Technical Note 33

Young Lives
An International Study of Childhood Poverty

The Reliability and Validity of Achievement Tests in the Second Young Lives School Survey in Ethiopia

Juan León and Yessenia Collahua

www.younglives.org.uk

SEPTEMBER 2015

The Reliability and Validity of Achievement Tests in the Second Young Lives School Survey in Ethiopia

Juan León and Yessenia Collahua

First published by Young Lives in September 2015

© Young Lives 2015

Printed on FSC-certified paper from traceable and sustainable sources.

About Young Lives

Young Lives is an international study of childhood poverty, following the lives of 12,000 children in four countries (Ethiopia, India, Peru and Vietnam) over 15 years. **www.younglives.org.uk**

Young Lives is funded from 2001 to 2017 by UK aid from the Department for International Development (DFID), It is co-funded by the Netherlands Ministry of Foreign Affairs from 2010 to 2014, and by Irish Aid from 2014 to 2015.

The views expressed are those of the author(s). They are not necessarily those of, or endorsed by, Young Lives, the University of Oxford, DFID or other funders.

Funded by

UKaid
from the British people

Ministry of Foreign Affairs of the Netherlands

Irish Aid
An Roinn Gnóthaí Eachtracha agus Trádála
Department of Foreign Affairs and Trade

Young Lives, Oxford Department of International Development (ODID), University of Oxford,
Queen Elizabeth House, 3 Mansfield Road, Oxford OX1 3TB, UK

Tel: +44 (0)1865 281751 • E-mail: younglives@younglives.org.uk

Contents

Abstract

This technical note gives details of the reliability and validity of the assessments used in the second school survey carried out by Young Lives in Ethiopia for the purpose of the construction of test scores on a common scale within each language for maths and reading comprehension. This document give details of the three-parameter model used to build the achievement scores in both content areas. We tested graphically for item fit and item bias (by gender and wave). Our results indicate that most of the items used have a good item fit as well as they did not show the presence of bias by wave or gender. Finally, we did an external validity analysis correlating the IRT scores (maths and reading comprehension) with individual and family characteristics, and the results showed that correlations were statistically significant with the expected signs.

The authors

Juan León is a PhD candidate in educational theory and policy and comparative and international education at The Pennsylvania State University. He has a Bachelor's degree in economics from the Pontificia Universidad Catolica of Peru and a diploma in liberal arts from the same institution. Juan is currently an Associate Researcher at GRADE in the education field. Additionally, he is a Lecturer in the Department of Psychology at the Universidad Antonio Ruiz de Montoya in Lima.

Yessenia Collahua has a Bachelor’s degree in social sciences, with economics as her specialist subject, from the Pontificia Universidad Católica of Peru. She is working as a Research Assistant at GRADE.

Acknowledgements

Caine Rolleston and Zoe James gave helpful advice on the analysis of data and provided useful comments to a preliminary version of this technical note. We are grateful to both of them.

1. Introduction

This technical note gives details of the procedures followed to equate the scores in the learning assessments administered at the beginning (Wave 1) and towards the end of the school year (Wave 2) as part of the second Young Lives Ethiopia school survey.[1] It outlines the reliability and validity of the assessments for the purpose of the construction of test scores on a common scale within each of the languages in which the test was administered, for use by the project research team.

Reliability is considered through an analysis of the internal consistency of the assessments (in maths and reading comprehension), essentially the inter-item reliability of test items, using the Cronbach's alpha and the Kuder-Richardson indexes (since a strict examination of test–retest reliability would not be possible).[2] In both cases, values above 0.70 indicate an adequate reliability (Nunnally and Bernstein 1994). Test validity refers to the extent to which the assessments measure what they set out to measure, such that sound conclusions can be drawn from the results. Validity has a variety of different facets, but for the purposes of this paper we focus on a statistical analysis of the construct and predictive validity of the assessments, looking at the correlations with socio-demographic variables.

In addition, we run tests for fairness and bias, checking specifically for differential item functioning (DIF) of test items by wave and gender. We assume that item bias is present when one of the groups (wave or gender) shows a different likelihood of answering an item correctly, even after comparing group members with the same level of ability on the domain in question. DIF analysis therefore offers the opportunity to minimise potential error introduced through working with measures in different waves.[3] We perform item bias analysis by wave and gender using a three-parameter model to arrive at a statistical estimate of DIF. We then compare item response theory (IRT) scores with and without DIF correction.

One advantage of using IRT analysis is to have comparable measures since assessments will be equated across time periods, anchoring the common items that are stable across time (i.e. those repeated items that have an absence of DIF in item parameters) as well as those free of gender bias. This method of equating is possible as a result of the linking design developed for these assessments. This was a vertical equating design that is commonly used to equate scores across grades or time periods. The aim is to keep a core of common items across forms (i.e. in Waves 1 and 2) while the remaining items are replaced with more advanced or difficult items in the second form. For this particular study, two-thirds of the items were kept in both forms while a third of them were replaced in the second form to avoid possible ceiling effects between measures and to adequately capture students' learning over time.

[1] For further details of Young Lives education research, please see www.younglives.org.uk/our-themes/education. Brief details of the second Ethiopia school survey are given in Section 2.1 of this paper.

[2] The test–retest reliability method is a way of estimating test or scale reliability or consistency. In order to estimate this index, we have to administer the same test or alternative test forms to the same individuals at two different times, then we correlate both test scores (Crocker and Algina 1986).

[3] It could introduce measurement error since ítem difficulty could change over time or wave; therefore DIF analysis allows us to ensure that ítem difficulty does not vary statistically over time.

This technical note has five sections. This first section is the introduction; Section 2 describes the data and the methods used for the analysis of the assessments; Section 3 sets out the results for the maths tests; Section 4 gives the results for the reading comprehension tests; and the last section sets out the final considerations about the assessment scores, based on the results of our analysis.

2. Data and methods

2.1. Young Lives Ethiopia school survey

Young Lives is a longitudinal study of childhood poverty in four developing countries (Ethiopia, India, Peru and Vietnam). Through a regular survey of all the children and their caregivers, combined with more in-depth interviews with a sub-sample of the children, we collect a wealth of information about their social and material circumstances, and their hopes and aspirations, set against the environmental and social reality of their communities. A school survey was introduced in 2010 (following the third round of the household survey) to collect data about the character and quality of education received by the Young Lives children.

Young Lives is following the lives of two cohorts of children: a Younger Cohort, consisting of approximately 2,000 children in each country who were born in 2001–02, and an Older Cohort, comprising around 1,000 children in each country born in 1994–95. In Ethiopia, the sample for the second school survey is much larger than these, and comprises approximately 10,000 children from 30 sites, instead of from the 20 sites covered in the main household survey and the first Ethiopia school survey. The first school survey took place in 2009–10, and the second one was carried out in 2012–13, when the children sampled were aged 11–12 and were in Grades 4 and 5.

Ethiopian schools teach predominantly in the child's mother tongue, with most moving towards English-medium education as the child progresses through secondary school. The sampled children were tested in seven different languages, depending on the medium of instruction in their school. For further information about the design of the two Ethiopia school surveys, the make-up of the sample and the way the survey was carried out, see Aurino et al. (2015). Frost and Rolleston (2013) give an overview of the first Ethiopia school survey.

2.2. Test theories

To assess the reliability of the test scores, we use a combination of classical test theory (CTT) and item response theory (IRT), also known as modern test theory. Below are brief descriptions of these two theories.

CTT: This theory is based on three main indicators, which we describe as follows:

- Item difficulty: the percentage of examinees who answer an item correctly. This index ranges from 0 to 1. Values close to 0 mean that few examinees answer the item correctly, and values close to 1 mean that a lot of examinees answer the item correctly. An item is considered extremely difficult if it has an item difficulty range from 0 to 0.20 and extremely easy if it ranges from 0.80 to 1 (Nunnally and Bernstein 1994). Simply put, this is the percentage of respondents who answer an item correctly.

- Item discrimination: this indicates how much an item discriminates between examinees with high and low ability. This index ranges from −1 to 1, where positive values indicate that the item is discriminating in favour of high achievers, and negative values indicate that the item is discriminating against high achievers. Items with a discrimination index lower than 0.10 have to be discarded because they do not help to differentiate between high and low achievers (Nunnally and Bernstein 1994).
- Reliability coefficient: this index measures the overall consistency of a test or scale. There are different types of measures of reliability (e.g. test-retest) but the most commonly used is a measure of the internal consistency, which assesses the multiple item correlation in a test or scale. There are different methods to measure the internal consistency but the most commonly used are Cronbach's alpha and the Kuder-Richardson index (KW20). Both indexes range from 0 to 1, and values closer to 0 indicate low reliability among the items, while values closer to 1 indicate high inter-item reliability. Index values above 0.70 indicate an adequate reliability (Nunnally and Bernstein 1994).

Thus, we use this group of indicators to check the adequacy of the items in each assessment wave and to identify those items that should be deleted according to CTT.

IRT: The IRT models rely on two main assumptions. One is the local independence assumption, which asserts that the probability of an individual answering an item correctly depends on his/her ability only and not by his/her answer to other items. Secondly, the model assumes unidimensionality. In other words, it assumes that only one latent trait is measureable across all items or at least one dominant factor is observed behind the set of items tested. Of these two assumptions, the latter is the most difficult to demonstrate since different factors could be affecting an individual's performance (e.g. test anxiety).[4]

This model assumes that an individual's ability depends on three item parameters – item difficulty, item discrimination, and item guessing. This last parameter refers to the chances that an individual with low levels of ability has to get an item right. This parameter is mainly considered for multiple choice tests since these allow examinees to guess. The equation for this model is:

$$P_i(\theta) = c_i + (1 - c_i)\frac{e^{a_i(\theta - b_i)}}{1 + e^{a_i(\theta - b_i)}} \qquad i = 1,2,\dots,n$$

$P_i(\theta)$: the probability that an individual with ability θ get right the item i

a_i : item discrimination

b_i : the item difficulty

c_i : guessing parameter

n : the number of items in the test

θ : the individual's ability parameter

The two-parameter model uses the same equation but assumes the guessing parameter (c_i) equal to zero while the one-parameter model not only assumes a guessing parameter (c_i) of zero but also the item discrimination (a_i) to be constant across items.

[4] For further information, see Cueto et al. (2009) and Cueto and Leon (2013).

