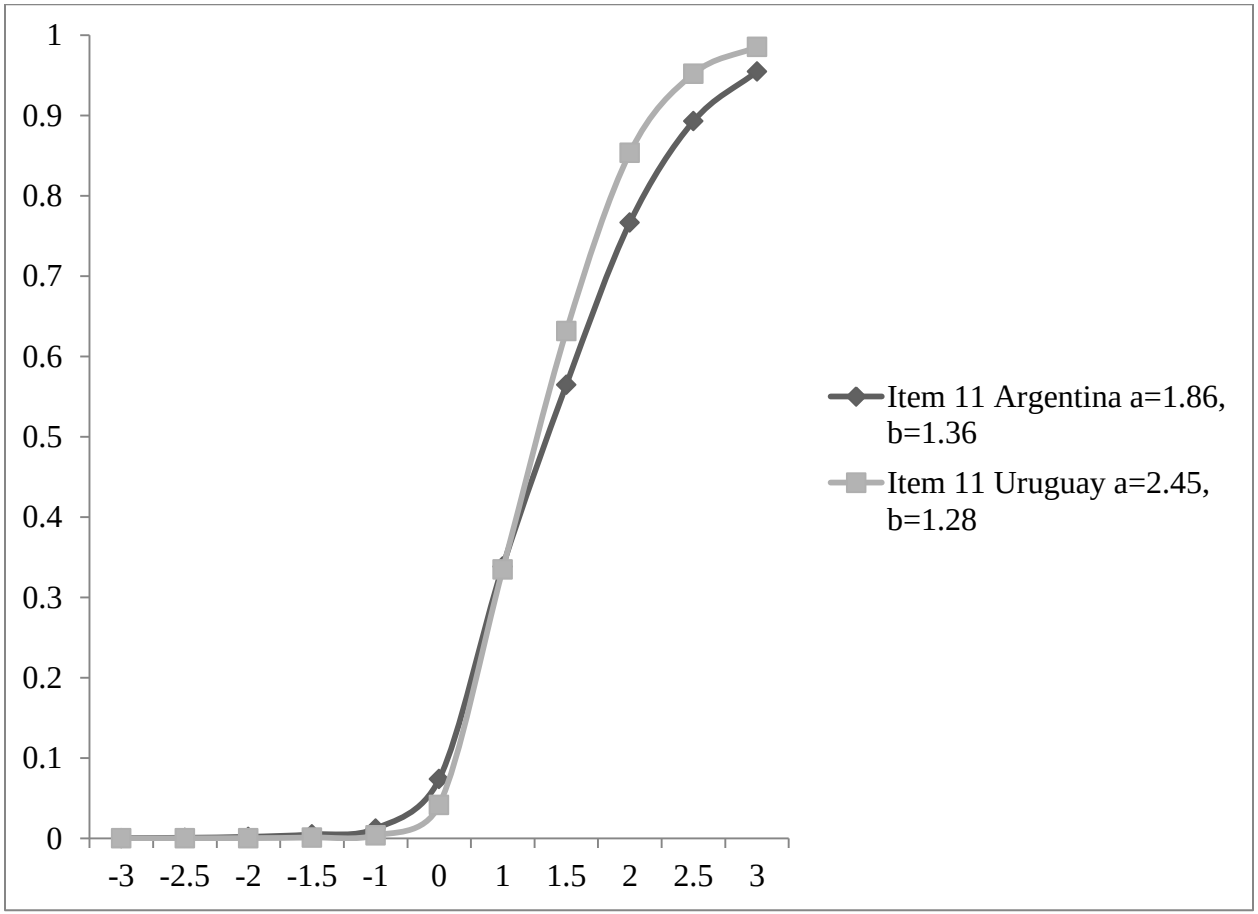


Figure 42 Argentina by Uruguay test information curve

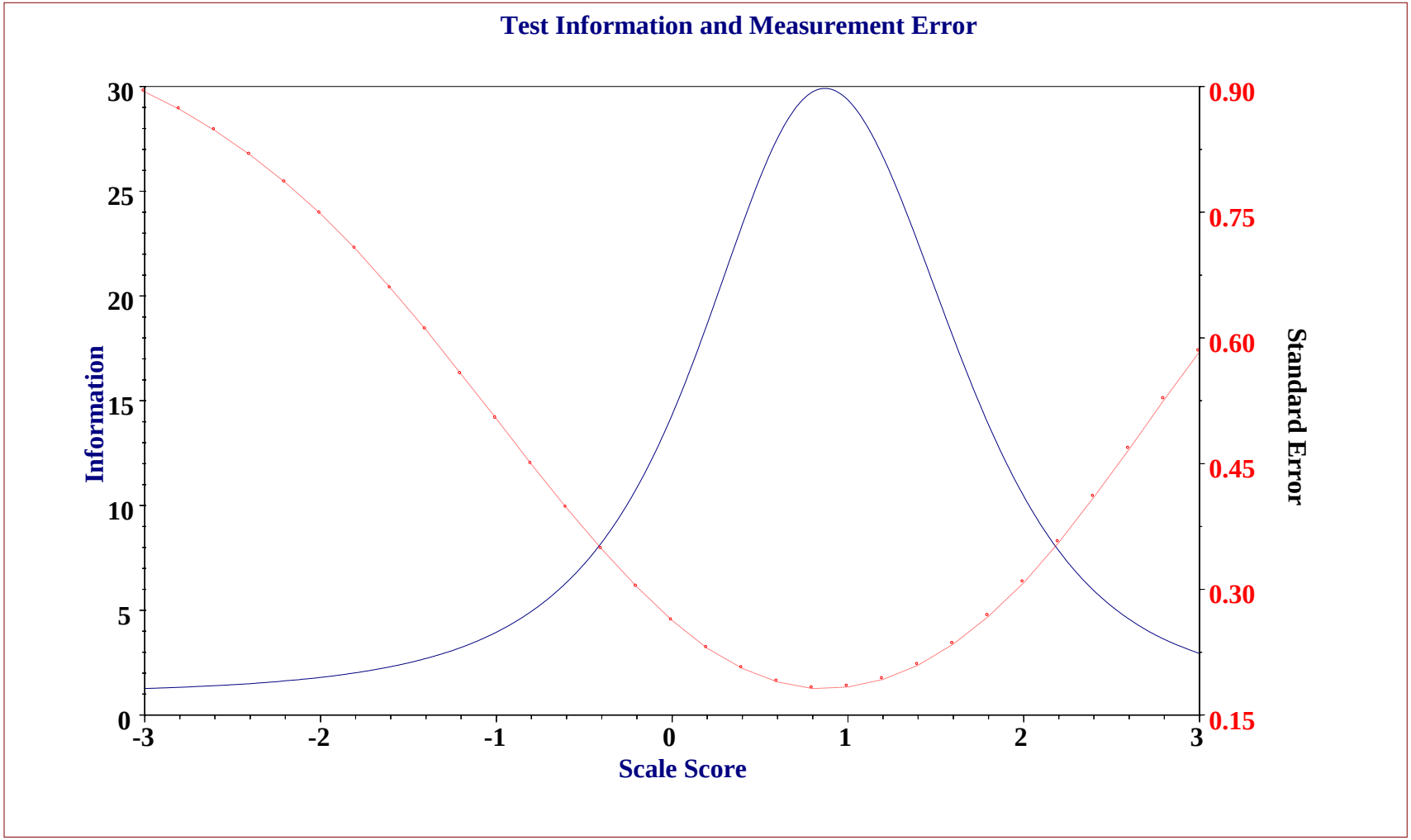


Figure 43 Argentina by Chile item 2

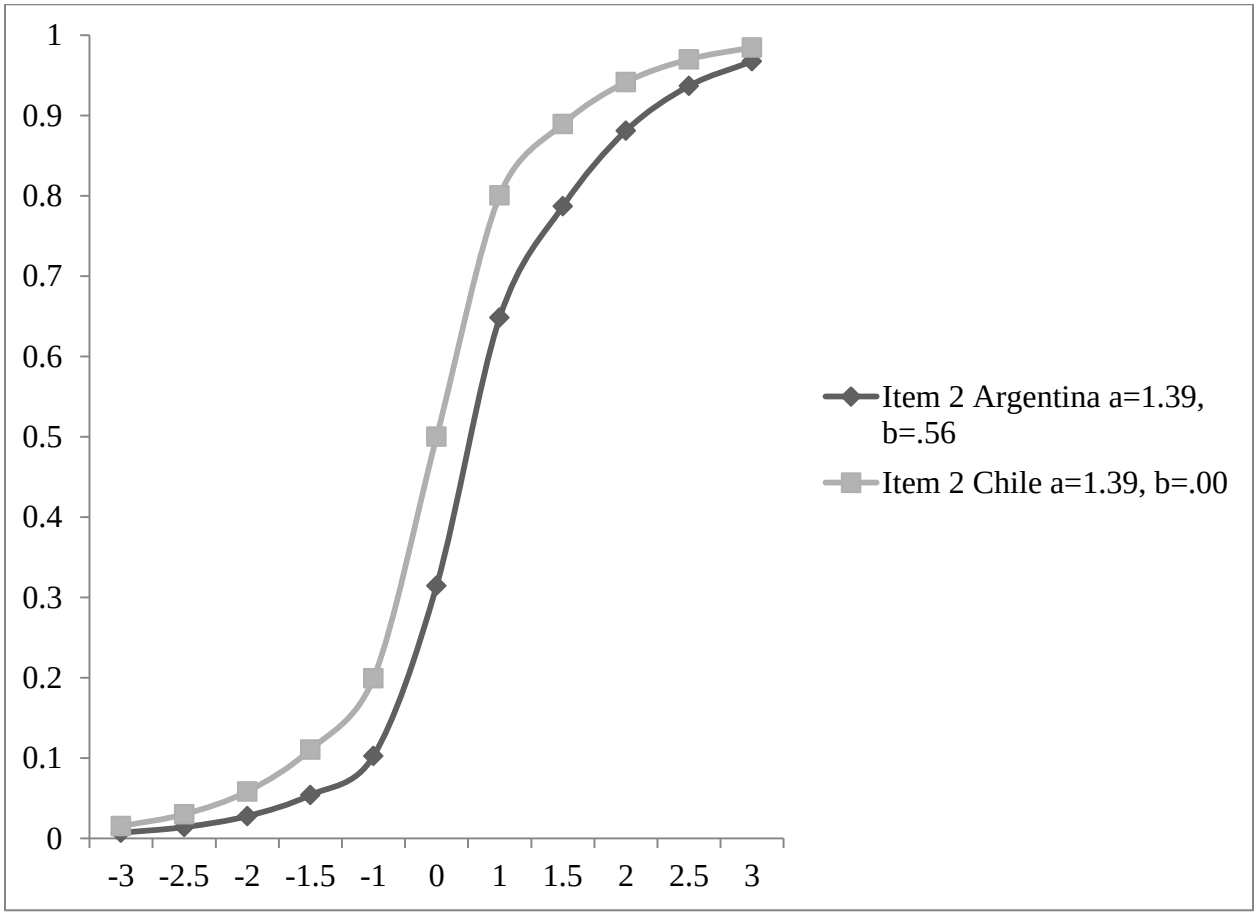


Figure 44 Argentina by Chile item 7

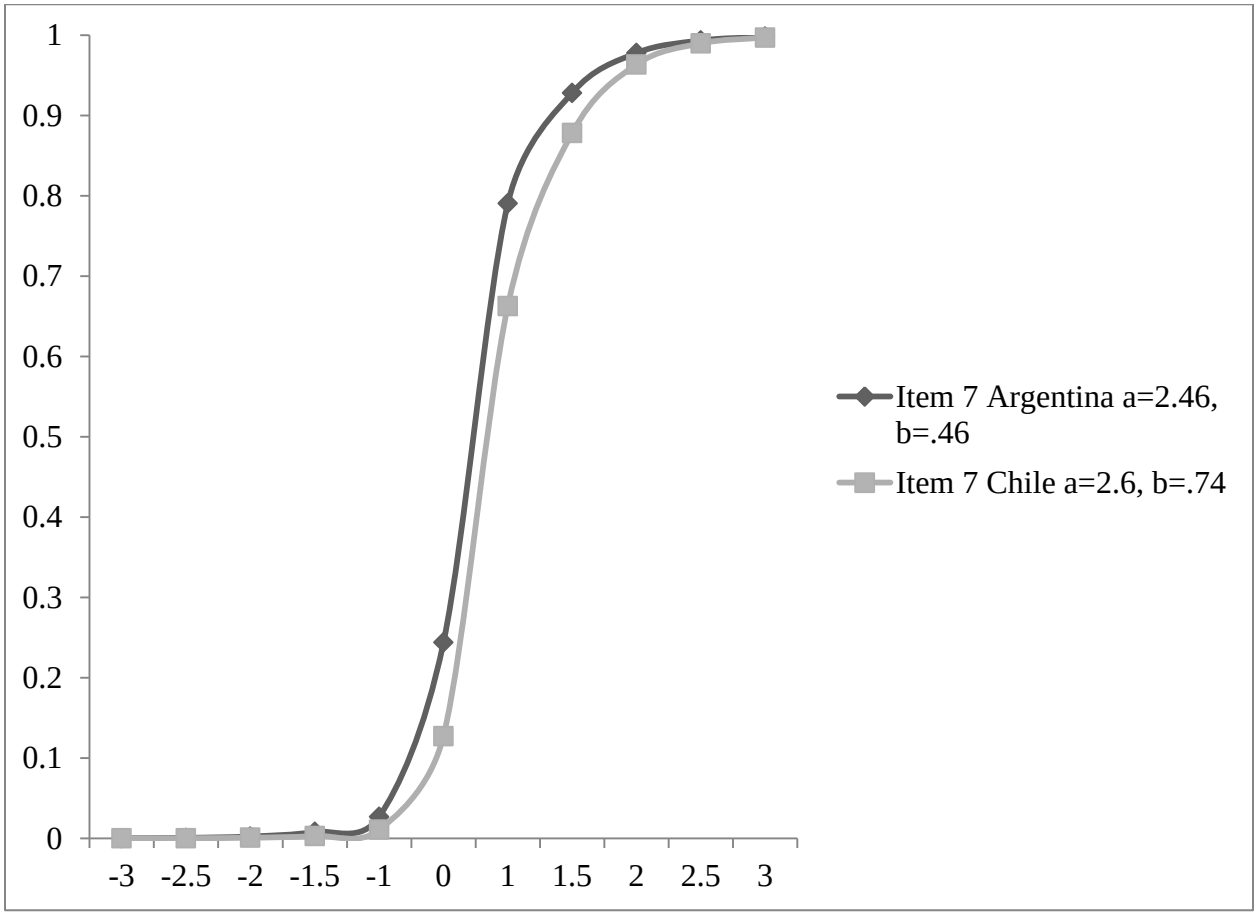


Figure 45 Argentina by Chile test information curve

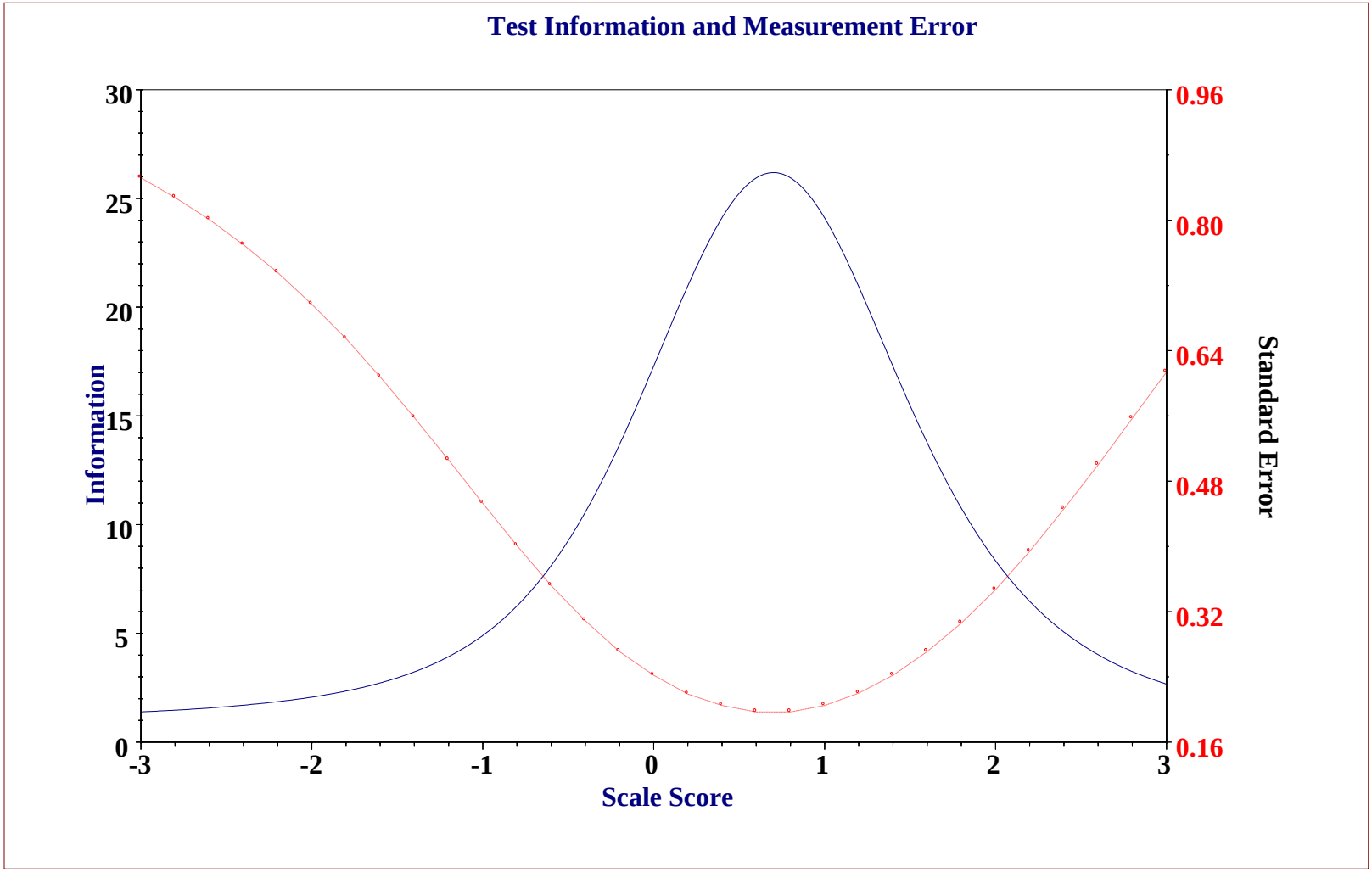


Figure 46 Argentina by Cuba item 2

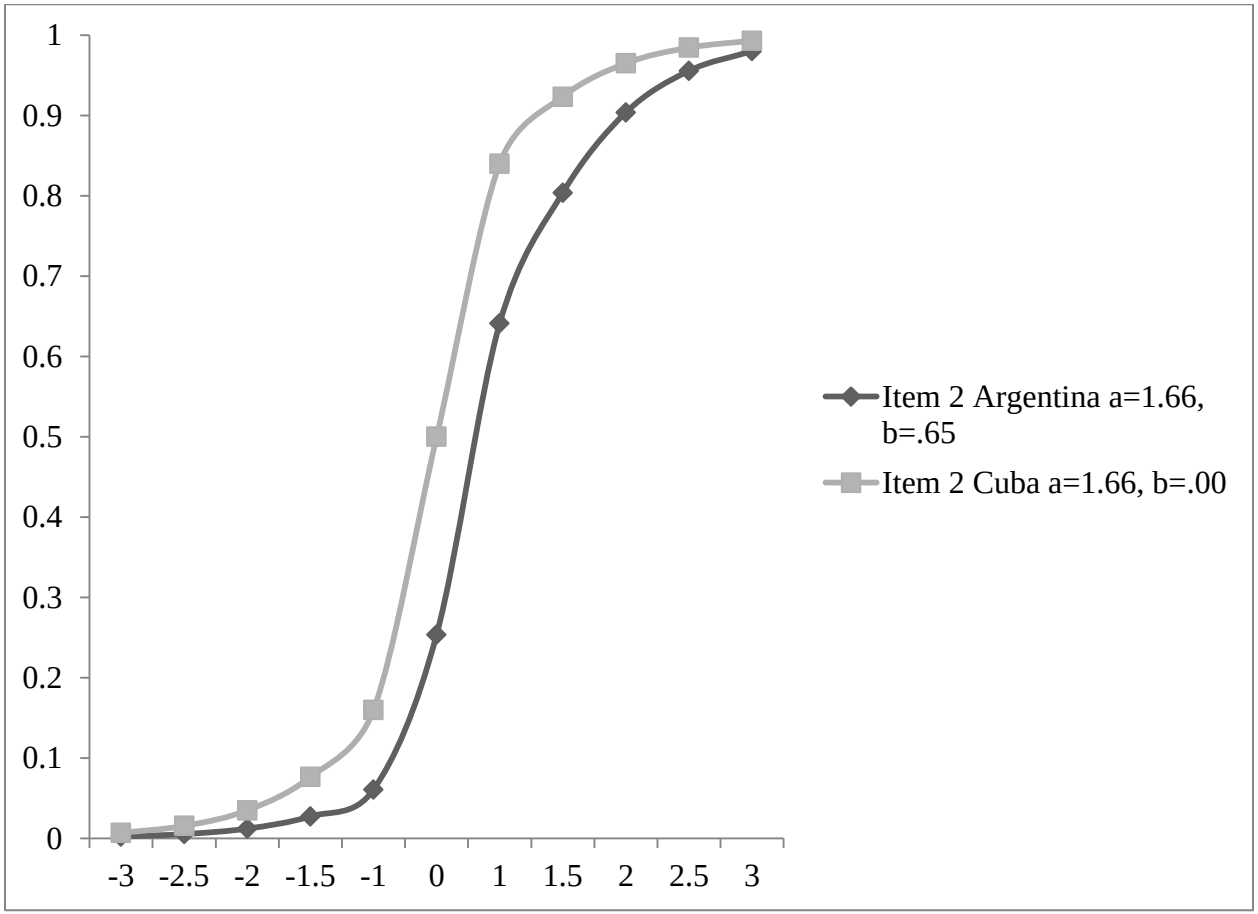


Figure 47 Argentina by Cuba item 5

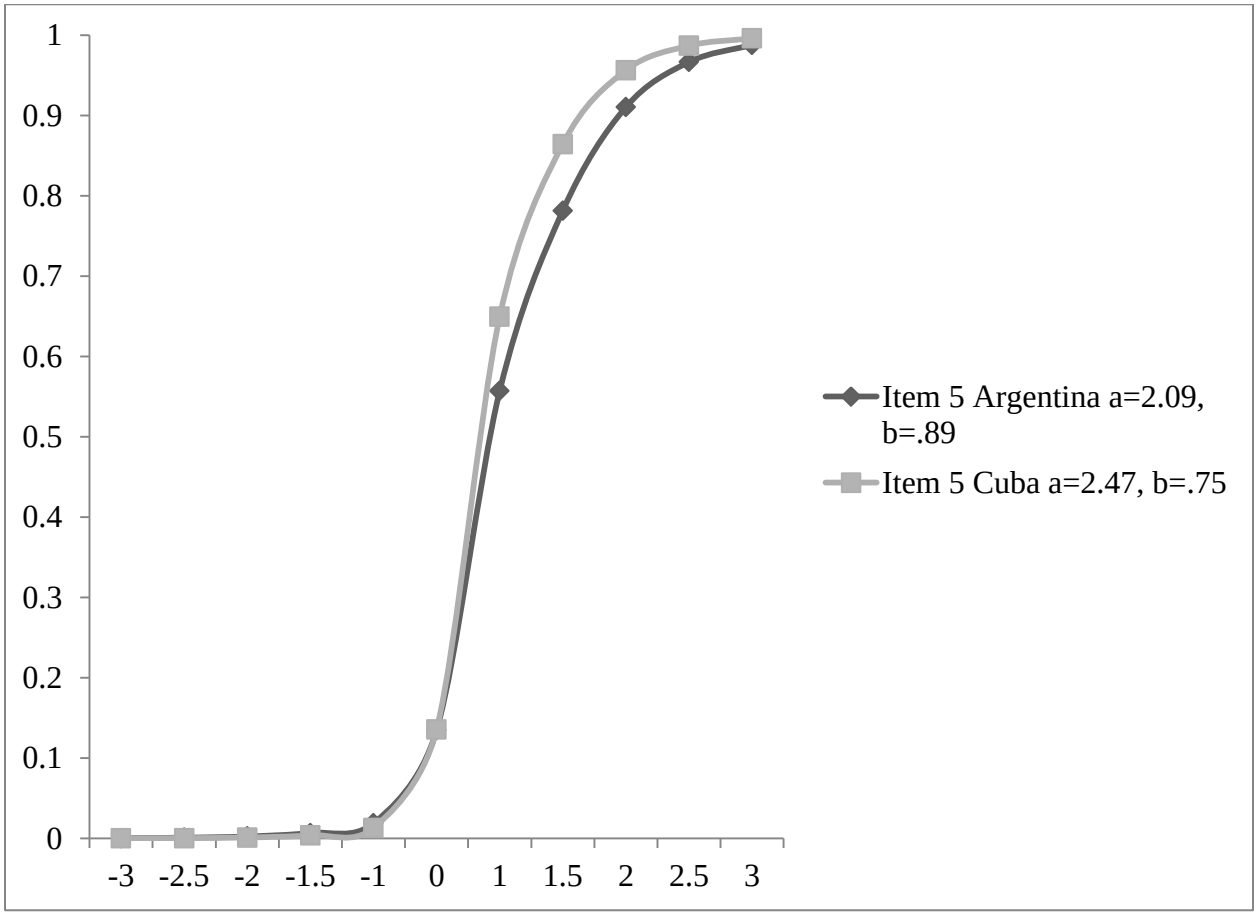


Figure 48 Argentina by Cuba item 7

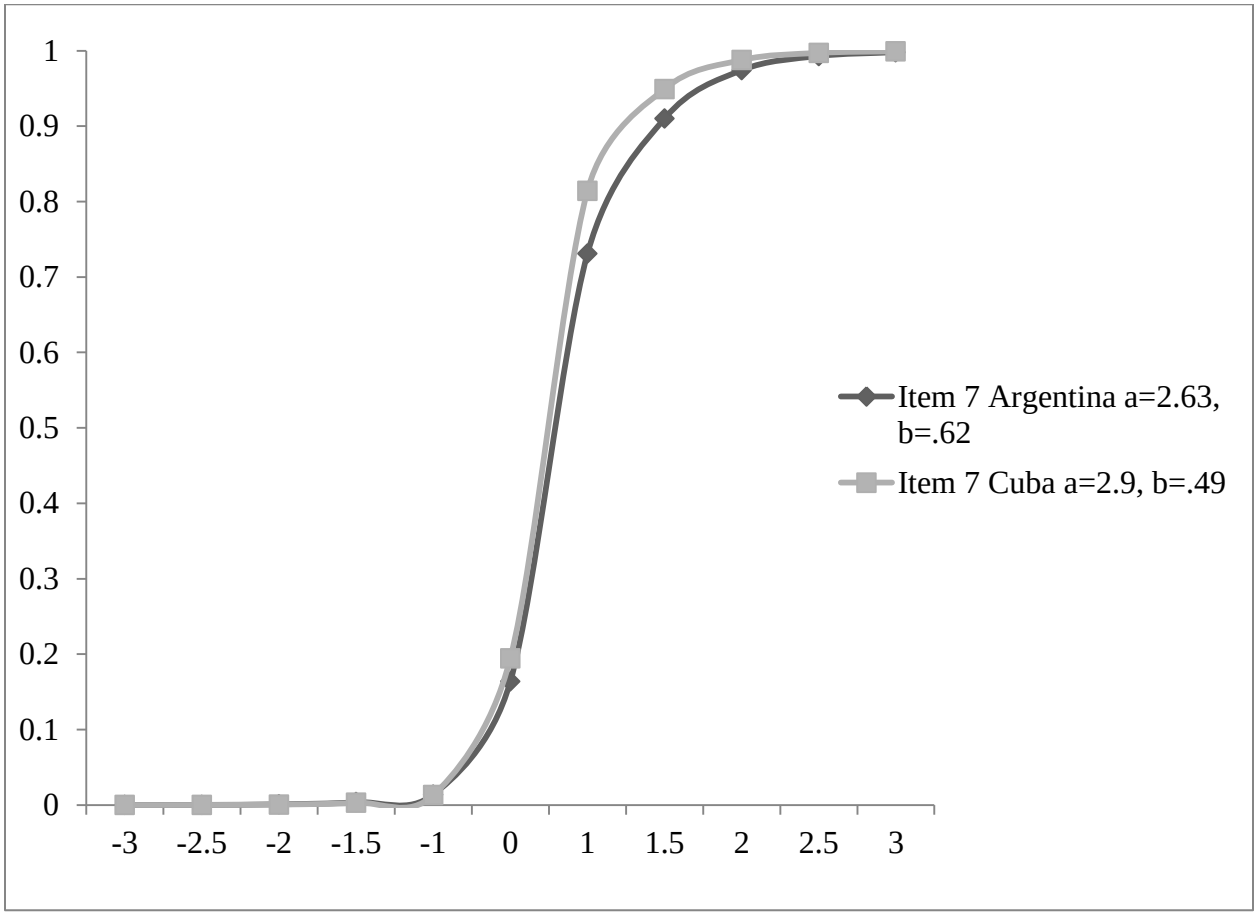


Figure 49 Argentina by Cuba test information curve

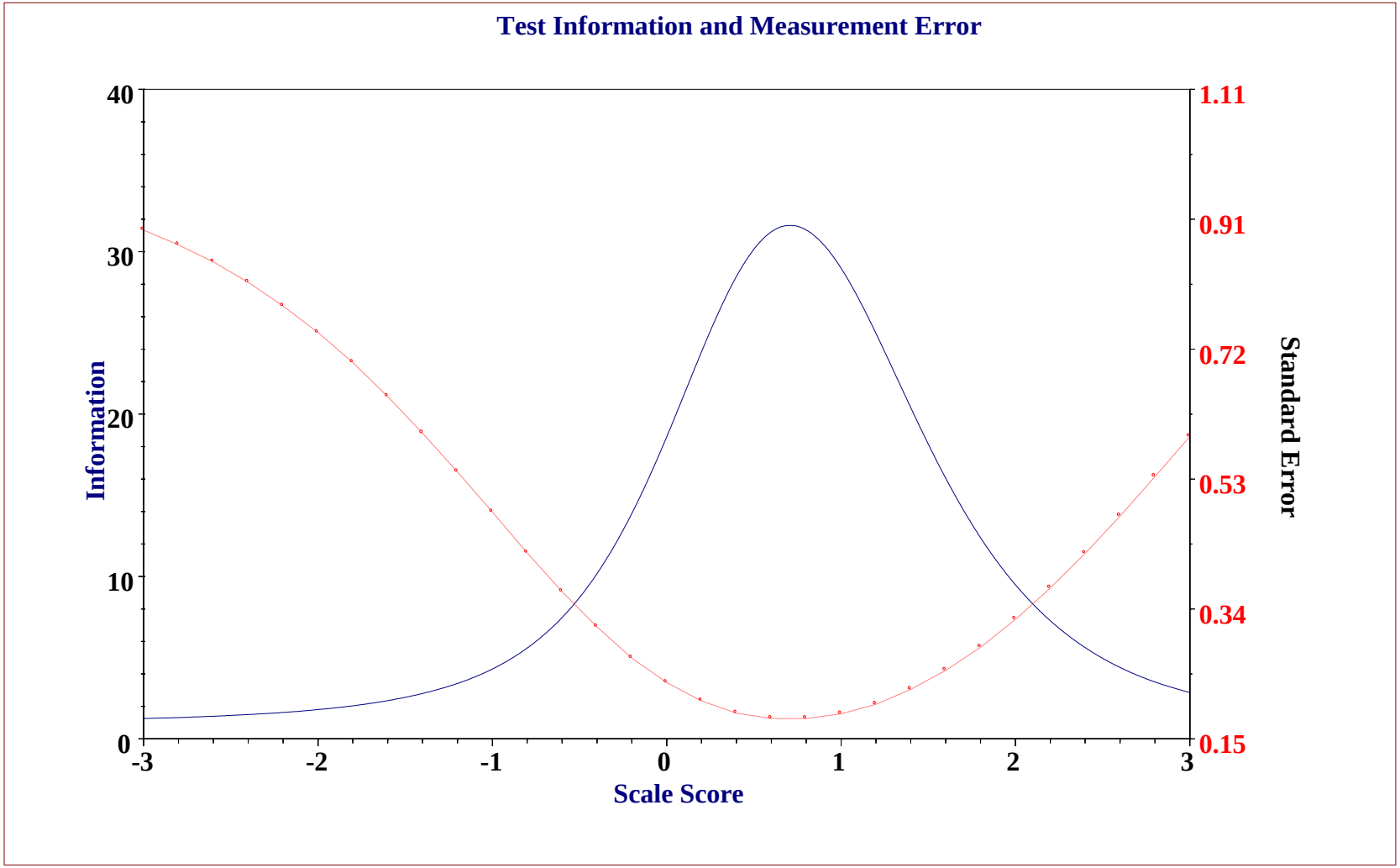


Figure 50 Mexico by Chile item 7

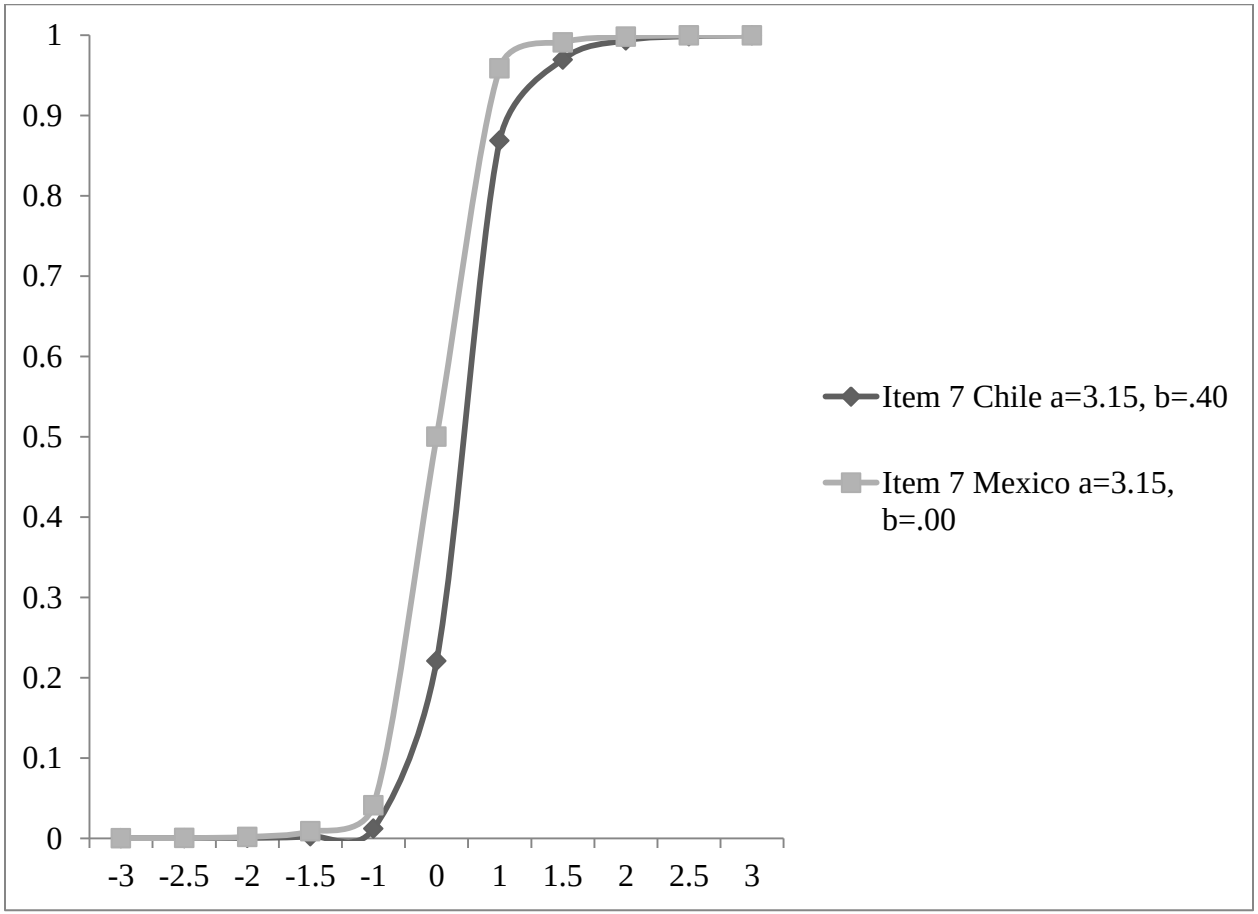


Figure 51 Mexico by Chile item 8

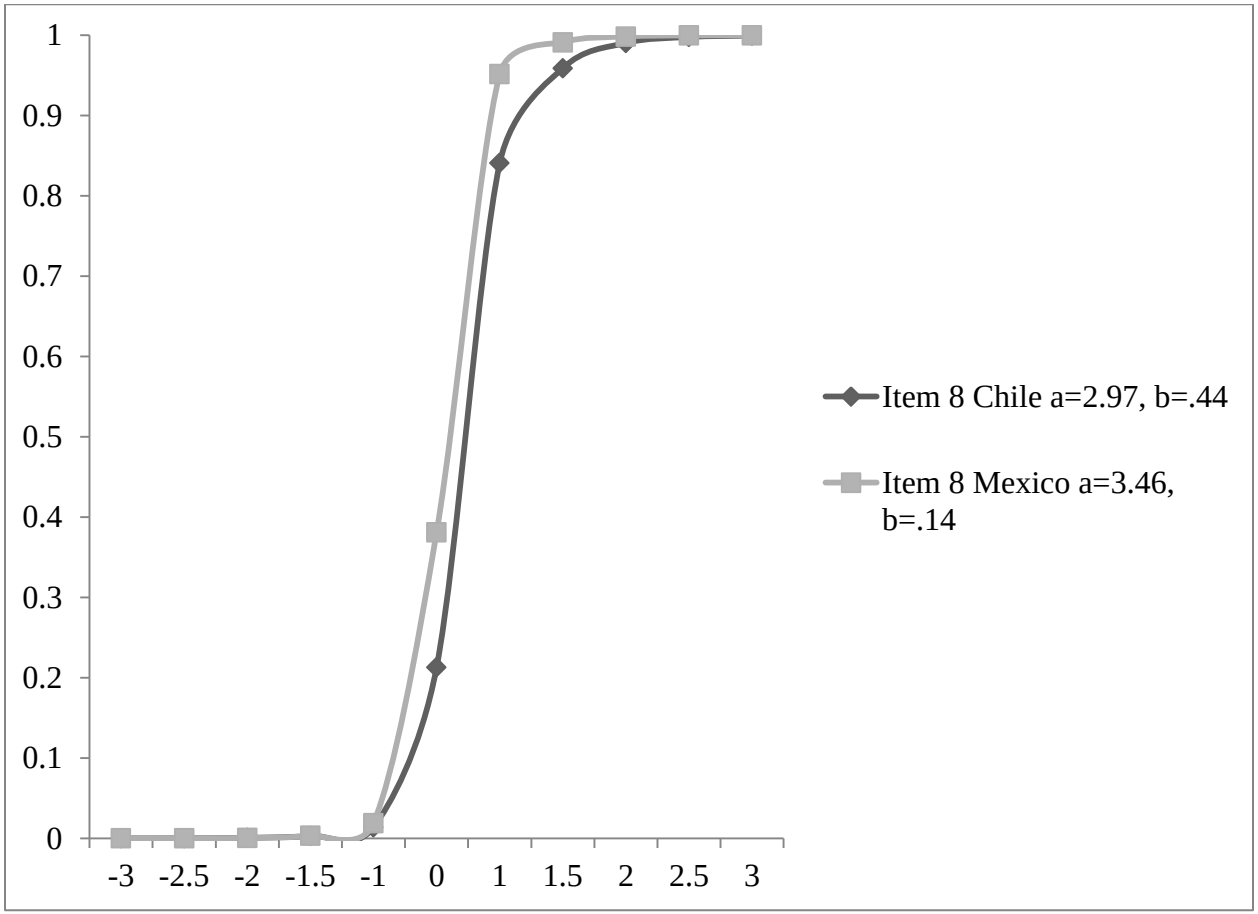


Figure 52 Mexico by Chile item 11

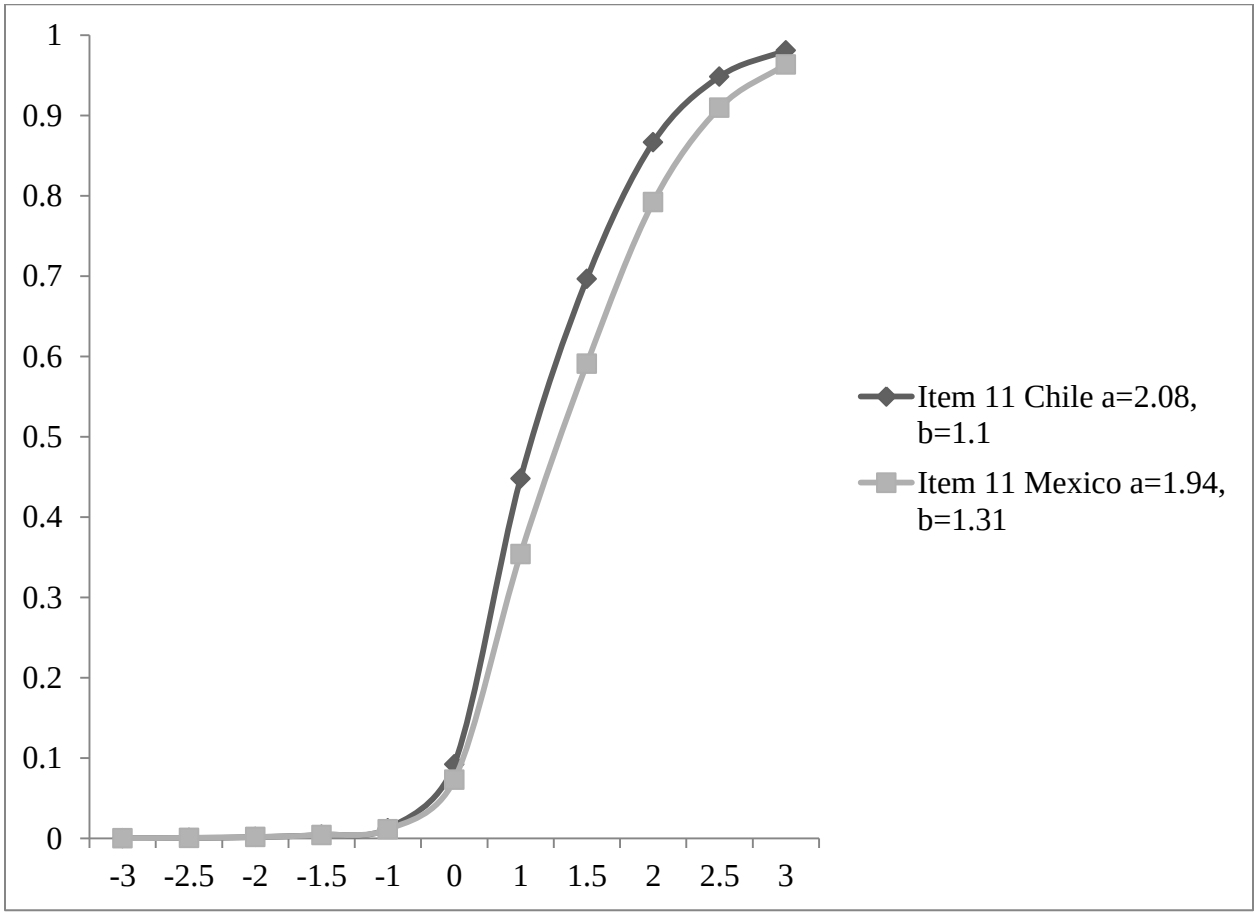


Figure 53 Mexico by Chile item 12

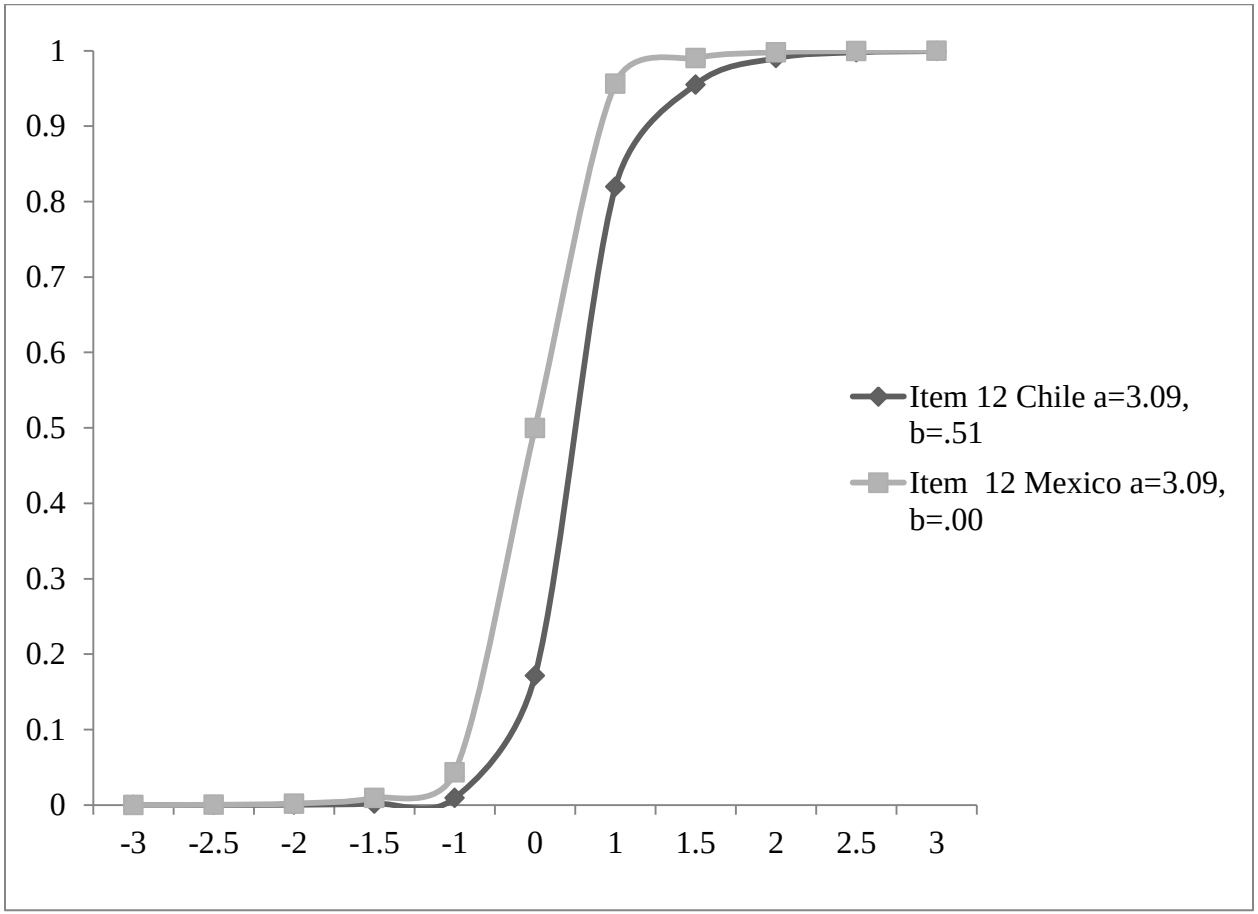


Figure 54 Mexico by Chile item 15

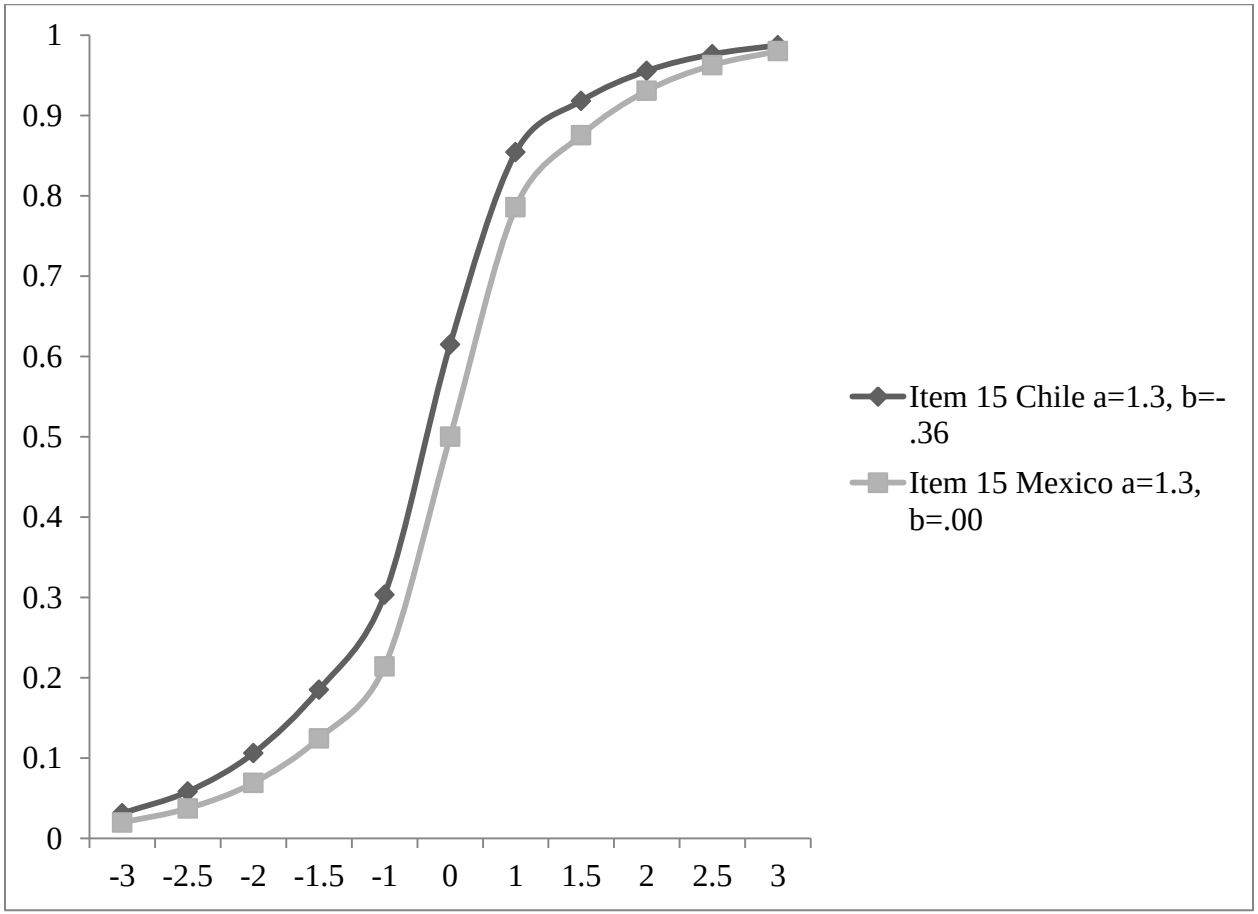


Figure 55 Mexico by Chile test information curve

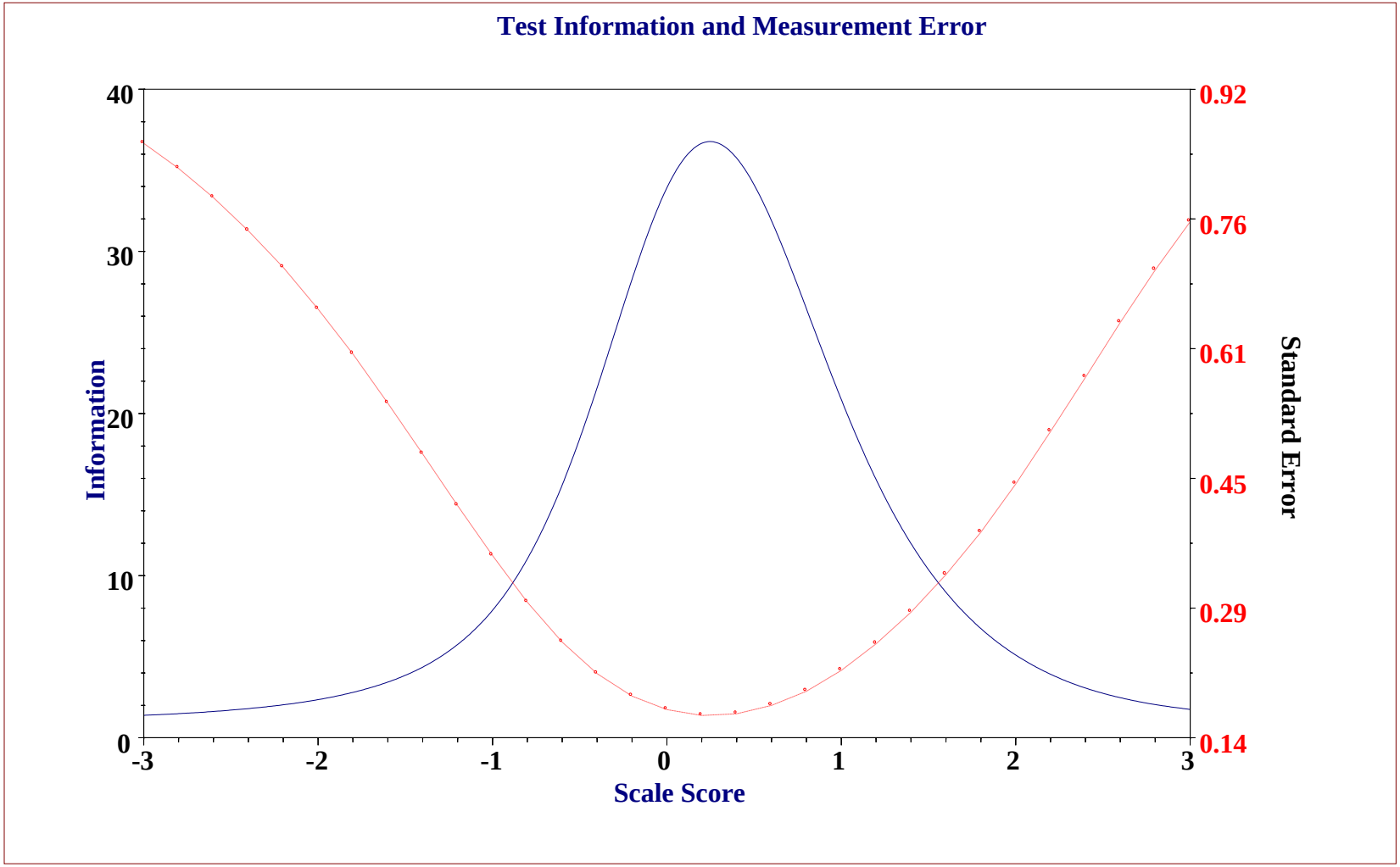


Figure 56 Mexico by Cuba item 2