One other aspect that it is important to check when using IRT models is the differential item function (DIF). An item has DIF if the probability of answering it correctly differs across groups or memberships (e.g. gender), controlling by individual's level of ability (Linacre 2008). However, DIF analysis could be sensitive to sample size since the standard errors of the item difficulty depend on the size of the groups that are being compared. Then, large sample sizes could lead to accepting even small differences between item difficulties as DIF. Therefore, it is necessary to use normalised standard errors in order to have better estimates of DIF between groups. For this technical note, we used a graphical approach to check for DIF. We estimated the item characteristic curve (ICC) for each item from the full sample and each group (cohort and wave) that we want to check for DIF and an item was considered to have DIF in each group if the ICC had a different shape than the ICC for the full sample.

Thus, the main differences between these two theories are as follows:

- CTT is sample-dependent while IRT is sample-independent (parameter invariance)
- CTT assumes linearity while IRT assumes non-linearity (logit link)
- CTT assumptions are weak while IRT assumptions are strong to meet with test data
- CTT does not need bigger sample sizes while IRT analysis requires bigger sample sizes (200 or above).

Finally, a main advantage of using IRT is that it helps to build comparable scores using common items. Hambleton (1989) indicates that if we have different tests (common items across them) and the items of those tests meet the IRT assumptions (good item fit indicators), then it is possible to estimate a score for each individual that is independent of the group of items that he/she answered. Thus, it is possible to use those items with adequate fit index as anchors in order to have a score that could be comparable across waves and cohorts.

2.3. Data-cleaning procedures

In both waves survey data underwent double data entry and comparison procedures in Ethiopia. Inconsistencies identified through this process prompted review of the paper questionnaires or tests, and were then corrected in the final dataset. Further cross-tabulations were conducted by the Oxford team, and inconsistencies identified were returned to the Ethiopia team for checking and correction.

Ahead of the analysis that follows, further cleaning and consistency checks were conducted, focusing particularly on the linking of child identification variables across Wave 1 and Wave 2 of the survey. As would be expected, attrition between the two samples is not insignificant (at around 15 per cent), as many Grade 4 and 5 children present on the first day of Wave 1 (who comprised the sample), were not present on the day on which the retest was conducted, owing to a myriad of reasons including holidays and festivals, seasonal migration and so on. Nonetheless attrition is low enough not to interfere with the estimation procedures unduly.

2.4. Sample

The sample for the second Ethiopia school survey covered all children in Grades 4 and 5 in all the schools in the 30 survey sites (the 20 Young Lives study sites, plus 10 further sites in two regions not covered in Young Lives – Somali and Afar). Thus, around a quarter of the Young Lives Younger Cohort children were included, because pupils attending schools outside the sentinel sites were excluded, as were pupils who were not in Grades 4 and 5.

The survey was conducted in two waves: one at the beginning of the school year and the other at the end of it. Students who were absent for the tests administered in Wave 2 were not included in our final sample for both waves. Students were tested in maths and reading comprehension, and, as stated above, they took the tests in the language they were normally taught in at school.

Tables 1 and 2 summarise the number of pupils who took the maths tests by grade and language of test for each wave. Note that children who were surveyed in Grade 3 (because in certain schools transitioning from Alternative Basic Education there was no Grade 4)[5] are included in the Grade 4 sample, and the six children for whom there are no data to show what language they took the test in are omitted.

Table 1. *Child sample in Waves 1 and 2 for maths, by language of test*

	Wave 1			Wave 2		
	Grade 4	Grade 5	Total	Grade 4	Grade 5	Total
Amharic	3,874	3,418	7,292	3,390	2,876	6,266
Hadiyya	143	167	310	117	153	270
Oromo	497	461	958	402	384	786
Sidamo	69	121	190	43	83	126
Somali	406	360	766	325	283	608
Tigrinya	812	980	1,792	723	885	1,608
Wolaytta	188	249	437	170	218	388
Total	5,989	5,756	11,745	5,170	4,882	10,052

Source: Ethiopia second school survey, Waves 1 and 2.

Table 2. *Child sample in Waves 1 and 2 for reading comprehension, by language of test*

	Wave 1			Wave 2		
	Grade 4	Grade 5	Total	Grade 4	Grade 5	Total
Amharic	3,903	3,411	7,314	3,389	2,866	6,255
Hadiyya	151	174	325	117	150	267
Oromo	485	456	941	400	384	784
Sidamo	73	126	199	43	83	126
Somali	411	358	769	325	282	607
Tigrinya	788	930	1,718	723	884	1,607
Wolaytta	198	266	464	170	218	388
Total	6,009	5,721	11,733	5,167	4,867	10,034

Source: Ethiopia second school survey, Waves 1 and 2.

Tables 3 and 4 present the characteristics of the sample, disaggregated by grade and language. They show that the average age of students in Grade 4 is 11 years old, and in Grade 5 it is 12 years old. In both grades, the sample is mostly balanced by gender except for students who took the test in Somali and Wolaytta, where the percentage of female students

[5] Alternative Basic Education (ABE) is a non-formal educational programme in Ethiopia. The main objective of ABE is to provide good education to children in Ethiopia, taking into consideration the local context realities.

was below 50 per cent, and for 5th grade students who took the test in Hadiyya, where the percentage of girls was less than 40 per cent. In terms of educational background, more than 50 per cent of the students who took the test in Amharic and Wolaytta had attended preschool, while students who took the test in Tigrinya had the lowest rates of preschool attendance, with 20 per cent and 21 per cent for 4th and 5th grade students respectively. Almost all the students had started primary school at the age of 7. Only students who took the test in Hadiyya had started primary school at an average age of 6, which could explain why the preschool attendance rate for Hadiyya-speaking children is lower.

Table 3. *Sample characteristics, by grade and language of test in maths*

	Age (years)	Female (%)	Wealth indexa	Age started primary	Attended preschool (%)	Repeated a grade (%)
Grade 4						
Amharic	11.0	53.5	0.19	6.9	62.8	28.6
Hadiyya	10.6	53.5	-1.29	5.9	32.5	16.7
Oromiffa	11.0	54.0	-0.02	7.0	41.2	13.3
Sidamo	11.6	60.0	-0.52	7.2	46.2	20.0
Somali	11.5	45.9	-0.08	7.3	39.1	9.8
Tigrinya	10.3	51.7	-0.50	6.9	19.9	10.4
Wolaytta	11.3	46.9	0.35	6.6	51.5	32.5
Total	11.0	52.6	0.02	6.9	52.3	23.4
Grade 5						
Amharic	12.1	51.9	0.27	6.6	67.9	28.3
Hadiyya	10.8	36.7	-1.11	5.5	35.4	11.6
Oromiffa	12.2	52.5	-0.20	7.2	35.6	13.2
Sidamo	13.0	52.5	-0.93	6.8	31.3	18.8
Somali	12.7	32.6	-0.19	7.5	28.4	5.9
Tigrinya	11.3	51.0	-0.51	7.0	20.6	16.3
Wolaytta	12.2	42.3	0.08	6.5	52.7	39.9
Total	12.0	49.8	-0.01	6.7	52.0	23.5

[a] Factor score built using information about the access of basic services at home, number of durable assets (e.g. fridge), and overcrowding.

Source: Ethiopia second school survey, Waves 1 and 2.

Table 4. *Sample characteristics by grade and language of test in reading comprehension*

	Age (years)	Female (%)	Wealth indexb	Age started primary	Attended preschool (%)	Repeated a grade (%)
Grade 4						
Amharic	11.0	53.4	0.18	6.9	62.9	28.7
Hadiyya	10.5	53.8	-1.30	5.9	32.5	18.8
Oromiffa	11.0	53.7	-0.03	7.0	41.2	13.4
Sidamo	11.5	60.5	-0.55	7.1	47.6	18.6
Somali	11.6	45.4	-0.09	7.3	39.0	9.7
Tigrinya	10.3	51.4	-0.50	6.9	20.3	10.5
Wolaytta	11.4	47.0	0.34	6.6	51.5	32.5
Total	11.0	52.5	0.02	6.9	52.6	23.6
Grade 5						
Amharic	12.1	51.8	0.27	6.6	67.8	28.4
Hadiyya	10.8	38.7	-1.11	5.5	35.3	12.0
Oromiffa	12.1	52.5	-0.20	7.2	35.3	13.3
Sidamo	13.0	53.0	-0.94	6.9	33.7	20.5
Somali	12.7	33.6	-0.18	7.5	28.6	5.9
Tigrinya	11.3	50.8	-0.49	7.0	20.6	16.2
Wolaytta	12.2	42.2	0.06	6.6	52.5	39.4
Total	12.0	49.8	-0.01	6.7	52.2	23.6

Source: Ethiopia second school survey, Waves 1 and 2.

3. Results for math tests

3.1. Raw scores

This section presents the item analysis performed on the maths tests in order to build the maths raw scores. We used CTT to analyse the item fit in terms of reliability. Tables 5 and 6 show the CTT reliability analysis performed, by language, in Waves 1 and 2 respectively. The Cronbach and KW20 indexes show that tests had adequate test reliability for all languages, with indexes around 0.69 and 0.80. We observe that tests administered in Oromo, Somali and Wolaytta had the highest reliability indexes in both waves, while the lowest reliability indexes in both waves are observed for the test administered in Sidamo.

Table 5. *Reliability indexes for the whole sample, by language, in maths for Wave 1*

Language	Initial		Final		Items deleted
	Cronbach	KW20	Cronbach	KW20	
Amharic	0.71	0.72	0.75	0.75	16, 21 & 24
Hadiyya	0.63	0.63	0.76	0.76	14, 18, 19, 20, 21, 22 & 23
Oromo	0.75	0.75	0.80	0.80	18, 19, 21, & 25
Sidamo	0.59	0.59	0.69	0.69	14,15,18, 19, , 21, 23, 24 & 25
Somali	0.76	0.76	0.78	0.78	20 & 25
Tigrinya	0.69	0.69	0.75	0.75	18,19, 20, 21 & 22
Wolaytta	0.79	0.79	0.81	0.81	14 & 19

Source: Ethiopia second school survey, Wave 1.

Table 6. *Reliability indexes for the whole sample, by language, in maths for Wave 2*

Language	Initial		Final		Items deleted
	Cronbach	KW20	Cronbach	KW20	
Amharic	0.72	0.73	0.80	0.80	25
Hadiyya	0.63	0.63	0.79	0.79	14, 18, 20, 22 & 23
Oromo	0.75	0.75	0.82	0.82	21 & 25
Sidamo	0.59	0.59	0.68	0.68	14,18, 21 & 25
Somali	0.74	0.74	0.72	0.72	16, 21& 25
Tigrinya	0.69	0.69	0.76	0.76	20, 21 & 22
Wolaytta	0.79	0.79	0.82	0.82	none

Source: Ethiopia second school survey, Wave 2.