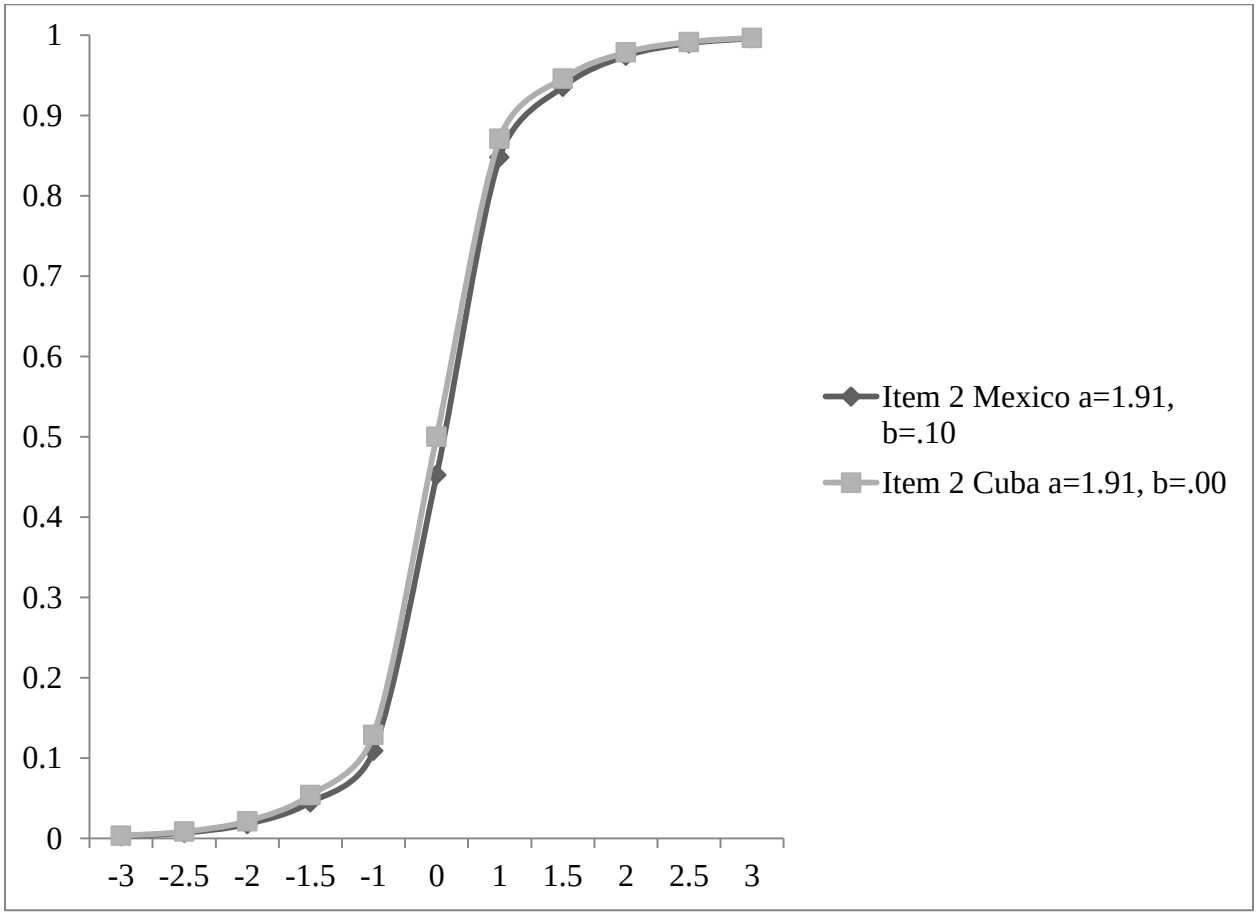


Figure 57 Mexico by Cuba item 4

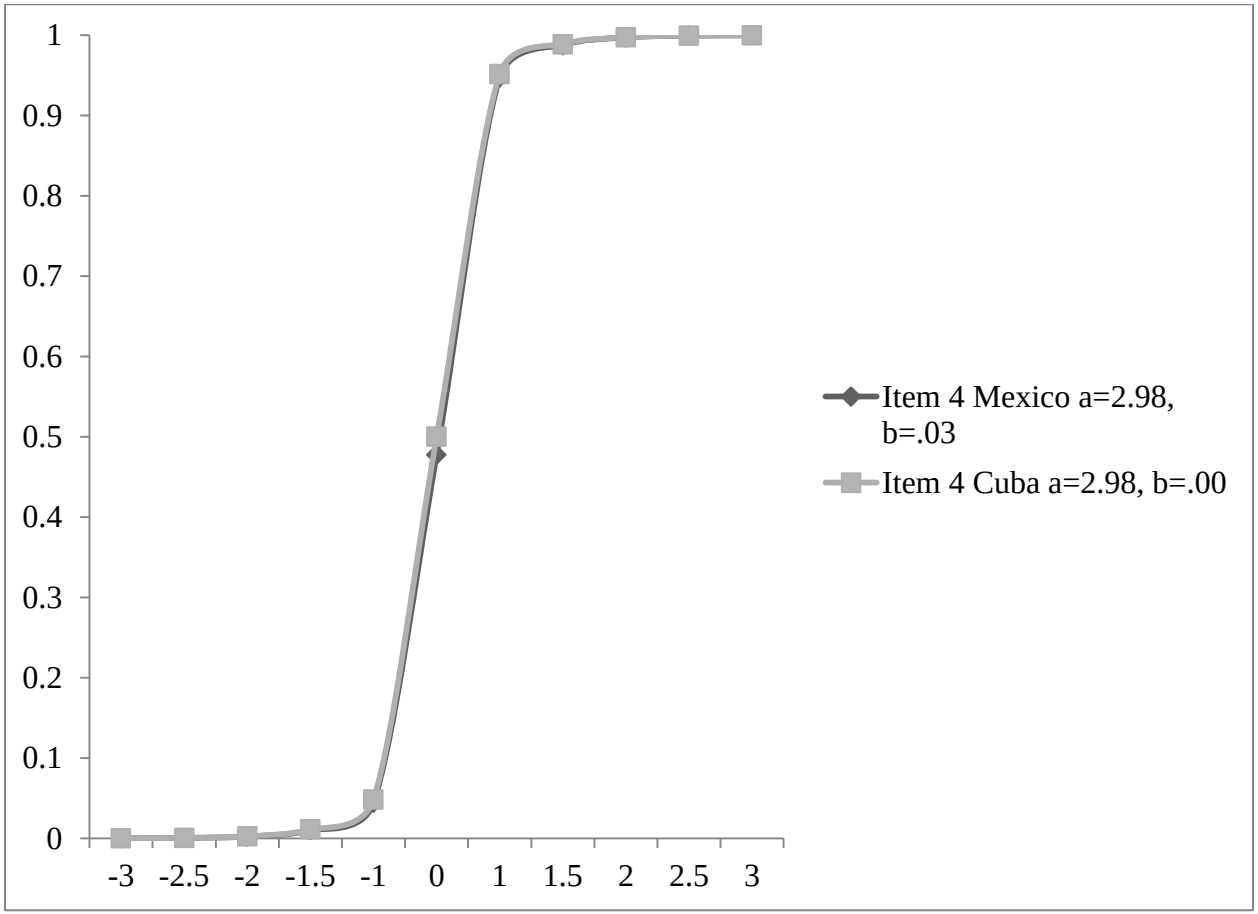


Figure 58 Mexico by Cuba item 7

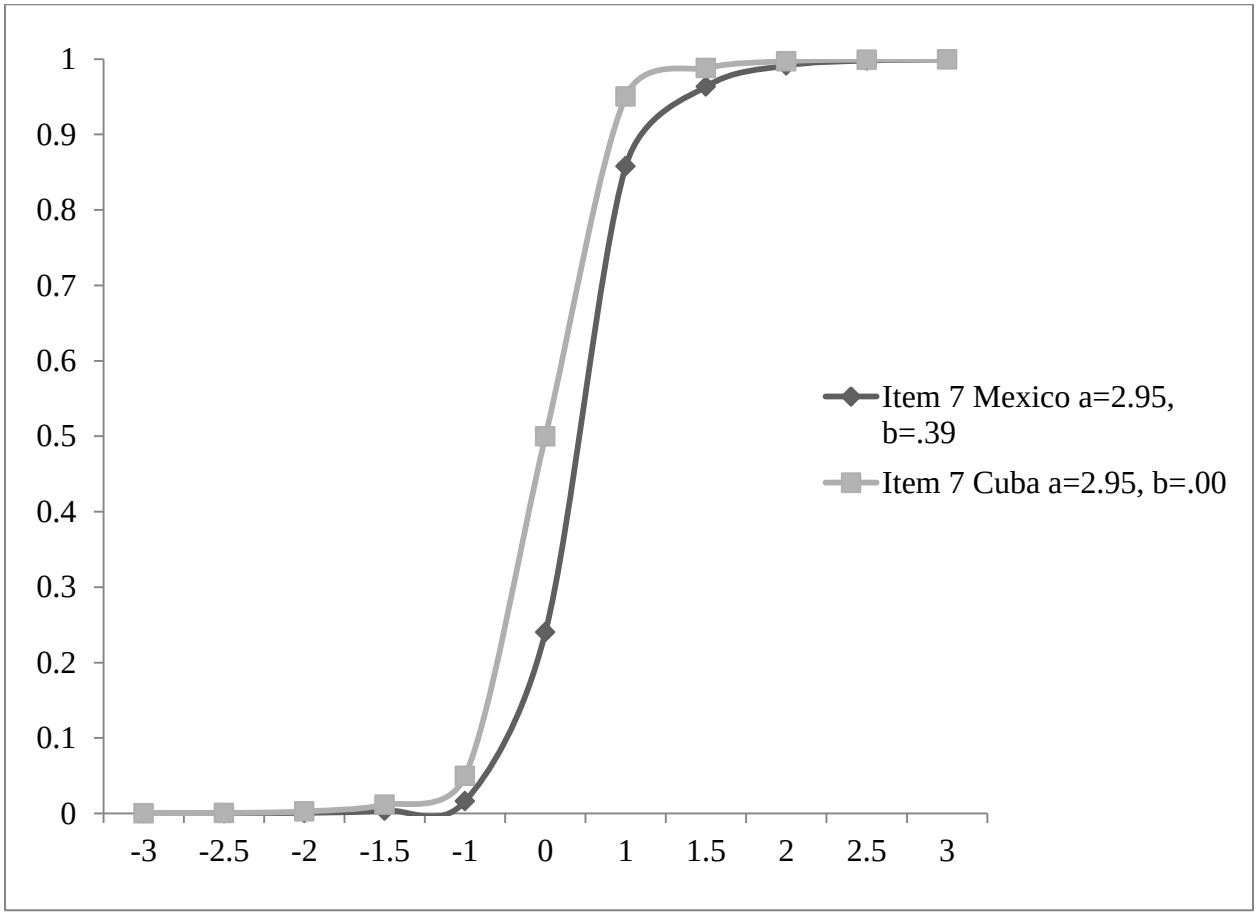


Figure 59 Mexico by Cuba item 11

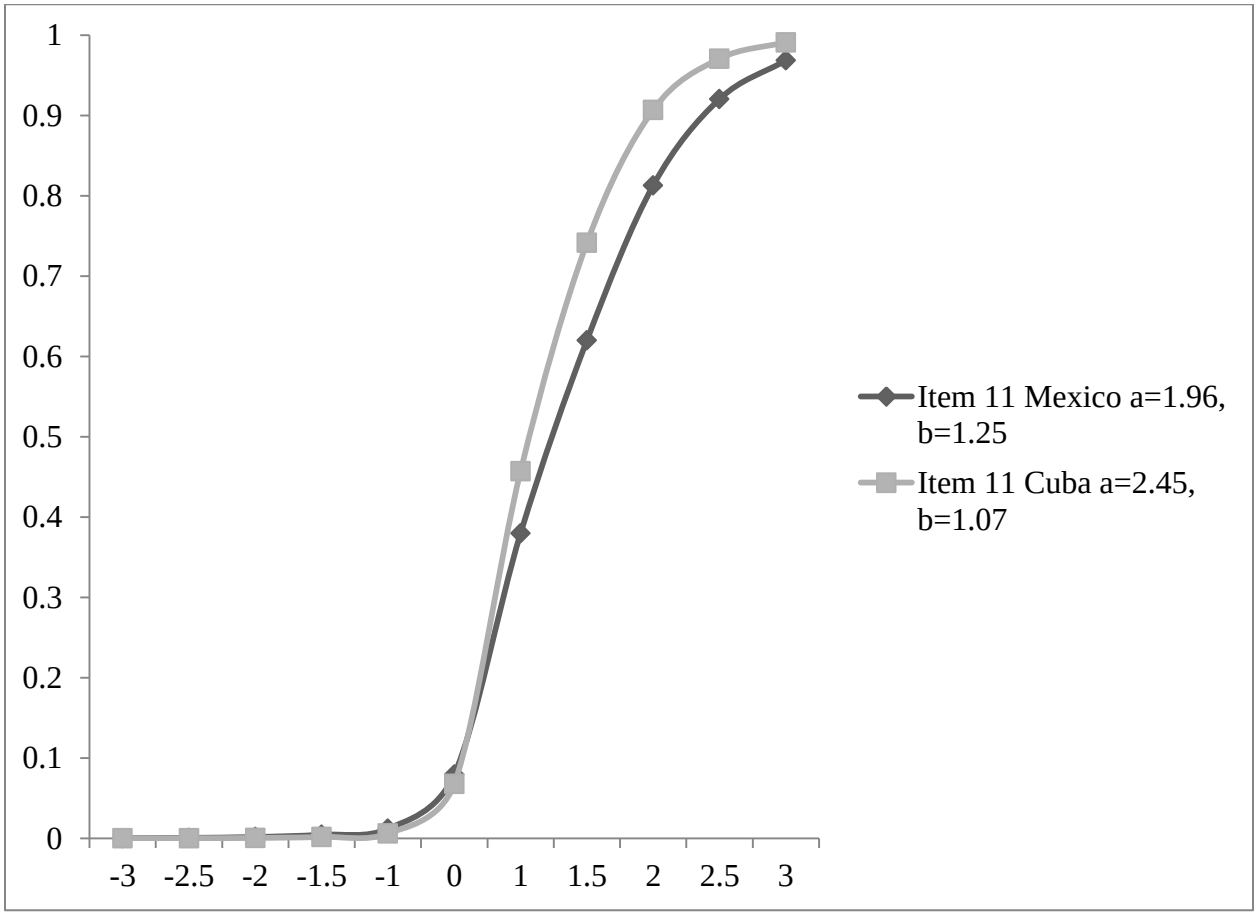


Figure 60 Mexico by Cuba item 12

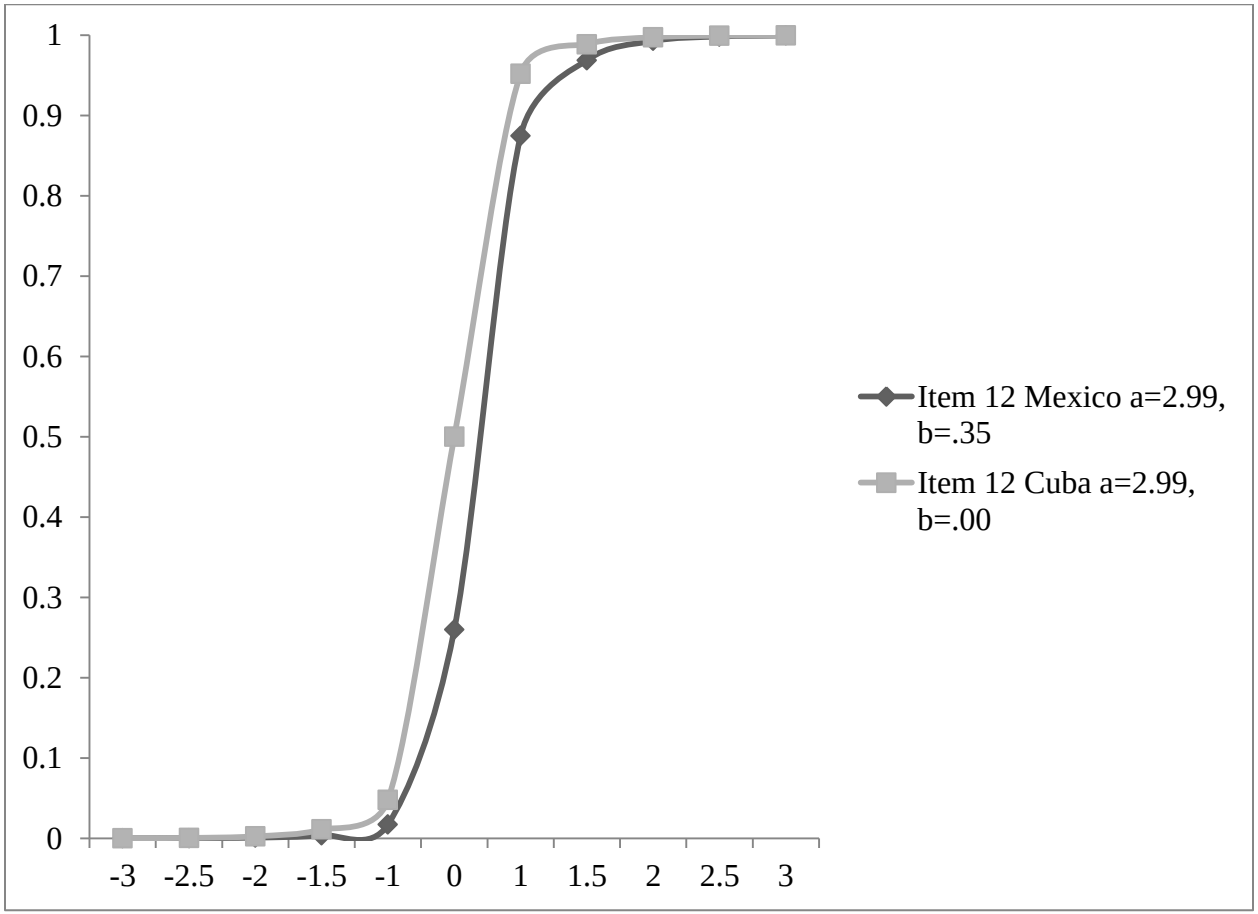


Figure 61 Mexico by Cuba item 14

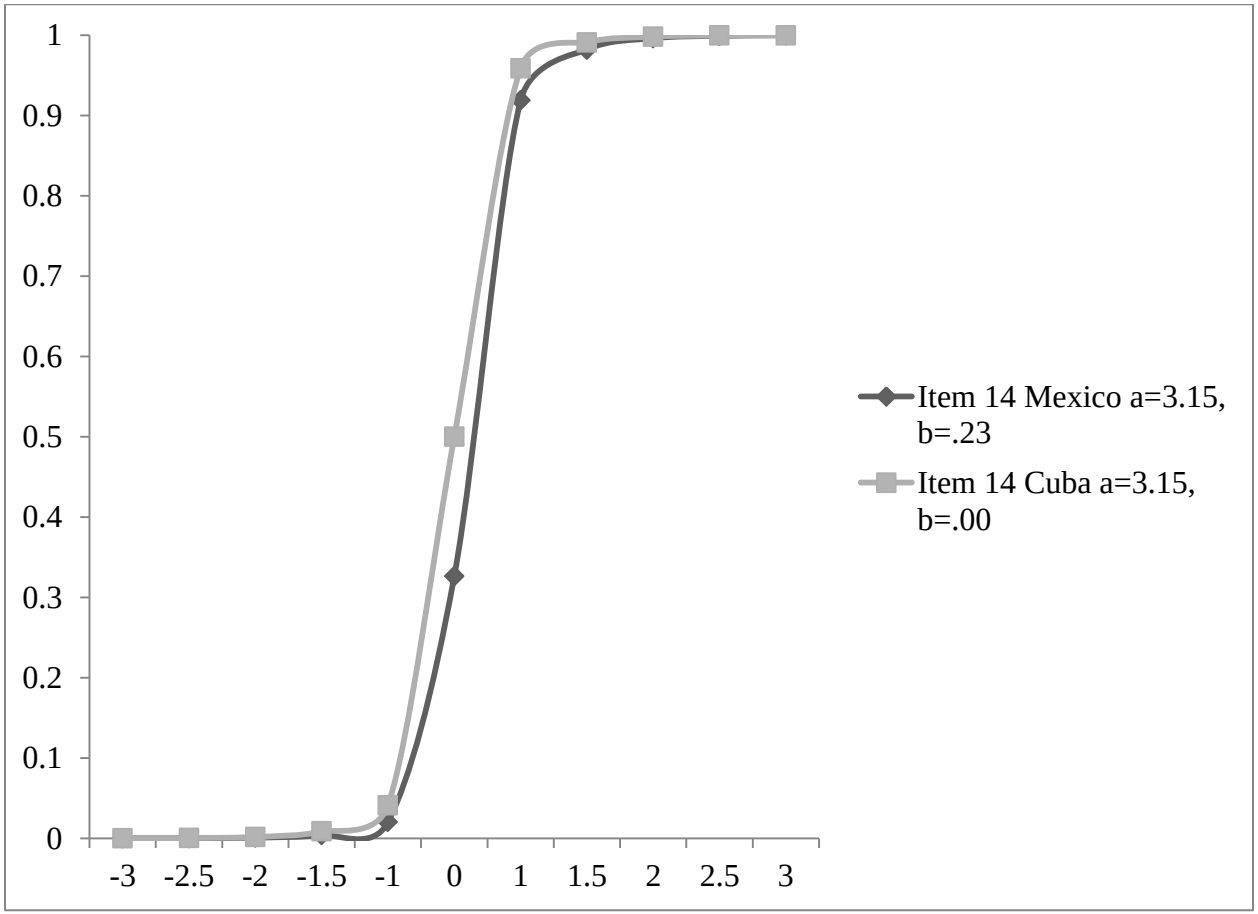


Figure 62 Mexico by Cuba test information curve

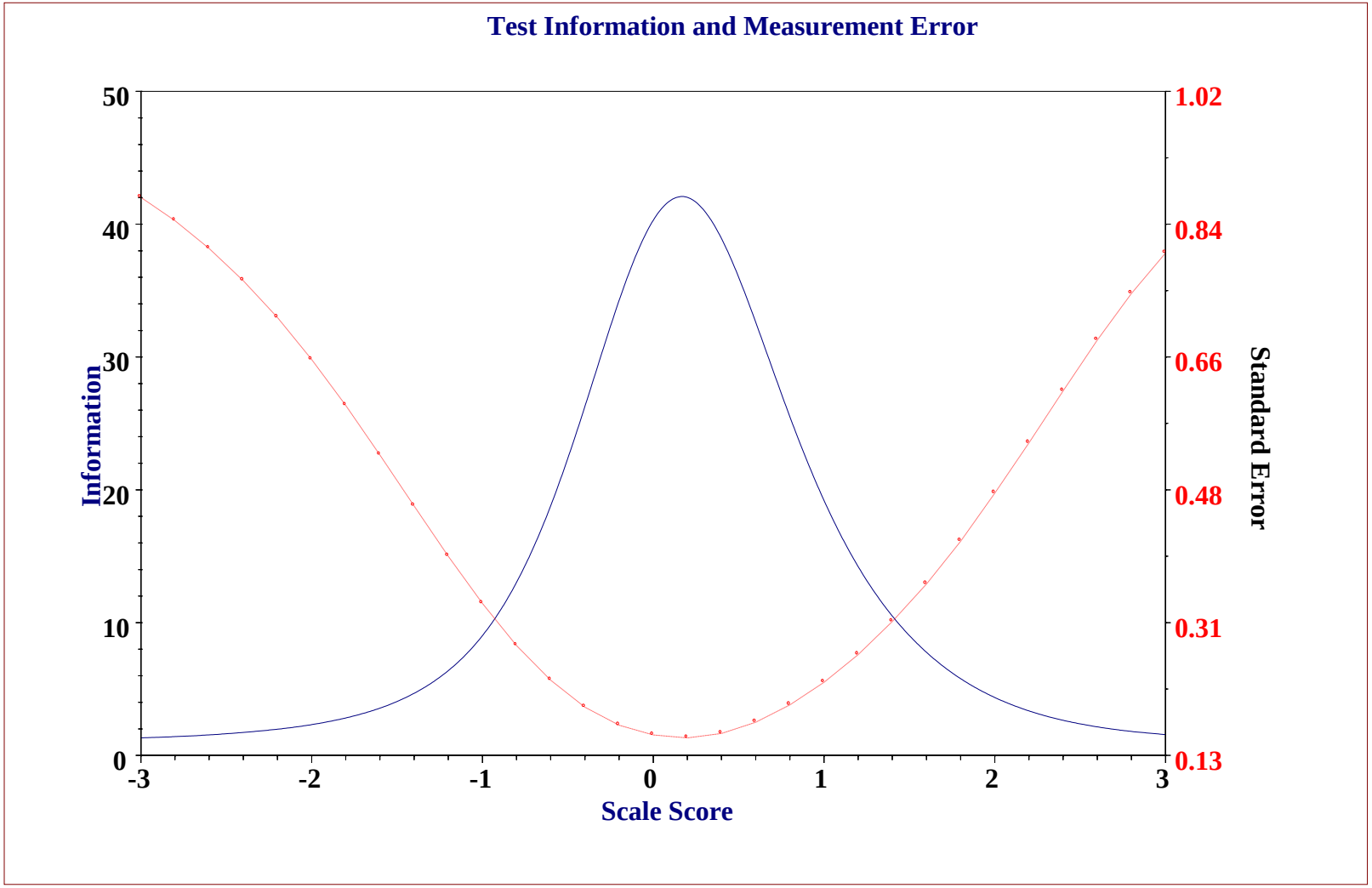


Figure 63 Uruguay by Chile item 2

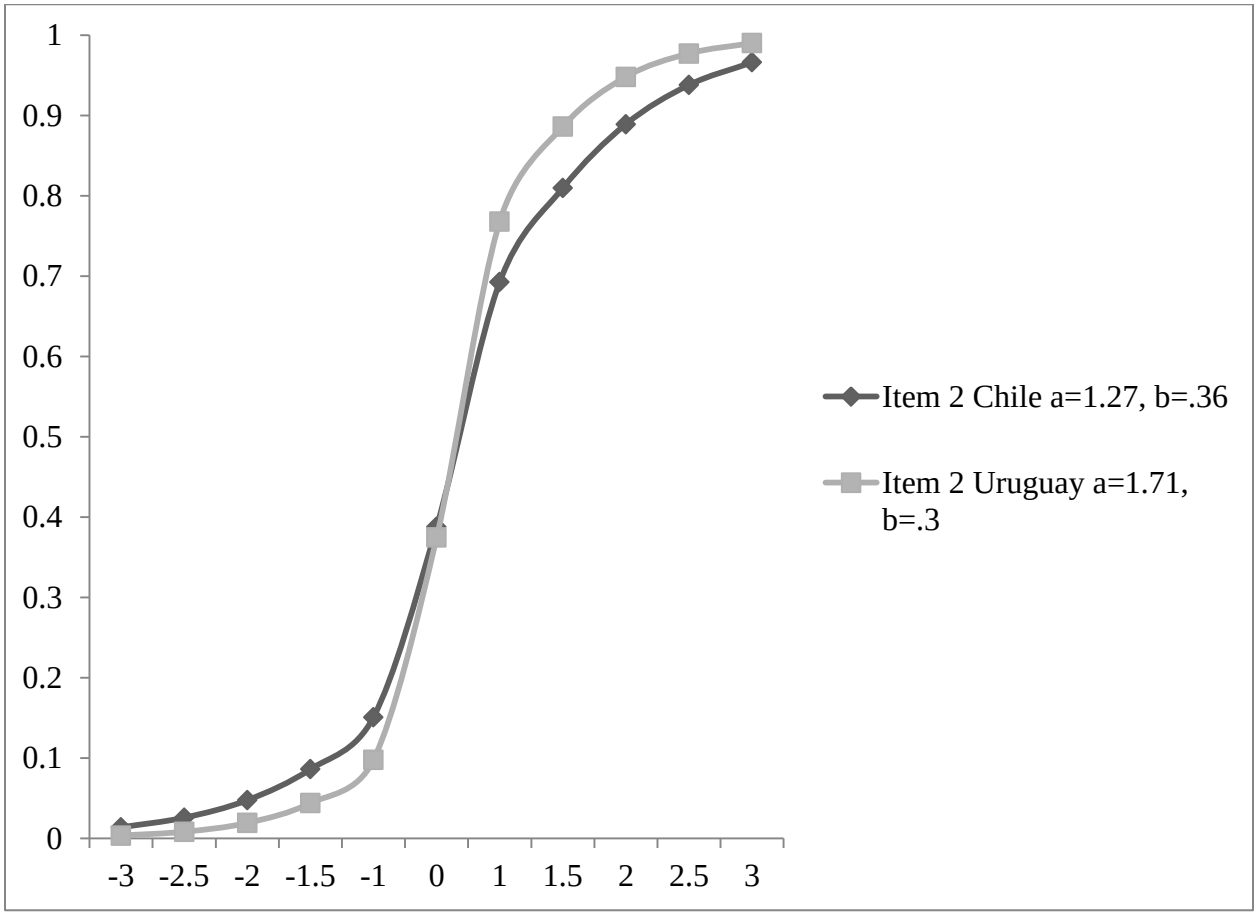


Figure 64 Uruguay by Chile item 3

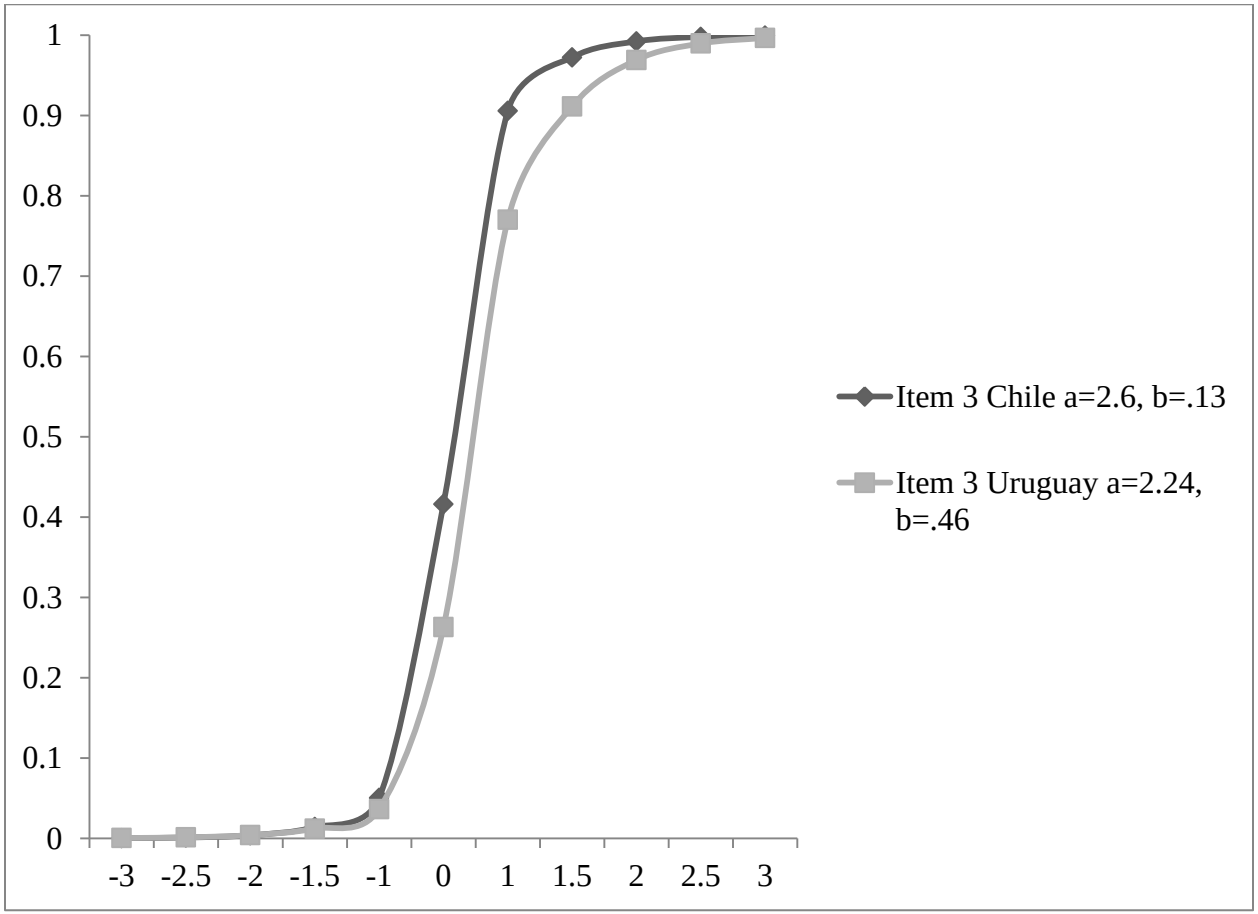


Figure 65 Uruguay by Chile item 10

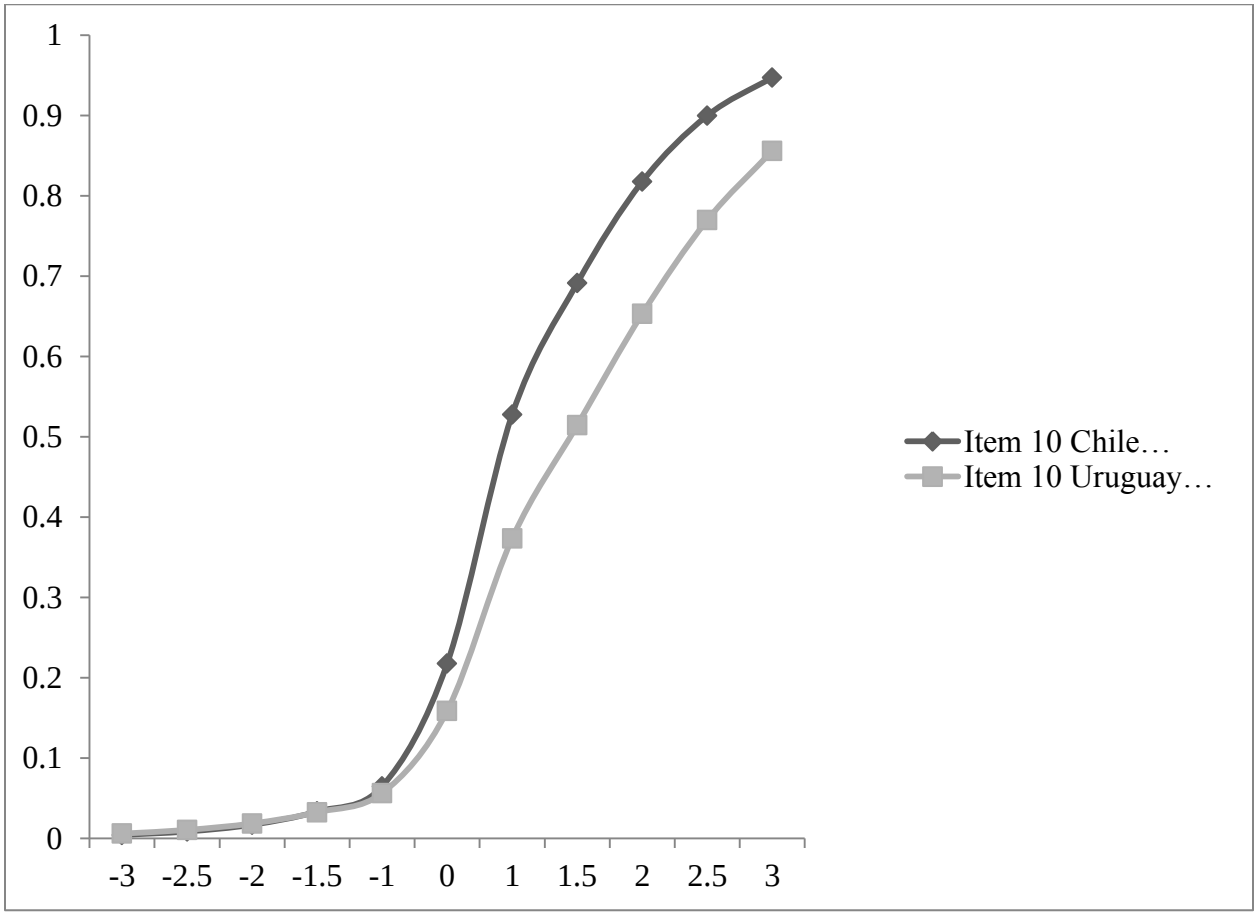


Figure 66 Uruguay by Chile item 14

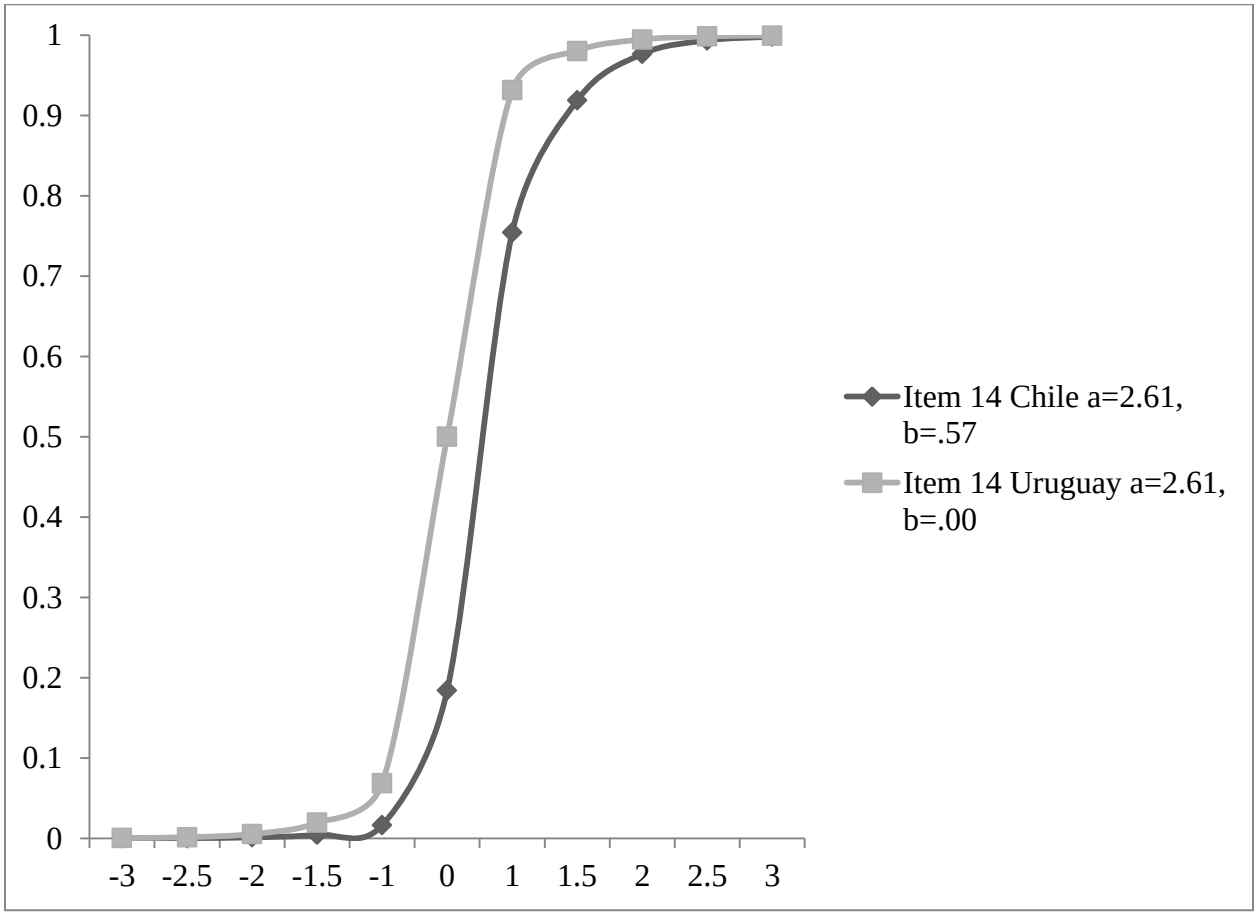


Figure 67 Uruguay by Chile test information curve

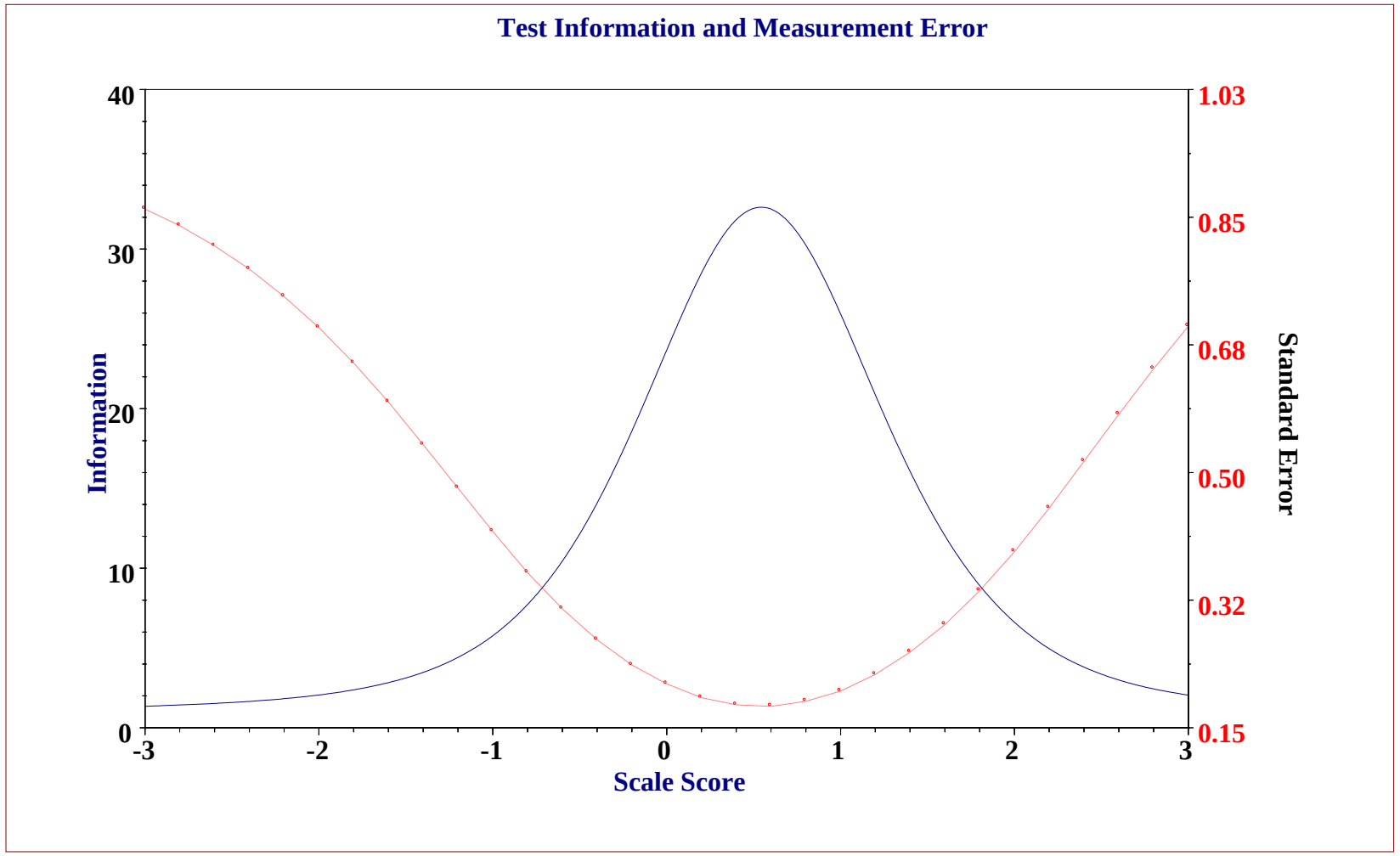


Figure 68 Uruguay by Cuba item 4

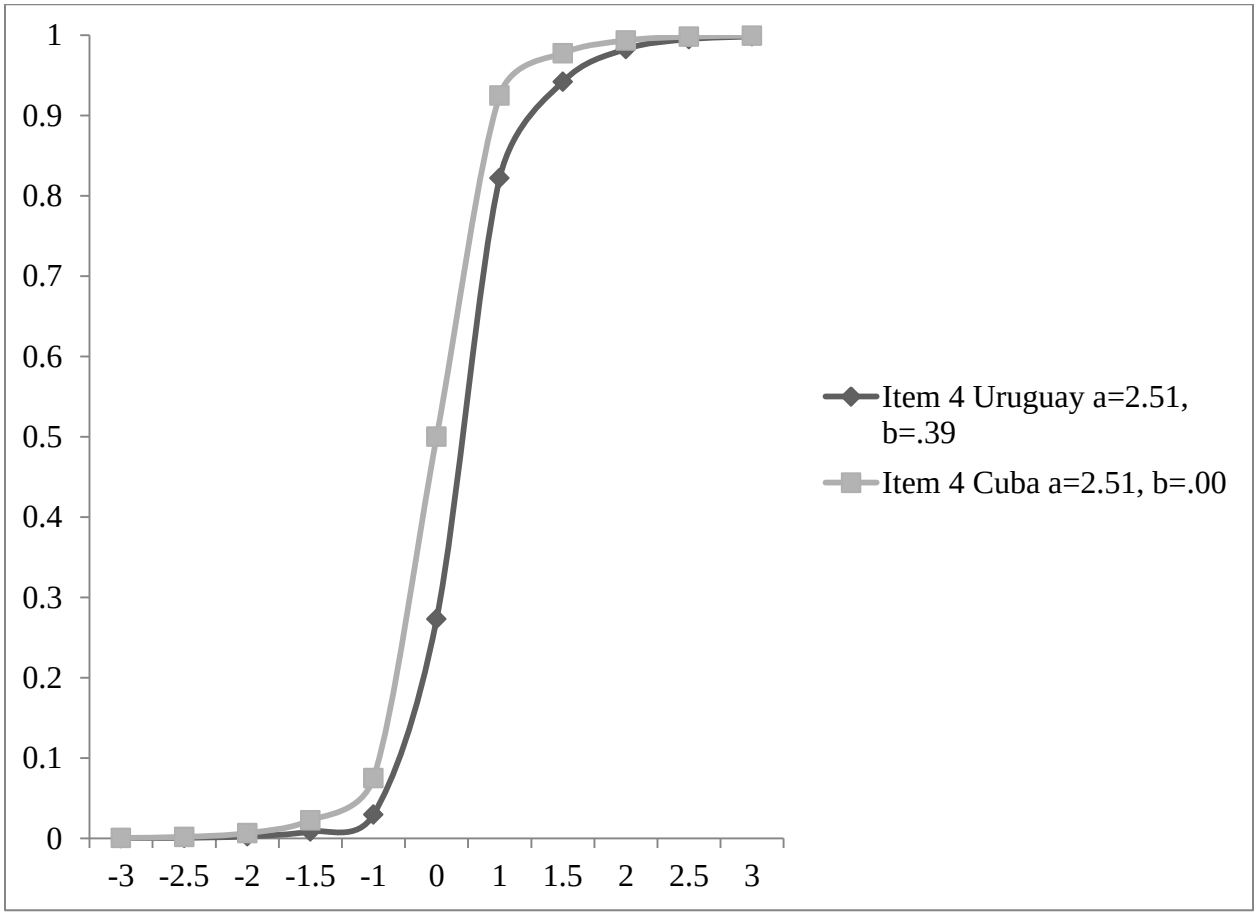


Figure 69 Uruguay by Cuba item 6

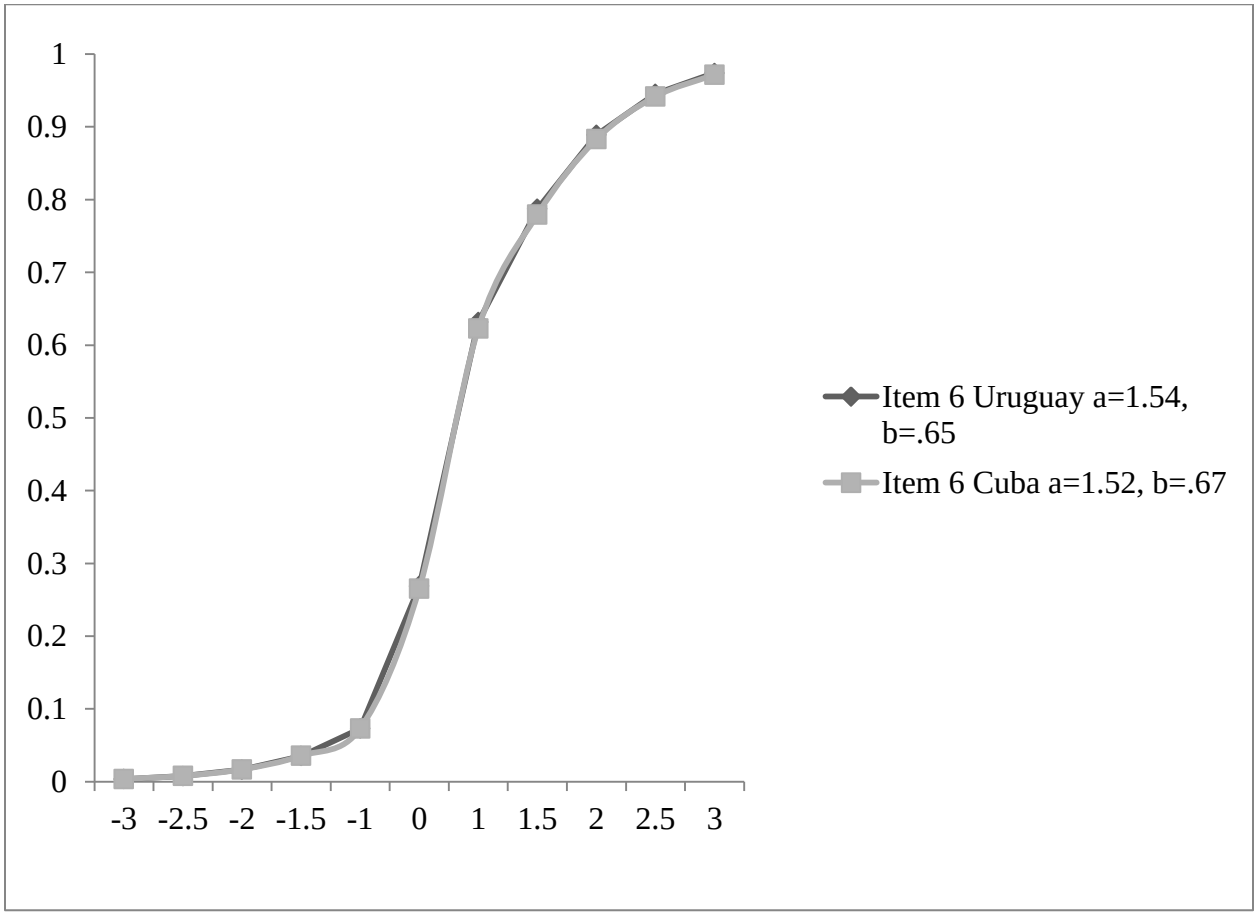


Figure 70 Uruguay by Cuba item 8

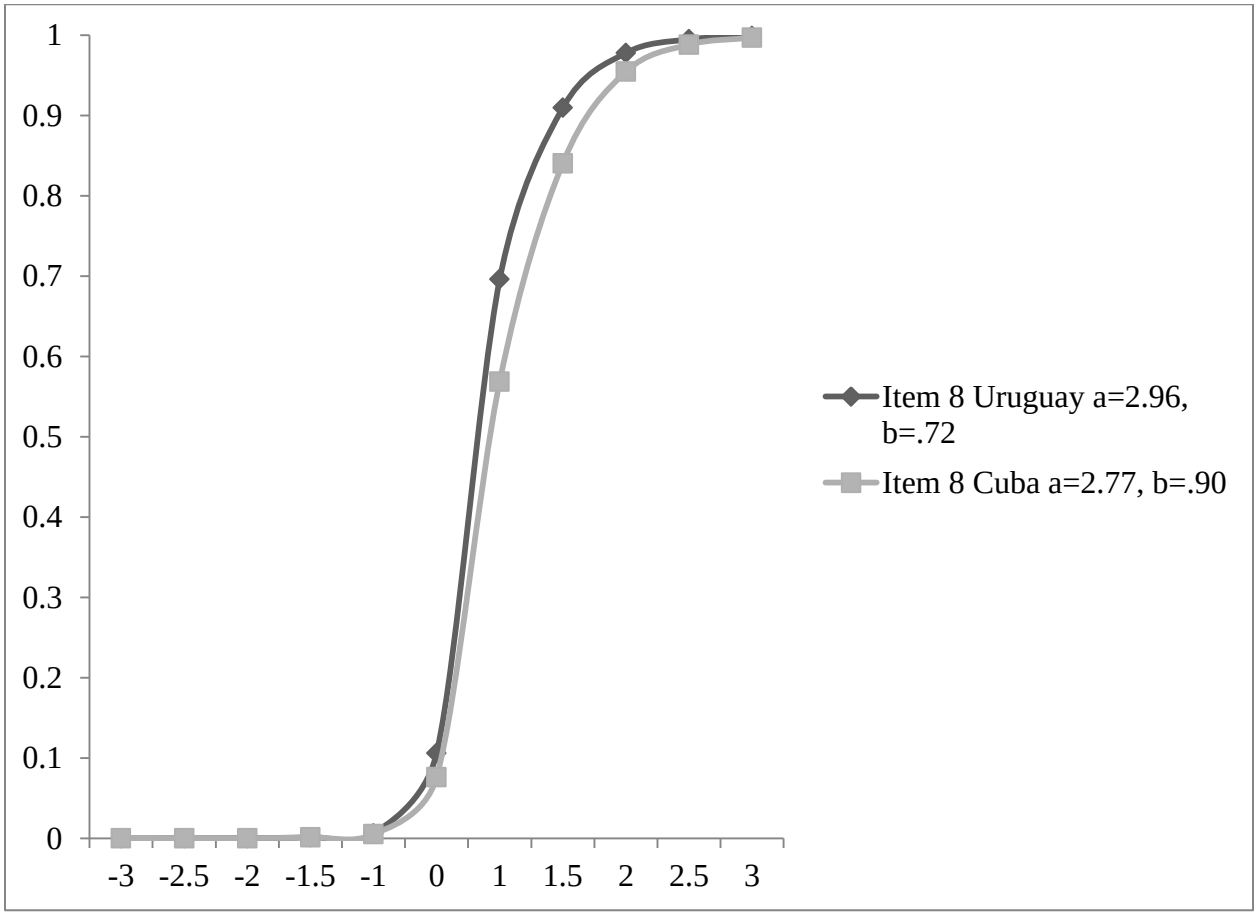
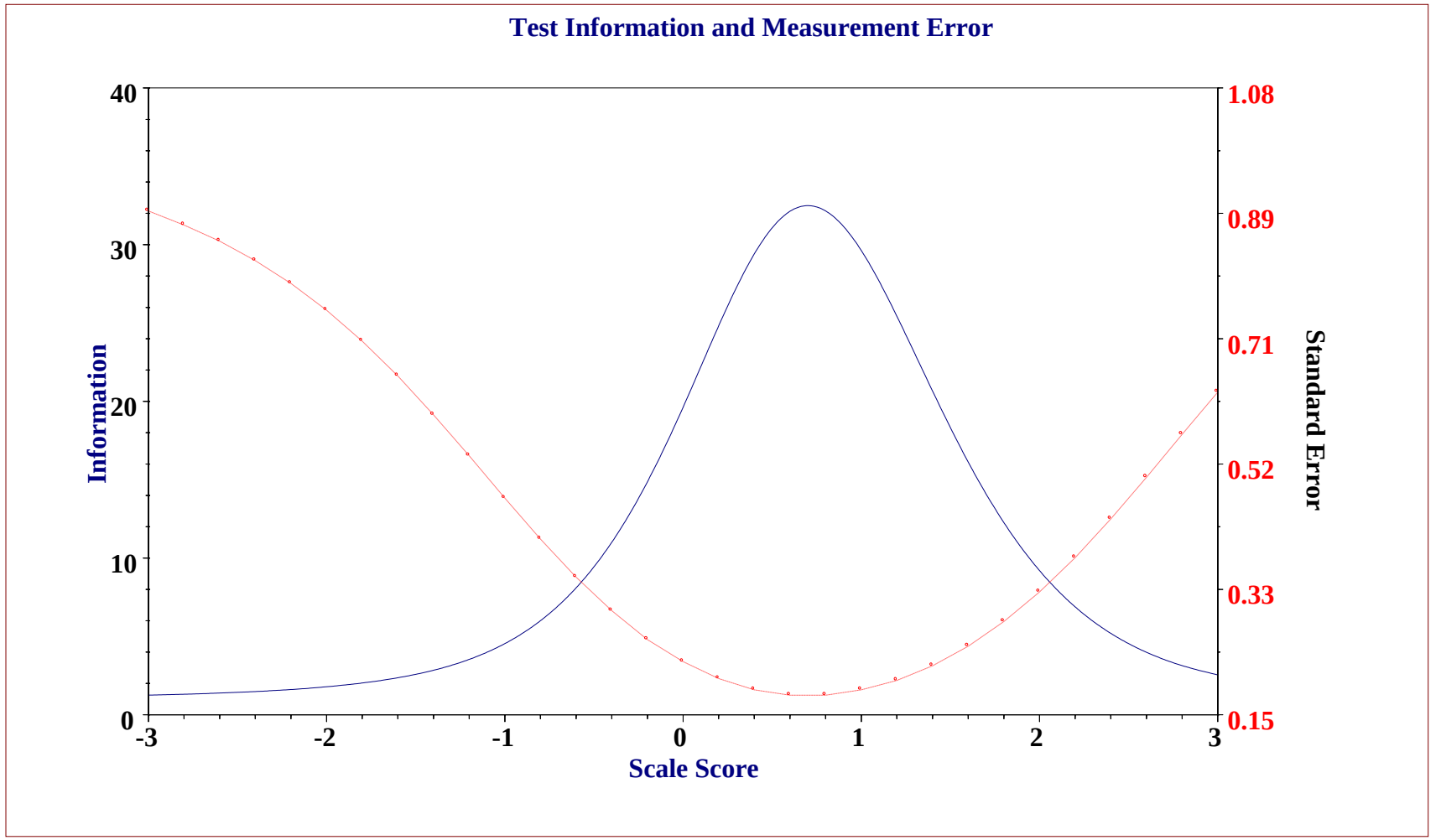


Figure 71 Uruguay by Cuba test information curve