Once the reliability index was established and items with poor fit were deleted from each scale for each language, we calculated the raw corrected scores. Table 7 presents mean raw corrected maths scores for Wave 1 by language, grade and gender. We observe that on average boys have higher scores than girls, with these differences not being statistically significant in most languages although the scores of Hadiyya-speaking boys are significantly higher those of girls who took the test in Hadiyya in both grades. Wolaytta-speaking students show the opposite pattern, since girls have higher scores than boys and this difference is statistically significant in both grades.

Table 7. *Raw mean scores in maths test, Wave 1, by language, gender and grade*

Language *(max. score)*	Grade 4			Grade 5			Total		
	Boys	Girls	Difference (G – B)	Boys	Girls	Difference (G – B)	Boys	Girls	Difference (G – B)
Amharic	12.8	12.7	-0.1	14.7	14.3	-0.4*	13.7	13.4	-0.3*
(max. 22)	(3.6)	(3.5)		(3.4)	(3.7)		(3.7)	(3.7)	
Hadiyya	8.2	7.1	-1.1*	8.5	7.2	-1.3*	8.4	7.1	-1.3*
(max. 15)	(3.0)	(2.7)		(3.0)	(2.8)		(3.0)	(2.7)	
Oromo	10.0	9.5	-0.5	11.9	11.1	-0.8*	10.9	10.3	-0.6*
(max. 17)	(3.5)	(3.6)		(3.7)	(3.8)		(3.7)	(3.7)	
Sidamo	10.4	10.1	-0.3	10.5	11.0	0.5	10.4	10.6	0.2
(max. 17)	(2.8)	(2.1)		(2.9)	(2.8)		(2.9)	(2.5)	
Somali	10.6	10.4	-0.2	12.7	12.4	-0.3	11.5	10.9	-0.6*
(max. 20)	(4.0)	(3.8)		(3.7)	(4.2)		(4.1)	(4.2)	
Tigrinya	8.5	8.4	-0.1	10.1	9.9	-0.2	9.4	9.2	-0.2
(max. 16)	(3.2)	(3.2)		(3.2)	(3.2)		(3.3)	(3.3)	
Wolaytta	9.6	11.1	1.5*	10.5	11.8	1.3*	10.1	11.5	1.4*
(max. 19)	(3.6)	(3.9)		(4.2)	(4.3)		(4.0)	(4.2)	

* Differences are statistically significant at 5% according the *t*-test for independent samples.
Note: Standard deviation in parentheses
Source: Ethiopia second school survey, Wave 1

A similar pattern is observed in Table 8 for Wave 2. For most of the languages, boys have higher maths scores than girls but these differences are not statistically significant in most cases. However, as in Wave 1, for the Wolaytta-speaking students, the gender differences favour girls, although this difference is statistically significant only in Grade 5.

Table 8. *Raw mean scores in maths test, Wave 2, by language, gender and grade*

Language *(max. score)*	**Grade 4**			**Grade 5**			**Total**		
	Boys	**Girls**	**Difference (G – B)**	**Boys**	**Girls**	**Difference (G – B)**	**Boys**	**Girls**	**Difference (G – B)**
Amharic *(max. 24)*	14.7 (4.5)	14.6 (4.2)	-0.1	16.3 (4.1)	15.9 (4.3)	-0.4*	15.5 (4.4)	15.2 (4.3)	-0.3*
Hadiyya *(max. 17)*	9.3 (3.6)	8.8 (3.4)	-0.5	10.3 (4.1)	9.8 (4.1)	-0.5	10.0 (3.9)	9.3 (3.7)	-0.7
Oromo *(max. 19)*	11.6 (4.5)	11.5 (4.2)	-0.1	13.2 (4.3)	11.9 (3.9)	-1.3*	12.4 (4.5)	11.7 (4.1)	-0.7*
Sidamo *(max. 19)*	11.1 (3.3)	10.9 (2.5)	-0.2	13.2 (3.4)	12.1 (3.6)	-1.1	12.6 (3.4)	11.7 (3.6)	-0.9
Somali *(max. 20)*	10.6 (3.7)	10.1 (3.7)	-0.5	12.5 (3.4)	12.4 (3.5)	-0.1	11.5 (3.7)	10.8 (3.9)	-0.7*
Tigrinya *(max. 18)*	9.3 (4.0)	9.1 (3.8)	-0.2	10.9 (3.6)	10.4 (3.6)	-0.5*	10.2 (3.9)	9.8 (3.8)	-0.4*
Wolaytta *(max. 21)*	10.7 (4.7)	11.2 (4.5)	0.5	11.1 (4.6)	12.6 (4.6)	1.5*	10.9 (4.6)	12.0 (4.6)	1.1*

* Differences are statistically significant at 5% according the *t*-test for independent samples.
Note: Standard deviation in parentheses
Source: Ethiopia second school survey, Wave 2.

After that, we calculated the mean raw score for the common items without item fit problems for Waves 1 and 2. Table 9 presents the maths scores gains (difference between Waves 1 and 2) for all languages in Grades 4 and 5. These range from 0.1 to 1.4 points and almost all of them are statistically significant for both grades. However, for Sidamo in Grade 4 and Wolaytta in Grades 4 and 5 the gain scores are not statistically significant.

Table 9. *Raw mean scores in maths test, Waves 1 and 2, by language and grade (common items)*

Language *(max. score)*	**Grade 4**			**Grade 5**		
	Wave 1	**Wave 2**	**Difference (W2 – W1)**	**Wave 1**	**Wave 2**	**Difference (W2 – W1)**
Amharic *(max.16)*	9.4 (2.9)	10.6 (3.2)	1.2*	10.8 (2.9)	11.4 (2.8)	0.6*
Hadiyya *(max. 12)*	5.9 (2.7)	7.1 (2.4)	1.2*	6.2 (2.8)	7.6 (3.1)	1.4*
Oromo *(max. 14)*	7.4 (3.0)	8.7 (3.4)	1.3*	9.0 (3.3)	9.7 (3.1)	0.7*
Sidamo *(max. 12)*	7.6 (2.1)	7.8 (2.6)	0.2	8.1 (2.7)	8.8 (2.3)	0.7*
Somali *(max. 14)*	7.3 (2.9)	7.8 (2.9)	0.5*	8.7 (2.6)	9.4 (2.4)	0.7*
Tigrinya *(max. 13)*	6.7 (2.7)	7.3 (3.2)	0.6*	8.1 (2.8)	8.7 (2.9)	0.6*
Wolaytta *(max. 15)*	8.0 (3.3)	8.5 (3.6)	0.5	8.9 (3.6)	9.0 (3.6)	0.1

* Differences are statistically significant at 5% according to the *t-test* for independent samples.
Note: Standard deviation in parentheses
Source: Ethiopia second school survey, Waves 1 and 2.

Finally, we checked whether the raw scores followed a normal distribution pattern. Table 10 presents the skewness and kurtosis statistics for the corrected raw scores in Waves 1 and 2. As we could observe, no score follows a normal distribution pattern since we reject the null hypothesis of normality according to the test performed and most of the scores are negatively skewed since the skewness statistic is negative in most cases (see Appendix A for distribution graphs).

Table 10. *Characteristics of the distribution of raw scores for maths by language*

Language	Wave 1		Wave 2	
	Skewness	Kurtosis	Skewness	Kurtosis
Amharic	-0.33*	2.82*	-0.33*	2.64
Hadiyya	-0.03	2.04*	-0.00	2.17*
Oromo	-0.30*	2.15*	-0.21*	2.04*
Sidamo	-0.89*	4.15*	-0.40*	2.73
Somali	-0.25	2.41*	-0.20*	2.29
Tigrinya	-0.22*	2.30*	-0.06*	2.18*
Welayitta	-0.18	2.07*	0.03	2.08*

* Statistically significant at 5% according the Jarque-Bera test for normality.
Source: Ethiopia second school survey, Waves 1 and 2.

3.2. IRT scores

As we mentioned above, we used IRT techniques to produce equated scores across waves for each language, which would not have been possible using techniques such as CTT. We estimated a three-parameter IRT model, which assumes that children's ability depends on item difficulty, item discrimination and item guessing. First, we ran the three-parameter model for the pooled sample (Waves 1 and 2); second, we examined and corrected for possible item misfit and DIF by gender and wave; and finally we ran the final three-parameter model calculation to equate the students' scores between Waves 1 and 2 for each language.[6]

Table 11 shows the IRT scores in maths by language, grade and gender for Wave 1. We were able to observe score differences between girls and boys for each language by grade. In Grade 4, we observe that for most of the languages boys outperformed girls; however this difference is only statistically significant for Hadiyya-speaking students. In Grade 5, we observe a similar pattern, with boys outperforming girls in maths in four of the seven languages; this difference is statistically significant for Amharic-, Hadiyya- and Oromo-speaking students.

[6] See Appendices for details of item fit and item bias.

Table 11. *IRT scores in maths test, Wave 1, by language, gender and grade*

Language *(max. score)*	Grade 4			Grade 5		
	Boys	Girls	Difference (G – B)	Boys	Girls	Difference (G – B)
Amharic *(max. 22)*	-0.26 (0.84)	-0.29 (0.80)	-0.03	0.18 (0.84)	0.08 (0.90)	-0.10*
Hadiyya *(max. 16)*	0.03 (0.85)	-0.37 (0.77)	-0.40*	-0.01 (0.83)	-0.40 (0.83)	-0.39*
Oromo *(max. 19)*	-0.31 (0.82)	-0.37 (0.77)	-0.06	0.26 (0.96)	0.01 (0.86)	-0.25*
Sidamo *(max. 19)*	0.05 (0.81)	-0.35 (0.66)	-0.40	-0.03 (1.01)	0.06 (0.89)	0.09
Somali *(max. 18)*	-0.07 (0.87)	-0.25 (0.81)	-0.18	0.29 (0.75)	0.32 (0.90)	0.03
Tigrinya *(max. 18)*	-0.30 (0.82)	-0.30 (0.80)	0.00	0.14 (0.83)	0.09 (0.82)	-0.05
Wolaytta *(max. 21)*	-0.24 (0.83)	0.01 (0.86)	0.25	-0.07 (0.92)	0.25 (0.96)	0.32*

* Differences are statistically significant at 5% according the *t*-test for independent samples.
Note: Standard deviation in parentheses.
Source: Ethiopia second school survey, Wave 1.