Appendix

GERIATRIC DEPRESSION SCALE SHORT FORM

1. Are you basically satisfied with your life?
2. Have you dropped many of your activities and interests?
3. Do you feel that your life is empty?
4. Do you often get bored?
5. Are you in good spirits most of the time?
6. Are you afraid that something bad is going to happen to you?
7. Do you feel happy most of the time?
8. Do you often feel helpless?
9. Do you prefer to stay at home, rather than going out and doing new things?
10. Do you feel that you have more problems with memory than most?
11. Do you think that it is wonderful to be alive now?
12. Do you feel worthless the way you are now?
13. Do you feel full of energy?
14. Do you feel that your situation is hopeless?
15. Do you think that most people are better off than you are?

GERIATRIC DEPRESSION SCALE SHORT FORM SPANISH VERSION

1. En general , est· satisfecho/a con su vida?
2. Ha abandonado muchas de sus tareas habituales y aficiones?
3. Siente que su vida est· vacÌa?
4. Se siente con frecuencia aburrido/a?
5. Se encuentra de buen humor la mayor parte del tiempo?
6. Teme que algo malo pueda ocurrirle?
7. Se siente feliz la mayor parte del tiempo?
8. Con frecuencia se siente desamparado/a, desprotegido/a?
9. Prefiere usted quedarse en casa, m·s que salir y hacer cosas nuevas?
10. Cree que tiene m·s problemas de memoria que. la mayorÌa de la gente?
11. En estos momentos, piensa que es estupendo estar vivo?
12. Actualmente se siente una in_til?
13. Se siente lleno la de energÌa?
14. Se siente sin esperanza en este momento?
15. Piensa que la mayorÌa de la gente est· en mejor situaciÛn que usted?

References

References

Adams, K. B. (2001). Depressive symptoms, depletion, or developmental change? Withdrawal, apathy, and lack of vigor in the Geriatric Depression Scale. *The Gerontologist, 41*(6), 768-777.

Adams, K. B., Matto, H. C., & Sanders, S. (2004). Confirmatory Factor Analysis of the Geriatric Depression Scale. *The Gerontologist, 44*(6), 818-826.

Angst, J. (1992). How predictable is depressive illness? In S. Montgomery & F. Rouillon (Eds.), *Long-term treatment of depression* (pp. 1-15). New York, NY: Wiley.

Ayotte, B. J., Yang, F. M., & Jones, R. N. (2010). Physical Health and Depression: A Dyadic Study of Chronic Health Conditions and Depressive Symptomatology in Older Adult Couples. *Journal of Gerontology: Psychological Sciences, 65B*(4), 438-448.

Baumgartner, H., & SteenKamp, J.-B. E. M. (2001). Response Styles in Marketing Research:A Cross-National Investigation. *Journal of Marketing Research, 38*(2), 143-156.

Beckstead, J. W., Yang, C. Y., & Lengacher, C. A. (2008). Assessing cross-cultural validity of scales: A methodological review and illustrative example. *International Journal of Nursing Studies, 45*(1), 110-119.

Beekman, A. T. F., Kriegsman, D. M. W., Deeg, D. J. H., & Tilburg, W. (1995). The association of physical health and depressive symptoms in the older population: age and sex differences. *Social Psychiatry and Psychiatric Epidemiology, 30*(1), 32-38.

Beekman, A. T. F., Penninx, B. W. J. H., Deeg, D. J. H., Ormel, J., Braam, A. W., & Van Tilburg, W. (1997). Depression and physical health in later life: results from the Longitudinal Aging Study Amsterdam (LASA). *Journal of Affective Disorders, 46*, 219-231.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B, 57*, 289-300.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107*, 238-246.

Bentler, P. M., & Bonnet, D. G. (1980). Significance tests and goodness of fit on the analysis of covariance structures. *Psychological Bulletin, 88*, 588-606.

Berkman, L. F., Berkman, C. S., Kasl, S., Freeman, D. H., Leo, L., Ostfeld, A. M., et al. (1986). Depressive Symptoms in Relation to Physical Health and Functioning in the Elderly *American Journal of Epidemiology, 124*(3), 372-388.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. *Psychometrika, 46*, 443-460.

Bollen, K. A. (1989). *Structural equations with latent variables*. New York: Wiley.

Bontempo, D. E. (2007). Polytomous factor analytic models in developmental research. *Dissertation Abstracts International: Section B: The Sciences and Engineering, 67*(8-B), 4754.

Bontempo, D. E., & Hofer, S. M. (2007). Assessing factorial invariance in cross-sectional and longitudinal studies. In A. D. v. D. Ong, Manfred H (Ed.), *Oxford handbook of methods in positive psychology. Series in positive psychology* (pp. 153-175). New York, NY: Oxford University Press.

Bontempo, D. E., Hofer, S. M., Mackinnon, A., Gray, K., Einfeld, S., Tonge, B., et al. (2008). Factor structure of the Developmental Behavior Checklist using confirmatory factor analysis of polytomous items. *Journal of Applied Measurement, 9*(3), 265-280.

Braam, A. W., Prince, M. J., Beekman, A. T. F., Delespaul, P., Dewey, M. E., Geerlings, S. W., et al. (2005). Physical health and depressive symptoms in older Europeans: Results from EURODEP. *British Journal of Psychiatry, 187*, 35-42.

Braswell, J. S., Lutkus, A. D., Grigg, W. S., Santapau, S. L., Tay-Lim, B., & Johnson, M. (2001). *The nation's report card: Mathematics 2000. NCES 2001-517*.

Washington, DC: U.S. Department of Education, Office of Educational Research and Improvement, National Center for Education Statistics.

Brink, T., Yesavage, J. A., Lum, O., Heersema, P. H., Adey, M., & Rose, T. L. (1982). Screening Tests For Geriatric Depression. *Clinical Gerontologist, 1*(1), 37-43.

Brown, L. M., & Schinka, J. A. (2005). Development and initial validation of a 15-item informant version of the Geriatric Depression Scale. [Peer Reviewed]. *International Journal of Geriatric Psychiatry, 20*(10), 911-918. doi: 10.1002/gps.1375

Brown, P. J., Woods, C. M., & Storandt, M. (2007). Model stability of the 15-item Geriatric Depression Scale across cognitive impairment and severe depression. *Psychology and Aging, 22*(2), 372-379.

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. B. J. S. Long (Ed.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Byrne, B. M. (2008). Testing for multigroup equivalence of a measuring instrument: A walk through the process. *Psicothema, 20*(4), 872-882.