Table 12 shows the IRT results for maths for Wave 2. As in Wave 1, we calculated the average IRT scores by gender and grade. In Grade 4, we observed that for most languages boys outperformed girls but that these differences were not statistically significant. A similar pattern was observed in Grade 5, with boys outperforming girls, and these differences were statistically significant for Amharic- and Oromo-speaking students.

Table 12. *IRT scores in maths test, Wave 2, by language, gender and grade*

Language *(max. score)*	Grade 4			Grade 5		
	Boys	Girls	Difference (G – B)	Boys	Girls	Difference (G – B)
Amharic *(max. 24)*	0.01 (0.95)	-0.01 (0.88)	-0.02	0.34 (0.86)	0.26 (0.91)	-0.08*
Hadiyya *(max. 19)*	0.17 (0.85)	0.08 (0.77)	-0.09	0.37 (0.98)	0.18 (0.99)	-0.19
Oromo *(max. 23)*	-0.04 (0.99)	-0.01 (0.93)	-0.03	0.41 (0.98)	0.14 (0.86)	-0.27*
Sidamo *(max. 20)*	-0.05 (0.79)	-0.02 (0.77)	0.03	0.31 (0.87)	0.19 (0.84)	-0.12
Somali *(max. 21)*	0.02 (0.83)	-0.05 (0.86)	-0.07	0.46 (0.72)	0.52 (0.74)	0.06
Tigrinya *(max. 21)*	-0.14 (0.94)	-0.18 (0.88)	-0.04	0.34 (0.92)	0.25 (0.91)	-0.09
Wolaytta *(max. 25)*	-0.07 (0.97)	0.05 (0.91)	0.12	0.01 (0.96)	0.30 (0.93)	0.29*

* Differences are statistically significant at 5% according the *t*-test for independent samples.
Note: Standard deviation in parentheses.
Source: Ethiopia second school survey, Wave 2.

Table 13 presents children's progress in maths between Waves 1 and 2 by language and grade. The learning progress is statistically significant for most of the languages: this increment ranges from 0.07 to 0.45. In Grade 4, we observe that Oromo-speaking students have the highest increment in maths scores over time, at 0.36 SD (p<0.05), while the lowest increment was for Somali- and Tigrinya-speaking students, at 0.14 SD (p<0.05). In Grade 5, for most languages the learning increment was statistically significant, with Hadiyya-speaking students being those who showed the highest increment, at 0.45 SD (p<0.05), and Oromo-speaking students having the lowest one, at 0.14 SD (p<0.05).

Table 13. *IRT scores in maths test, Waves 1 and 2, by language and grade*

Language *(max. score)*	**Grade 4**			**Grade 5**		
	Wave 1	**Wave 2**	**Difference (W2 – W1)**	**Wave 1**	**Wave 2**	**Difference (W2 – W1)**
Amharic	-0.27	-0.00	0.27*	0.13	0.30	0.17*
(max. 17)	(0.82)	(0.92)		(0.87)	(0.89)	
Hadiyya	-0.18	0.13	0.31*	-0.15	0.30	0.45*
(max. 13)	(0.83)	(0.81)		(0.85)	(0.99)	
Oromo	-0.34	0.02	0.36*	0.13	0.27	0.14*
(max. 16)	(0.79)	(0.96)		(0.91)	(0.93)	
Sidamo	-0.19	-0.04	0.15	0.01	0.25	0.24
(max. 14)	(0.75)	(0.77)		(0.95)	(0.86)	
Somali	-0.15	-0.01	0.14*	0.31	0.48	0.17*
(max. 13)	(0.85)	(0.84)		(0.80)	(0.72)	
Tigrinya	-0.30	-0.16	0.14*	0.11	0.29	0.18*
(max. 14)	(0.81)	(0.91)		(0.82)	(0.92)	
Wolaytta	-0.13	-0.01	0.12	0.06	0.13	0.07
(max. 17)	(0.85)	(0.94)		(0.95)	(0.96)	

* Differences are statistically significant at 5% according the *t*-test for independent samples.
Note: Standard deviation in parentheses.
Source: Ethiopia second school survey, Waves 1 and 2.

As with the raw scores, we checked whether the IRT scores followed a normal distribution pattern. Table 14 presents the skewness and kurtosis statistics for the IRT equated scores. We could observe that we rejected the null hypothesis of normality according to the test performed;[7] however, in comparison with the raw scores, most of the IRT scores are not negatively skewed, showing a symmetrical distribution (see Appendix E for distribution graphs).

[7] The Jarque-Bera test is used to check if a variable (test scores) is normally distributed. To check for normal distribution two statistics are generally used: skewness (lack of symmetry) and kurtosis (tallness or flatness). A variable is considered normally distributed if skewness is equal to zero and kurtosis equal to three. Thus, the Jarque-Bera test checks statistically if the skewness is equal to or different from zero and the kurtosis equal to or different from three.

Table 14. *Characteristics of the distribution of IRT scores for maths, by language*

Language	Original		Corrected	
	Skewness	Kurtosis	Skewness	Kurtosis
Amharic	-0.15*	3.02	-0.15*	3.01
Hadiyya	0.09	2.42*	0.07	2.36*
Oromo	0.07	2.48*	0.05	2.43*
Sidamo	-0.30*	3.02	-0.13	2.45*
Somali	-0.10	2.32*	-0.13*	2.29*
Tigrinya	-0.01	2.59*	-0.04	2.47*
Wolaytta	0.09	2.30*	0.02	2.14*

* Statistically significant at 5 % according the Jarque-Bera test for normality.
Source: Ethiopia school survey, Waves 1 and 2.

Finally, we correlated the IRT scores (Waves 1 and 2) with children's background characteristics. Tables 15 and 16 show the correlation between IRT scores in the two waves and children's gender, age, grade, their family's wealth index, the age they started primary school, whether they attended preschool and whether they repeated a grade. As we expected, the maths IRT scores are positively correlated with grade, age and wealth index for most languages; and negatively correlated with student retention for all languages; and with respect to the other variables (gender, age the child started primary school, and preschool attendance), the correlations vary by language.

Table 15. *Correlation between maths IRT scores, Wave 1, and children's demographic and educational background variables*

Variable	Amharic	Hadiyya	Oromo	Sidamo	Somali	Tigrinya	Wolaytta
Female	-0.04*	-0.23*	-0.09*	-0.04	-0.11*	-0.02	0.15*
Age	0.07*	0.09	0.26*	-0.04	0.07	0.15*	-0.01
Grade	0.24*	0.01	0.26*	0.11	0.37*	0.24*	0.10*
Wealth index	0.15*	-0.15*	0.08*	-0.20*	0.26*	0.22*	0.15*
Age started primary school	-0.02	0.14*	0.28*	0.05	-0.16*	0.13*	0.09
Attended preschool	0.14*	-0.13*	0.05	0.03	-0.01	0.14*	0.14*
Repeated a grade	-0.22*	-0.14*	-0.22*	-0.53*	-0.20*	-0.21*	-0.14*

* Statistically significant at 5%.
Source: Ethiopia second school survey, Waves 1 and 2.

Table 16. *Correlation between IRT maths scores, Wave 2, and children's demographic and educational background variables*

Variable	Amharic	Hadiyya	Oromo	Sidamo	Somali	Tigrinya	Wolaytta
Female	-0.03*	-0.09*	-0.08*	-0.05	-0.07	-0.04	0.11*
Age	0.00	0.08	0.15*	0.02	0.10*	0.16*	0.04
Grade	0.17*	0.09	0.13*	0.15	0.38*	0.24*	0.08
Wealth index	0.16*	-0.20*	0.12*	-0.21*	0.28*	0.22*	0.15*
Age started primary school	-0.07*	0.09	0.20*	0.09	-0.11*	0.14*	0.16*
Attended preschool	0.16*	-0.14*	0.12*	-0.05	-0.03	0.13*	0.01
Repeated a grade	-0.23*	-0.03	-0.20*	-0.38*	-0.20*	-0.21*	-0.10

* Statistically significant at 5 %.
Source: Ethiopia second school survey, Waves 1 and 2.

4. Results for reading comprehension tests

4.1. Raw scores

We performed the same range of tests on the reading comprehension assessment scores. Tables 17 and 18 show the CTT reliability analysis performed, by language of administration, in Waves 1 and 2 respectively. The Cronbach and KW20 indexes show that tests had adequate test reliability for all the languages, with indexes around 0.66 and 0.88. Also, we could observe that tests administered in Amharic, Oromo, Somali and Tigrinya have the highest reliability indexes in both waves, while the lowest reliability is evident for tests administered in Sidamo in both waves.

Table 17. *Reliability indexes for the whole sample, by language, in reading comprehension for Wave 1*

Language	Initial		Final		Items deleted
	Cronbach	KW20	Cronbach	KW20	
Amharic	0.86	0.86	0.86	0.86	none
Hadiyya	0.77	0.77	0.80	0.80	15 & 24
Oromo	0.85	0.85	0.86	0.86	24
Sidamo	0.66	0.66	0.76	0.76	2, 20, 21, 24 & 25
Somali	0.88	0.88	0.88	0.88	none
Tigrinya	0.84	0.84	0.88	0.88	21, 22 & 24
Wolaytta	0.79	0.79	0.81	0.81	25

Source: Ethiopia second school survey, Wave 1.

Table 18. *Reliability indexes for the whole sample by language in reading comprehension for Wave 2*

Language	Initial		Final		Items deleted
	Cronbach	KW20	Cronbach	KW20	
Amharic	0.79	0.79	0.83	0.83	22, 24, 25
Hadiyya	0.74	0.74	0.78	0.78	18, 22 & 25
Oromo	0.82	0.82	0.84	0.84	18, & 24
Sidamo	0.63	0.63	0.74	0.74	14, 15, 18, 23 & 25
Somali	0.85	0.85	0.88	0.88	22, 23 & 25
Tigrinya	0.73	0.73	0.83	0.83	15, 16, 18, 22, 23, 24 & 25
Wolaytta	0.75	0.75	0.81	0.81	20, 22, 23 & 25

Source: Ethiopia second school survey, Wave 2.

Table 19 presents mean raw scores in reading comprehension for Wave 1 by language, grade and gender. In Grade 4, in most languages, boys have higher scores than girls but this difference is statistically significant only for Amharic-speaking students. However, in Grade 5, we have the opposite pattern since for three of the seven languages, girls have higher scores than boys although this difference is statistically significant only for Wolaytta-speaking students.

Table 19. *Raw mean scores in reading comprehension test, Wave 1, by language, gender and grade*

Language *(max. score)*	Grade 4			Grade 5			Total		
	Boys	Girls	Difference (G – B)	Boys	Girls	Difference (G – B)	Boys	Girls	Difference (G – B)
Amharic	17.39	17.80	0.40*	18.70	18.76	0.06	18.0	18.2	0.2*
(max. 23)	4.30	4.19		3.57	3.84		(4.0)	(4.1)	
Hadiyya	12.47	10.32	-2.14*	14.52	13.45	-1.06	13.7	11.7	-2.0*
(max. 23)	4.42	3.49		4.22	4.44		(4.4)	(4.2)	
Oromo	15.27	15.41	0.14	17.57	17.17	-0.40	16.4	16.2	-0.2
(max. 24)	4.95	4.82		5.00	4.92		(5.1)	(4.9)	
Sidamo	13.03	12.18	-0.85	13.45	13.66	0.21	13.3	13.1	-0.2
(max. 19)	3.58	3.40		2.68	3.20		(3.0)	(3.3)	
Somali	18.38	18.03	-0.34	20.39	20.65	0.26	19.0	18.6	-0.4
(max. 25)	5.12	5.01		4.27	4.80		(5.1)	(5.4)	
Tigrinya	14.88	14.56	-0.32	16.85	16.62	-0.23	15.9	15.7	-0.2
(max. 21)	4.84	5.11		4.06	4.26		(4.5)	(4.8)	
Wolaytta	14.16	13.76	-0.40	14.40	15.74	1.34*	14.3	14.8	0.5
(max. 21)	3.77	3.96		4.08	3.97		(3.9)	(4.1)	

* Differences are statistically significant at 5% according the *t*-test for independent samples.
Note: Standard deviation in parentheses.
Source: Ethiopia second school survey, Wave 1.

Table 20 presents the mean raw scores for Wave 2. In both grades, we observe for most of the languages that girls have higher scores than boys; however, only a few of these differences are statistically significant. Only for Amharic-speaking students in Grade 4 and Somali-speaking students in Grade 5 are these differences statistically significant.

Table 20. *Raw mean scores in reading comprehension test, Wave 2, by language, gender and grade*

Language *(max. score)*	Grade 4			Grade 5			Total		
	Boys	Girls	Difference (G – B)	Boys	Girls	Difference (G – B)	Boys	Girls	Difference (G – B)
Amharic	14.19	14.69	0.49*	15.17	15.25	0.08	14.6	14.9	0.3*
(max. 19)	(3.66)	(3.35)		(2.92)	(3.16)		(3.4)	(3.3)	
Hadiyya	11.91	10.63	-1.27	12.55	11.78	-0.78	12.3	11.2	-1.1*
(max. 21)	(4.42)	(3.33)		(4.03)	(4.53)		(4.2)	(3.9)	
Oromo	12.38	13.13	0.75	14.53	14.09	-0.44	13.5	13.6	0.1
(max. 20)	(4.30)	(4.44)		(4.25)	(4.32)		(4.4)	(4.4)	
Sidamo	13.18	12.08	-1.10	13.08	12.45	-0.62	13.1	12.3	-0.8
(max. 19)	(2.53)	(2.95)		(3.08)	(3.54)		(2.9)	(3.3)	
Somali	16.13	15.30	-0.83	17.40	18.36	0.95*	16.5	15.9	-0.6
(max. 21)	(4.52)	(4.61)		(3.45)	(2.64)		(4.3)	(4.7)	
Tigrinya	12.41	12.59	0.18	13.99	14.19	0.20	13.3	13.5	0.2
(max. 18)	(3.86)	(4.03)		(3.50)	(3.49)		(3.7)	(3.8)	
Wolaytta	9.71	10.10	0.39	10.48	11.22	0.75	10.2	10.7	0.5
(max. 20)	(3.77)	(3.97)		(3.70)	(3.66)		(3.7)	(3.8)	

* Differences are statistically significant at 5% according the *t*-test for independent samples.
Note: Standard deviation in parentheses.
Source: Ethiopia second school survey, Wave 1.

We then calculated the mean raw score for the common items without item fit problems for Waves 1 and 2. Table 21 shows the increment over time in reading comprehension scores for all languages in Grades 4 and 5. These increments range from 0.1 to 1.5 points and more than half of them are statistically significant for both grades. Only for Sidamo, Somali and Wolaytta in Grade 4 and Hadiyya, Sidamo and Wolaytta in Grade 5 is the learning gain not statistically significant.

Table 21. *Raw mean scores in reading comprehension test, Waves 1 and 2, by language and grade (common items)*

Language *(max. score)*	**Grade 4**			**Grade 5**		
	Wave 1	**Wave 2**	**Difference (W2 – W1)**	**Wave 1**	**Wave 2**	**Difference (W2 – W1)**
Amharic	7.9	8.6	0.7*	8.4	8.9	0.5*
(max. 9)	(2.2)	(2.0)		(1.9)	(1.7)	
Hadiyya	5.8	7.3	1.5*	7.3	7.8	0.5
(max. 11)	(2.6)	(2.1)		(2.4)	(2.5)	
Oromo	7.7	8.9	1.1*	9.0	9.6	0.6*
(max. 12)	(2.9)	(2.8)		(2.9)	(2.5)	
Sidamo	7.2	7.9	0.7	7.5	7.9	0.4
(max. 10)	(2.0)	(1.7)		(2.0)	(1.7)	
Somali	10.5	10.8	0.5	11.5	11.8	0.3*
(max. 13)	(2.8)	(2.8)		(2.2)	(1.7)	
Tigrinya	6.8	7.7	0.9*	7.8	8.4	0.6*
(max. 10)	(2.6)	(2.3)		(2.1)	(1.7)	
Wolaytta	6.7	6.8	0.1	7.4	7.5	0.1
(max. 10)	(2.3)	(2.6)		(2.2)	(2.3)	

* Differences are statistically significant at 5% according to the *t*-test for independent samples.
Note: Standard deviation in parentheses.
Source: Ethiopia second school survey, Waves 1 and 2.

Then, we checked if the raw scores followed a normal distribution pattern. Table 22 shows the skewness and kurtosis statistics for the corrected raw scores in Waves 1 and 2. We could observe that no score follows a normal distribution pattern since we reject the null hypothesis for normality according to the test performed, and most of scores are negatively skewed (see Appendix A for distribution graphs).

Table 22. *Characteristics of the distributions of raw scores, by language*

Language	**Wave 1**		**Wave 2**	
	Skewness	**Kurtosis**	**Skewness**	**Kurtosis**
Amharic	-1.41*	4.44*	-1.49*	4.74*
Hadiyya	0.07	1.93*	0.16	2.18*
Oromo	-0.39*	2.15*	-0.39*	2.07*
Sidamo	-0.48*	2.75	-0.32*	2.55
Somali	-0.97*	3.06*	-1.15*	3.55*
Tigrigna	-0.95*	2.94*	-0.91*	3.16*
Wolaytta	-0.33*	2.15*	-0.24*	2.07*

* Statistically significant at 5% according to the Jarque-Bera test for normality.
Source: Ethiopia second school survey, Waves 1 and 2.

4.2. IRT scores

To estimate the IRT scores, we followed the same procedures as for the maths tests.[8] Table 23 shows the IRT scores in reading comprehension by language, grade and gender for Wave 1. In Grade 4, we observe that gender differences, favouring boys, are statistically significant for Hadiyya- and Sidamo-speaking students, while for Amharic-speaking students the difference is significant favouring girls. In Grade 5, we observe statistically significant gender differences for Amharic-speaking students favouring girls and for Haddiya-speaking students favouring boys.

Table 23. *IRT scores in reading comprehension test Wave 1 by language, gender and grade*

Language	Grade 4			Grade 5		
	Boys	Girls	Difference (G – B)	Boys	Girls	Difference (G – B)
Amharic	-0.27 (0.88)	-0.17 (0.89)	0.10*	-0.01 (0.82)	0.08 (0.86)	0.09*
Hadiyya	-0.12 (1.05)	-0.65 (0.74)	-0.53*	0.21 (0.89)	-0.09 (0.94)	-0.30*
Oromo	-0.33 (0.93)	-0.27 (0.94)	0.06	0.15 (0.99)	0.03 (0.97)	-0.12
Sidamo	0.29 (1.05)	-0.42 (0.94)	-0.71*	0.03 (0.91)	0.04 (0.91)	0.01
Somali	-0.04 (0.89)	-0.17 (0.87)	-0.13	0.26 (0.76)	0.34 (0.81)	0.08
Tigrinya	-0.30 (0.96)	-0.36 (0.98)	-0.06	0.10 (0.88)	0.05 (0.90)	0.05
Wolaytta	-0.06 (0.86)	-0.18 (0.90)	-0.12	-0.03 (0.95)	0.30 (0.89)	0.33*

* Differences are statistically significant at 5% according the *t*-test for independent samples.
Note: Standard deviation in parentheses.
Source: Ethiopia second school survey, Wave 1.

Table 24 presents the results for Wave 2. We observe a similar pattern as in Wave 1. In Grade 4, gender differences were statistically significant in the Amharic and Hadiyya assessments, with boys outperforming girls in the former and vice versa in the latter; while in Grade 5, the gender differences were significant in Wolaytta, favouring girls.

[8] See Appendices for details of item fit and item bias.