Byrne, B. M., & Campbell, T. L. (1999). Cross-cultural comparisons and the presumption of equivalent measurement and theoretical structure - A look beneath the surface. *Journal of Cross-Cultural Psychology, 30*(5), 555-574.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. *Psychological Bulletin, 105*(3), 456-466.

Byrne, B. M., & Watkins, D. (2003). The issue of measurement invariance revisited. *Journal of Cross-Cultural Psychology, 34*(2), 155-175.

Chaaya, M., Sibai, A. M., El Roueiheb, Z., Chemaitelly, H., Chahine, L. M., Al-Amin, H., et al. (2008). Validation of the Arabic version of the short Geriatric Depression Scale (GDS-15). *International Psychogeriatrics, 20*(3), 571-581.

Chau, J., Martin, C. R., Thompson, D. R., Chang, A. M., & Woo, J. (2006). Factor structure of the Chinese version of the Geriatric Depression Scale. *Psychology, Health & Medicine, 11*(1), 48-59.

Chen, Y. H., Rendina-Gobioff, G., & Dedrick, R. F. (2010). Factorial Invariance of a Chinese Self-Esteem Scale for Third and Sixth Grade Students: Evaluating Method Effects Associated with Positively and Negatively Worded Items. *The International Journal of Educational and Psychological Assessment, 6*(1), 21-35.

Cheng, S.-T., & Chan, A. C. M. (2004). A Brief Version of the Geriatric Depression Scale for the Chinese. *Psychological Assessment, 16*(2), 182-186.

Cheung, G. W., & Rensvold, R. B. (2002). Evaluating goodness-of-fit indexes for testing measurement invariance. *Structural Equation Modeling, 9*(2), 233-255.

Coleman, R. M., Miles, L. E., Guilleminault, C., Zarcone, V. P., Van Der Hoed, J., & Dement, W. C. (1981). Sleep-wake disorders in the elderly: a polysomnographic analysis. *J. Am. Geriat. Soc. , 29*, 289–296.

Djernes, J. K. (2006). Prevalence and predictors of depression in populations of elderly: a review. *Acta Psychiatrica Scandinavica, 113*(5), 372-387.

Eberhardt, M. S., & Pamuk, E. P. (2004). The importance of place of residence: Examining health in rural and nonrural areas. *American Journal of Public Health, 94*, 1682-1686.

Edelen, M. O., Thissen, D., Teresi, J. A., Kleinman, M., & Ocepek-Welikson, K. (2006). Identification of Differential Item Functioning Using Item Response Theory and the Likelihood-Based Model Comparison Approach: Application to the Mini-Mental State Examination. *Medical Care. Special Issue: Measurement in a multi-ethnic society, 44*(11, Suppl 3), S134-S142.

Edwards, M. C. (2009). An Introduction to Item Response Theory Using the Need for Cognition Scale. *Social and Personality Psychology Compass, 3*(4), 507-529.

Embretson, S. E. R., S.P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological Methods, 9*, 466-491.

Flowers, C., Oshima TC, & Raju, N. (1999). A description and demonstration of the polytomous DFIT framework. *Applied Psychological Measurement, 23*, 309-326.

Friedman, B., Heisel, M. J., & Delavan, R. L. (2005). Psychometric Properties of the 15-Item Geriatric Depression Scale in Functionally Impaired, Cognitively Intact, Community-Dwelling Elderly Primary Care Patients. *Journal of the American Geriatrics Society, 53*(9), 1570-1576.

Ganguli, M., Dube, S., Johnston, J. M., Pandav, R., Chandra, V., & Dodge, H. H. (1999). Depressive symptoms, cognitive impairment and functional impairment in a rural elderly population in India: A Hindi version of the Geriatric Depression Scale *International Journal of Geriatric Psychiatry, 14*(10), 807-820.

Hartley, D. (2004). Rural health disparities, population health, and rural culture. *American Journal of Public Health, 94*, 1675-1678.

Hays, R. D., Morales, L. S., & Reise, S. P. (2000). Item Response Theory and Health Outcomes Measurement in the 21st century. *Medical Care, 28*(9SII), II-28-II-42.

Holland, P. W., & Thayer, D. T. (1988). Differential Item Performance and the Mantel-Haenszel Procedure. In H. Wainer & H. I. Braun (Eds.), *Test Validity* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum Associates.

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*, 179-185.

Horn, J. L. (1991). Comments on Issues in Factorial Invariance. In L. M. C. a. J. L. Horn (Ed.), *Best Methods for the Analysis of Change* (pp. 114-125). Washington, DC: American Psychological Association.

Horn, J. L., McArdle, J. J., & Mason, R. (1983). When is invariance not invariant: A practical scientist's look at the ethereal concept of factor invariance. *Southern Psychologist, 1*, 179-188.

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*, 1-55.

Incalzi, R. A., Cesari, M., Pedone, C., & Carbonin, P. U. (2003). Construct validity of the 15-item Geriatric Depression Scale in older medical inpatients. *Journal of Geriatric Psychiatry and Neurology, 16*(1), 23-28.

Jöreskog, K. G. (1971). Simultaneous factor analysis in several populations. *Psychometrika, 36*, 409-426.

Jöreskog, K. G., & Moustaki, I. (2001). Factor analysis of ordinal variables: A comparison of three approaches. *Multivariate Behavioral Research, 36*(3), 347-387.

Lai, D., Tong, H. M., Zeng, Q., & Xu, W. Y. (2010). The factor structure of a Chinese Geriatric Depression Scale-SF: use with alone elderly Chinese in Shanghai, China. *International Journal of Geriatric Psychiatry, 25*(5), 503-510.

Lai, D. W. L., Fung, T. S., & Yuen, C. T. Y. (2005). The Factor Structure of A Chinese Version of The Geriatric Depression Scale. *International Journal of Psychiatry in Medicine, 35*(2), 137-148.

MacCallum, R. C., Roznowski, M., & Necowitz, L. B. (1992). Model Modifications in Covariance Structure Analysis: The Problem of Capitalization on Chance. *Psychological Bulletin, 111* (May), 490-504.

Malakouti, S. K., Fatollahi, P., Mirabzadeh, A., Salavati, M., & Zandi, T. (2006). Reliability, validity and factor structure of the GDS-15 in Iranian elderly. *International Journal of Geriatric Psychiatry, 21*, 588-593.

Meredith, W. (1993). Measurement Invariance, Factor-Analysis and Factorial Invariance. *Psychometrika, 58*(4), 525-543.

Millsap, R. E. (1995). Measurement invariance, predictive invariance, and the duality paradox. *Multivariate Behavioral Research, 30*(4), 577-605.

Millsap, R. E. (1997). Invariance in measurement and prediction: Their relationship in the single-factor case. *Psychological Methods, 2*(3), 248-260.

Millsap, R. E., & Everson, H. T. (1993). Methodology review: statistical approaches for assessing measurement bias. *Applied Psychological Measurement, 17*, 297-334.

Millsap, R. E., & Yun-Tein, J. (2004). Assessing Factorial Invariance in Ordered-Categorical Measures. *Multivariate Behavioral Research, 39*(3), 479-515.

Mitchell, J., Matthews, H. F., & Yesavage, J. A. (1993). A Multidimensional Examination of Depression Among the Elderly. *Research on Aging, 15*(2), 198-219.

Muthén, B., & Asparouhov, T. (2002). Latent Variable Analysis With Categorical Outcomes: Multiple-Group And Growth Modeling In Mplus. . *Mplus Web Notes: No. 4 Version 5.* Retrieved from http://www.statmodel.com/download/webnotes/CatMGLong.pdf website:

Muthen, B., & Christoffersson, A. (1981). SIMULTANEOUS FACTOR-ANALYSIS OF DICHOTOMOUS-VARIABLES IN SEVERAL GROUPS. *Psychometrika, 46*(4), 407-419.

Muthén, B., du Toit, S. H. C., & Spisic, D. (1997). Robust inference using weighted least squares and quadratic estimating equations in latent variable modeling with categorical and continuous outcomes. *Unpublished manuscript.*

Muthén, L., & Muthén, B. (2008). *Mplus user's guide* ((5th ed.) ed.). Los Angeles: Muthén & Muthén.

Muthén, L., & Muthén, B. O. (2008). *Mplus user's guide* ((5th ed.) ed.). Los Angeles: Muthén & Muthén.

Myers, M. B., Calantone, R. J., Page, T. J., & Taylor, C. R. (2000). Academic insights: An application of multiple-group causal models in assessing cross-cultural measurement equivalence. *Journal of International Marketing, 8*(4), 108-121.

Onishi, J., Suzuki, Y., Umegaki, H., Kawamura, T., Iguchi, A., & Endo, H. (2006). A Comparison of Depressive Mood of Older Adults in a Community, Nursing Homes, and a Geriatric Hospital: Factor Analysis of Geriatric Depression Scale. *Journal of Geriatric Psychiatry and Neurology, 19*(1), 26-31.

Ormel, J., Kempen, G. I. J. M., Penninx, B. W. J. H., Brilman, E. I., Beekman, A. T. F., & Van Sonderen, E. (1997). Chronic medical conditions and mental health in older people: disability and psychosoical resources mediate specific mental health effects. *Psychological Medicine, 27*, 1065-1077.

Oshima TC, & Morris, S. (2008). Raju's Differential Functioning of Items and Tests (DFIT). *NCME Instructional Module, Fall 2008,* 43-50.

Parmelee, P. A., & Katz, I. R. (1990). Geriatric Depression Scale. *Journal of the American Geriatrics Society, 38*(12), 1379.

Parmelee, P. A., Lawton, M. P., & Katz, I. R. (1989). Psychometric properties of the Geriatric Depression Scale among the institutionalized aged. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 1*(4), 331-338.

Pelaez, M., Palloni, A., Albala, C., Alfonso, J. C., Ham-Chande, R., Hennis, A., et al. (2004). SABE - Survey on Health, Well-Being, and Aging in Latin America and The Caribbean, 2000. Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

Pocinho, M. T. S., Farate, C., Dias, C. A., Yesavage, J. A., & Lee, T. T. (2009). Clinical and psychometric validation of the Geriatric Depression Scale (GDS) for Portuguese elders. *Clinical Gerontologist, 32*(2), 223-236.

Raju, N., van der Linden WJ, & Fleer, P. (1995). IRT-based internal measures of differential item functioning of items and tests. *Applied Psychological Measurement, 19,* 353-368.

Ramirez, M., Ford, M. E., Stewart, A. L., & Teresi, J. A. (2005). Measurement Issues in Health Disparities Research. *Health Services Research, 40*(5 Part 2), 1640-1657.

Reckase, M. D. (1979). Unifactor latent trait models applied to multi-factor tests: Results and implications. *Journal of Educational Measurement, 4,* 207-230.

Reise, S. P. (2005). Item response theory and its applications for cancer outcomes measurement. In J. Lipscomb, C. C. Gotay & C. Snyder (Eds.), *Outcomes assessment in cancer: Measures, methods, and applications* (pp. 425-444). New York, NY: Cambridge University Press.

Reise, S. P., Ainsworth, A. T., & Haviland, M. G. (2005). Item response theory - Fundamentals, applications, and promise in psychological research. *Current Directions in Psychological Science, 14*(2), 95-101.

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory Factor-Analysis And Item Response Theory - 2 Approaches For Exploring Measurement Invariance. *Psychological Bulletin, 114*(3), 552-566.

Roszkowski, M. J., & Soven, M. (2010). Shifting gears: consequences of including two negatively worded items in the middle of a positively worded questionnaire. *Assessment & Evaluation in Higher Education, 35*(1), 117-134.

Salamero, M., & Marcos, T. (1992). Factor study of the Geriatric Depression Scale. *Acta Psychiatrica Scandinavica, 86*(4), 283-286.

Schreiner, A. S., Morimoto, T., & Asano, H. (2001). Depressive symptoms among postroke patients in Japan: frequency distribution and factor structure of the GDS. *International Journal of Geriatric Psychiatry, 16*, 941-949.

Schriesheim, C. A., Eisenbach, R. J., & Hill, K. D. (1991). The Effect of Negation and Polar Opposite Item Reversals on Questionnaire Reliability and Validity: An Experimental Investigation. *Educational and Psychological Measurement, 51*, 67-78.

Sheikh, J. I., & Yesavage, J. A. (1986). Geriatric Depression Scale (GDS): Recent evidence and development of a shorter version. *Clinical Gerontologist, 5*(1-2), 165-173.

Sörbom, D. (1974). A general method for studying differences in factor means and factor structure between groups. *British J. of Mathematical and Stat. Psychology 27*, 229-239.

Sörbom, D. (1982). Structural equation models with structured means. In K. G. a. W. Jöreskog, H. (Ed.), *Systems under indirect observation: Causality, structure, prediction*. North-Holland, Amsterdam.: Elsevier Science.

Stalh, S. M., & Hahn, A. A. (2006). The National Institute on Aging's Resource Centers for Minority Aging Research. *Medical Care, 44*(11 Suppl 3), S1-S2.

Steele, R. G., Little, T. D., Ilardi, S. S., Rex F.R., Brody, G. H., & Hunter, H. L. (2006). A Confirmatory Comparison of the Factor Structure of the Children's Depression Inventory between European American and African American Youth. *J Child Fam Stud, 15*, 779-794.

SteenKamp, J.-B. E. M., & Baumgartner, H. (1998). Assessing Measurement Invariance in Cross-National Consumer Research. *Journal of Consumer Research 25*, 78-90.

Steiger, J. H. (1990). Structural Model Evaluation and Modification: An Interval Estimation Approach. *Multivariate Behavioral Research, 25*(April), 173-180.

Steinberg, L. (2001). The Consequences of Pairing Questions: Context Effects in Personality Measurement. *Journal of Personality and Social Psychology, 81*(2), 332-342.

Stewart, A. L., & Napoles-Springer, A. (2000). Health-related quality-of-life assessments in diverse population groups in the United States. *Medical Care, 38*(9), 102-124.

Swaminathan, H., & Rogers, H. J. (1990). Detecting Differential Item Functioning Using Logistic Regression Procedures. *Journal of Educational Measurement, 27*(4), 361-370.

Tang, W. K., Wong, E., Chiu, H. F. K., Ungvari, G. S., & Lum, C. M. (2005). The Geriatric Depression Scale should be shortened: Results of Rasch analysis. *International Journal of Geriatric Psychiatry, 20*(8), 783-789.

Teresi, J. A. (2006). Overview of Quantitative Measurement Methods: Equivalence, Invariance, and Differential Item Functioning in Health Applications. *Medical Care. Special Issue: Measurement in a multi-ethnic society, 44*(11, Suppl 3), S39-S49.

Teresi, J. A., Ramirez, M., Lai, J.-S., & Silver, S. (2008). Occurences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures: Description of DIF methods, and review of measures of depression, quality of life and general health. *Psycholoy Science Quarterly, 50*(4), 538-612.

Thissen, D. (2001). IRTLRDIF v2.0b Software for the Computation of the Statistics Involved in Item Response Theory Likelihood-Ratio Tests for Differential Item Functioning. Available on Dave Thissen's web page www.unc.edu/~dthissen.

Thissen, D., Steinberg, L., & Kuang, D. (2002). Quick and Easy Implementation of the Benjamini-Hochberg Procedure for Controlling the False Positive Rate in Multiple Comparisons. *Journal of Educational and Behavioral Statistics, 27*(1), 77-83.

Thissen, D., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. W. H. Braun (Ed.), *Test Validity* (pp. 147-169). Hillsdale, NJ: Erlbaum.

Thissen, D., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 67-113). Hillsdale, NJ: Lawrence Erlbaum Associates.

Thissen, D., W-H., & Bock, R. D. (2003). MULTILOG User's Guide. Multiple, Categorical Item Analysis and Test Scoring Using Item Response Theory. (Version 7). Lincolnwood, IL: Scientific Software International.

Thurstone, L. L. (1947). *Multiple-factor analysis: a development and expansion of The Vectors of Mind*. Chicago, IL: University of Chicago Press.

Tomarken, A. J., & Waller, N. G. (2003). Potential problems with "well fitting" models. *Journal of Abnormal Psychology, 112*, 578-598.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices and recommendations for organizational research. *Organizational Research Methods, 3*(1), 4-69.

Vandenberg, R. J., & Lance, C. E. (2000). A Review and Synthesis: of the Measurement Invariance Literature: Suggestions, Practices, and Recommendations for Organizational Research. *Organizational Research Methods, 3*(1), 4-70.

Wrobel, N. H., & Farrag, M. F. (2006). A Preliminary Report on the Validation of the Geriatric Depression Scale in Arabic. *Clinical Gerontologist, 29*(4), 33-45.

Yang, Y., Small, B. J., & Haley, W. E. (2001). Cross-cultural comparability of the Geriatric Depression Scale: comparison between older Koreans and old Americans. *Aging & Mental Health, 5*(1), 31-37.

Yesavage, J. A., Brink, T. L., Rose, T. L., Lum, O., Huang, V., & Adey, M. (1983). Development and validation of a geriatric depression screening scale: a preliminary report. *Journal of Psychiatry Research, 17*(1), 37-49.

Zunzunegui, M., Alvarado, B., Beland, F., & Vissandjee, B. (2009). Explaining health differences between men and women in later life: A cross-city comparison in Latin America and the Caribbean. *Social Science & Medicine, 68*, 235–242.