Table 24. *IRT scores in reading comprehension test, Wave 2, by language, gender and grade*

	Grade 4			Grade 5		
	Boys	Girls	Difference (G – B)	Boys	Girls	Difference (G – B)
Amharic	0.01 (0.88)	0.13 (0.87)	0.12*	0.25 (0.78)	0.29 (0.83)	0.04
Hadiyya	0.27 (0.91)	0.05 (0.67)	-0.32*	0.41 (0.86)	0.20 (0.98)	-0.21
Oromo	-0.03 (0.88)	0.12 (0.88)	0.15	0.36 (0.93)	0.33 (0.92)	-0.03
Sidamo	0.31 (0.72)	-0.02 (0.83)	-0.33	0.24 (0.93)	0.10 (0.93)	-0.14
Somali	0.13 (0.92)	-0.03 (0.97)	-0.16	0.42 (0.77)	0.60 (0.68)	0.18
Tigrinya	-0.08 (0.85)	-0.02 (0.93)	-0.06	0.31 (0.82)	0.37 (0.83)	0.06
Wolaytta	-0.08 (0.96)	-0.04 (1.03)	0.04	0.07 (0.93)	0.34 (0.96)	0.27*

* Differences are statistically significant at 5% according to the *t*-test for independent samples.
Note: Standard deviation in parentheses.
Source: Ethiopia second school survey, Wave 2.

Table 25 presents reading comprehension learning progress by language and grade. In Grade 4, for most of the languages, we observe a statistically significant increment in IRT scores over time. The learning gain was highest for Hadiyya-speaking students, at 0.50 SD (p<0.05) and lowest for Somali students, at 0.16 SD (p<0.05). In Grade 5, the biggest learning gain was for Tigrinya-speaking students, at 0.27 SD (p<0.05) and the lowest for Somali-speaking students, at 0.20 SD (p<0.05).

Table 25. *IRT scores in reading comprehension test, Waves 1 and 2, by language and grade*

Language	Grade 4			Grade 5		
	Wave 1	Wave 2	Difference (W2 – W1)	Wave 1	Wave 2	Difference (W2 – W1)
Amharic	-0.22 (0.89)	-0.07 (0.87)	0.29*	0.04 (0.84)	0.27 (0.80)	0.23*
Hadiyya	-0.41 (0.93)	0.09 (0.80)	0.50*	0.09 (0.92)	0.33 (0.91)	0.24*
Oromo	-0.31 (0.94)	0.05 (0.88)	0.36*	0.09 (0.98)	0.35 (0.92)	0.26*
Sidamo	-0.15 (1.04)	0.11 (0.79)	0.25	0.04 (0.90)	0.17 (0.93)	0.13
Somali	-0.10 (0.89)	0.06 (0.95)	0.16*	0.28 (0.78)	0.48 (0.75)	0.20*
Tigrinya	-0.33 (0.97)	-0.05 (0.89)	0.28*	0.07 (0.89)	0.34 (0.83)	0.27*
Wolaytta	-0.12 (0.88)	-0.06 (0.99)	0.06	0.11 (0.93)	0.18 (0.85)	0.07

* Differences are statistically significant at 5% according to the *t*-test for independent samples.
Note: Standard deviation in parentheses
Source: Ethiopia second school survey, Waves 1 and 2.

As with the raw scores, we checked to see whether the IRT scores followed a normal distribution pattern. Table 26 presents the skewness and kurtosis statistics for the equated IRT scores. We observe that no score follows a normal distribution since we reject the null hypothesis of normal distribution according the test performed; however, in comparison with the raw scores, most of the IRT scores are not skewed. Therefore most of them present a symmetrical distribution (see Appendix E for distribution graphs).

Table 26. *Characteristics of the distribution of IRT scores for reading comprehension, by language*

Language	Original		Corrected	
	Skewness	Kurtosis	Skewness	Kurtosis
Amharic	-0.62*	2.81*	-0.63*	2.79*
Hadiyya	0.11	2.50*	0.08	2.40*
Oromo	-0.16*	2.45*	-0.18*	2.44*
Sidamo	-0.16	2.43*	-0.03	2.16*
Somali	-0.46*	2.52*	-0.46*	2.51*
Tigrinya	-0.35*	2.50*	-0.34*	2.49*
Wolaytta	-0.10	2.22*	-0.07	2.19*

*Statistically significant at 5% according to the sk-test for independent samples.
Source: Ethiopia second school survey, Waves 1 and 2.

Finally, we correlate the IRT scores (Waves 1 and 2) with children's background characteristics. Tables 27 and 28 show the correlation between IRT scores in the two waves and children's gender, age, grade, their family's wealth index, the age they started primary school, whether they attended preschool and whether they repeated a grade. As we expected, the reading comprehension IRT scores are positively correlated with grade, age and wealth index for most languages; negatively correlated with student retention for all languages, and for almost all languages no correlation was found with gender; and with respect to the other variables (age the child started primary school and preschool attendance), the correlations vary by language.

Table 27. *Correlation between IRT scores, Wave 1, and children's demographic and educational background variables*

Variable	Amharic	Hadiyya	Oromo	Sidamo	Somali	Tigrinya	Wolaytta
Female	0.05*	-0.25*	-0.02	-0.13	-0.06	-0.03	0.06
Age	-0.03*	0.07	0.18*	-0.13	0.17*	0.17*	0.10*
Grade	0.15*	0.26*	0.20*	0.09	0.34*	0.22*	0.13*
Wealth index	0.21*	-0.01	0.13*	-0.25*	0.27*	0.21*	0.16*
Age started primary	-0.06*	0.13*	0.20*	-0.14	-0.02	0.17*	0.15*
Attended preschool	0.20*	-0.11	0.08*	-0.05	-0.03	0.13*	0.10
Repeated a grade	-0.26*	-0.22*	-0.24*	-0.43*	-0.25*	-0.21*	-0.09

* Statistically significant at 5%.
Source: Ethiopia second school survey, Waves 1 and 2.

Table 28. *Correlation between IRT scores, Wave 2, and children's demographic and educational background variables*

Variable	Amharic	Hadiyya	Oromo	Sidamo	Somali	Tigrinya	Wolaytta
Female	0.05*	-0.17*	0.03	-0.12	-0.06	0.03	0.08
Age	-0.09*	0.08	0.14*	-0.15	0.14*	0.14*	-0.02
Grade	0.12*	0.13*	0.16*	0.04	0.38*	0.23*	0.13*
Wealth index	0.21*	-0.16*	0.15*	-0.17	0.30*	0.19*	0.10
Age started primary	-0.10*	0.11	0.17*	-0.10	0.02	0.12*	0.09
Attended preschool	0.22*	-0.11	0.16*	0.02	0.13*	0.13*	-0.05
Repeated a grade	-0.22*	-0.14*	-0.19*	-0.34*	-0.01	-0.20*	-0.12*

* Statistically significant at 5%.
Source: Ethiopia second school survey, Waves 1 and 2.

5. Final remarks

This technical note gives details of the procedures followed to equate the maths and reading comprehension scores for the main languages in the school survey administered in Ethiopia, the main results being:

- It was possible to ensure an adequate equating of the maths and reading comprehension scores for all the main languages across waves in Ethiopia and the DIF analysis performed by wave indicates that most of the items used do not have presence of bias by wave or gender.
- In maths and reading comprehension, we did not observe gender differences for most languages, Amharic- and Hadiyya-speaking students being the only ones who showed gender differences in both grades.
- We observed for math and reading comprehension that scores gains (between Waves 1 and 2) were statistically significant for most languages, being in most cases around 0.2 SD.
- Finally, as we expected, the correlations between IRT scores (maths and reading comprehension) and age, grade and family wealth index were statistically significant with the expected signs (positively correlated with grade, age and wealth index for most languages and negatively correlated with student retention for all languages), and showed almost no correlation with gender, indicating that differences between boys' and girls' scores are not significant in any of the languages in which the tests were administered.

References

Aurino, E., Z. James and C. Rolleston (2015) *Young Lives Ethiopia School Survey 2012–13: Data Overview Report*, Working Paper 134, Oxford: Young Lives.

Crocker, L. and J. Algina (1986) *Introduction to Classical and Modern Test Theory*, Orlando, FL: Holt, Rinehart and Winston.

Cueto, S. and J. León (2013) *Psychometric Characteristics of Cognitive Development and Achievement Instruments in Round 3 of Young Lives*, Technical Note 25, Oxford: Young Lives.

Cueto, S., J. León, G. Guerrero and I. Muñoz (2009) *Psychometric Characteristics of Cognitive Development and Achievement Instruments in Wave 2 of Young Lives*, Technical Note 15, Oxford: Young Lives.

Frost, M. and C. Rolleston (2013) *Improving Education Quality, Equity and Access: A Report on Findings from the Young Lives School Survey (Wave 1) in Ethiopia*, Working Paper 96, Oxford, Young Lives.

Hambleton, R.K. (1989) 'Principles and Selected Applications of Item Response Theory' in R.L. Linn (ed.) *Educational Measurement*, 3rd ed., New York: Macmillan.

Linacre, J.M. (2008) *Winsteps: A Rasch Analysis Computer Program* (Version 3.68), http://www.winsteps.com.

Nunnally, J.C. and I.H. Bernstein (1994) *Psychometric Theory,* New York: McGraw-Hill.

Wright B.D. and GA. Douglas (1976) *Rasch Item Analysis by Hand*, MESA Research Memorandum Number 21, Chicago, IL: Statistical Laboratory, Department of Education, University of Chicago.

Appendix A: Distribution graphs for raw corrected scores by language and area

Maths

Figure A1. *Amharic maths raw scores by wave*

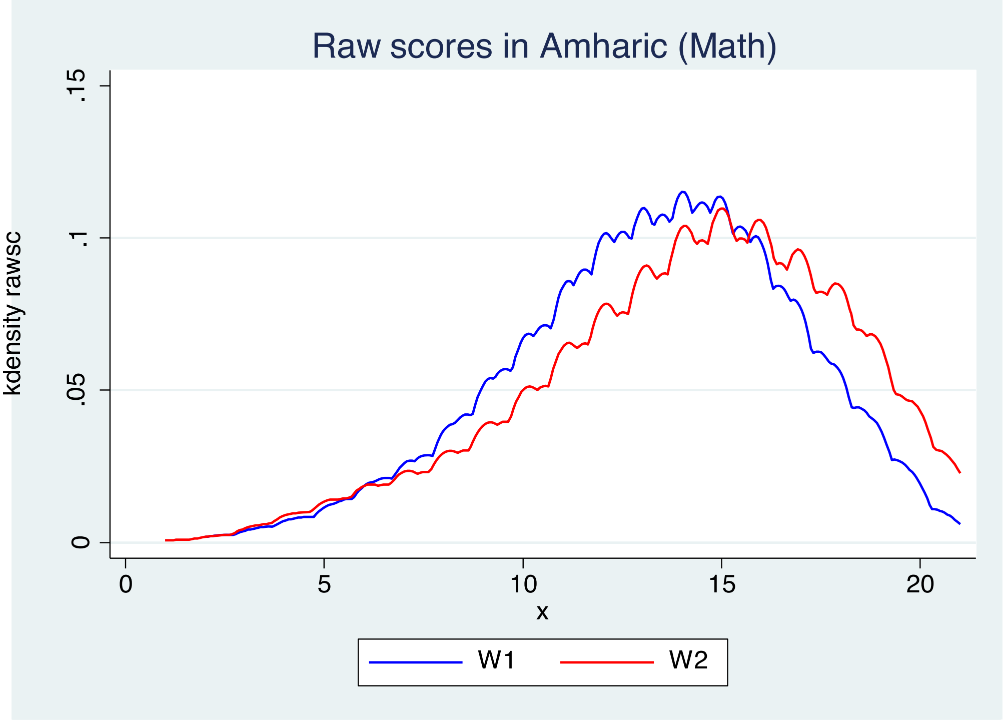


Figure A2. *Hadiyya maths raw scores by wave*

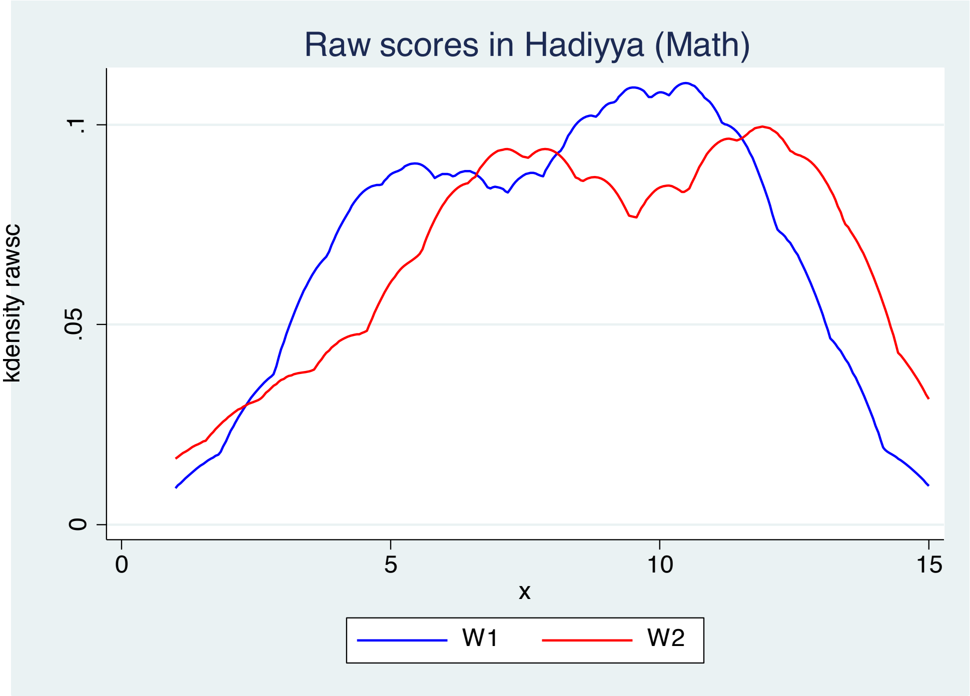


Figure A3. *Oromo maths raw scores by wave*

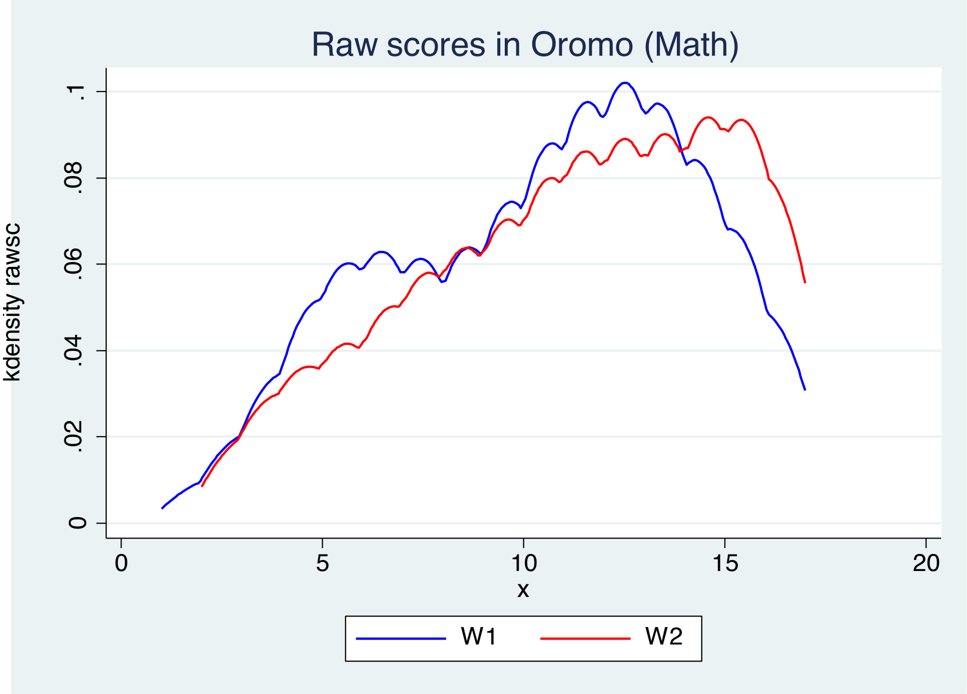


Figure A4. *Sidamo maths raw scores by wave*

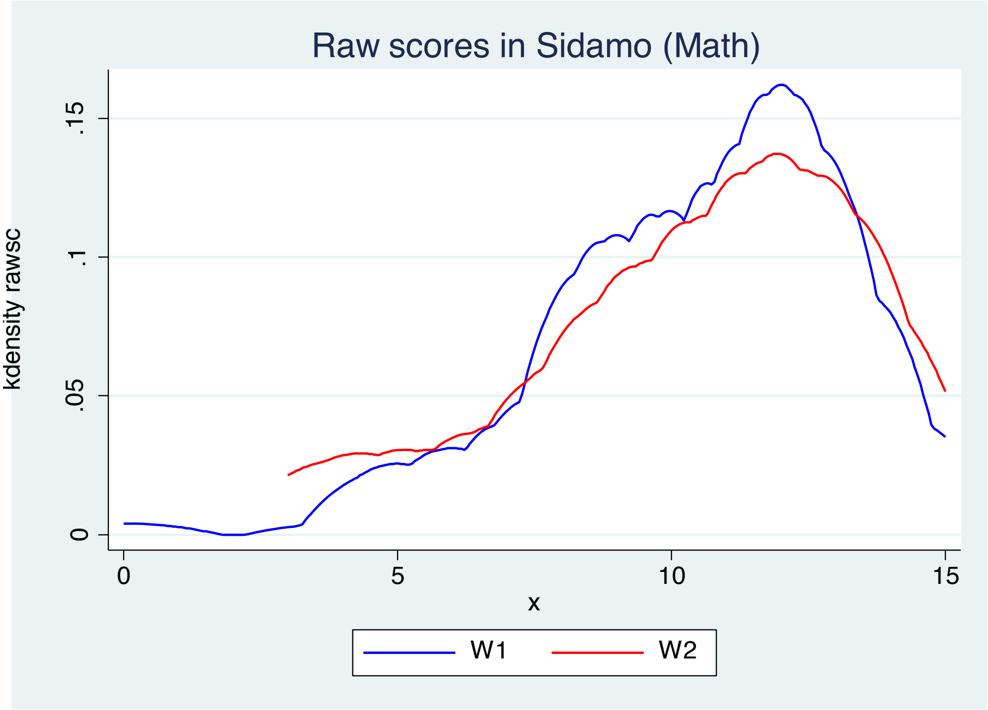


Figure A5. *Somali maths raw scores by wave*

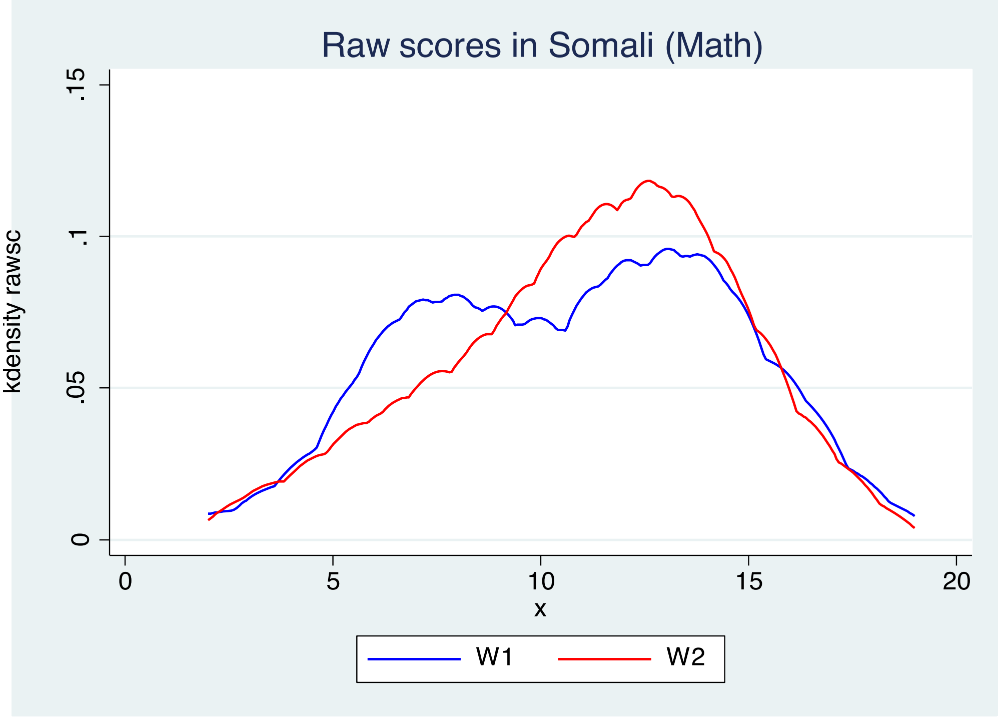


Figure A6. *Tigrinya maths raw scores by wave*

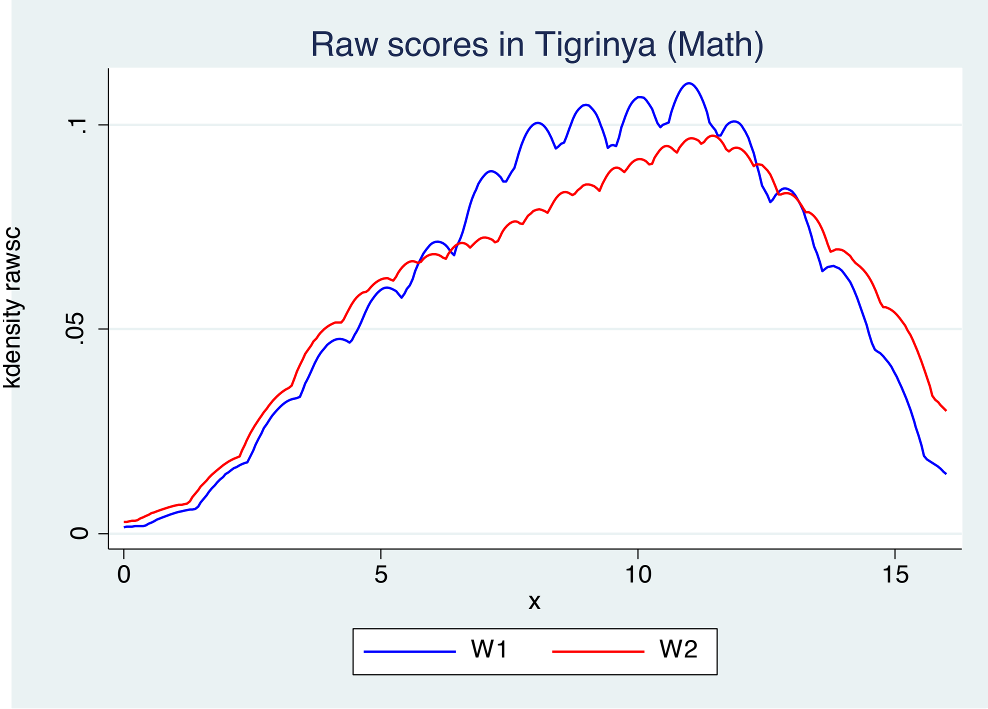
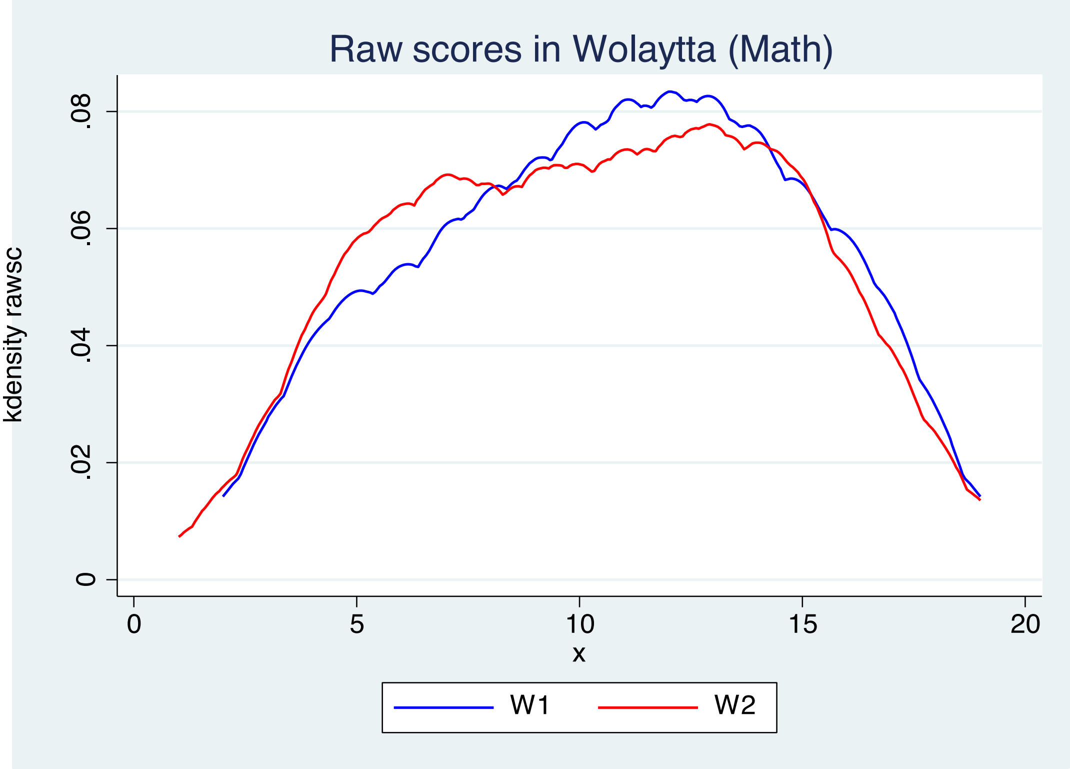


Figure A7. *Wolaytta maths raw scores by wave*



Reading comprehension

Figure A8. *Amharic maths raw scores by wave*

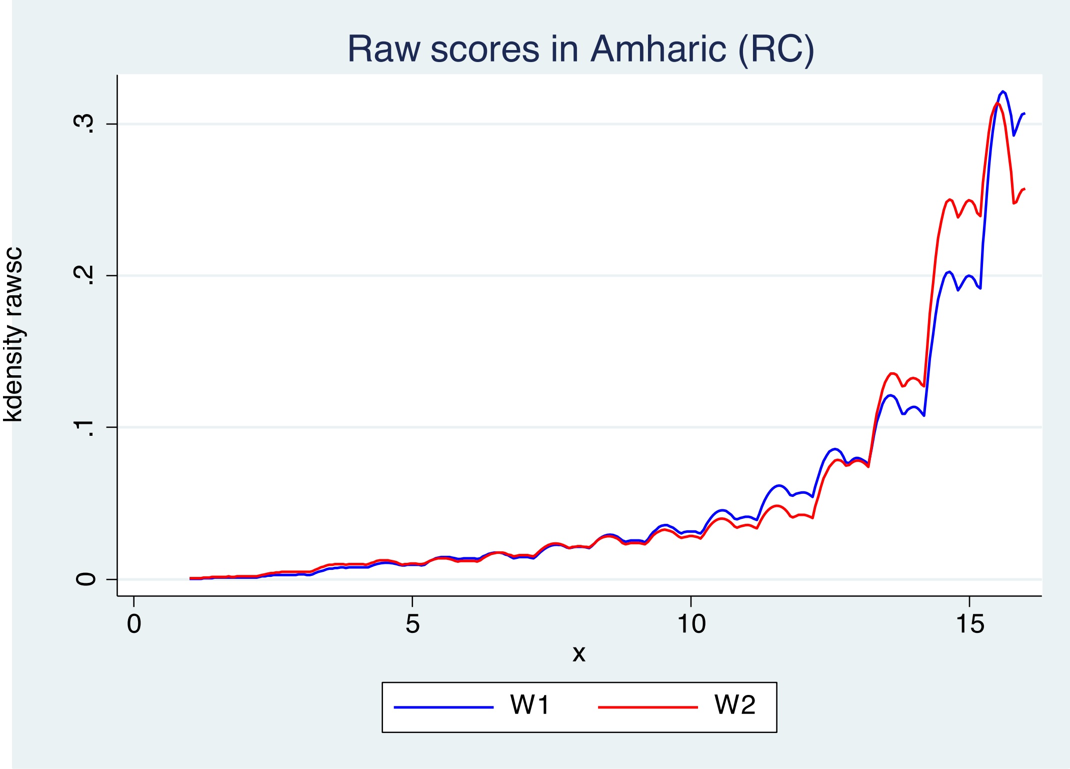


Figure A9. *Hadiyya maths raw scores by wave*

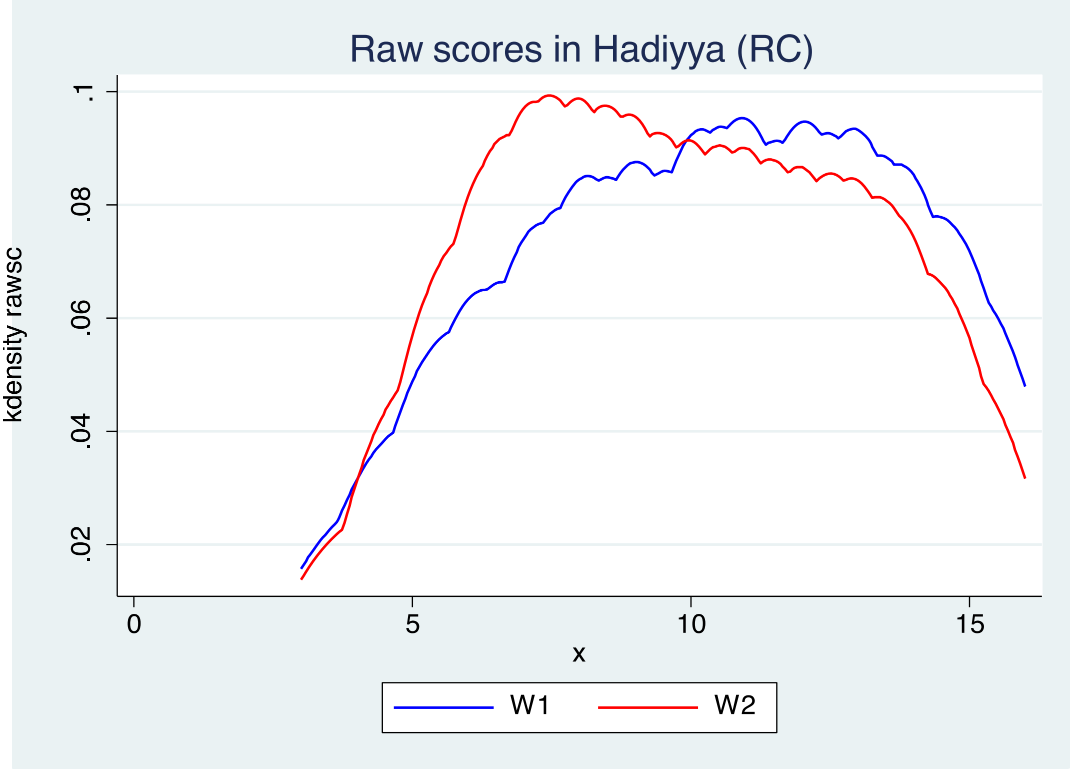


Figure A10. *Oromo maths raw scores by wave*

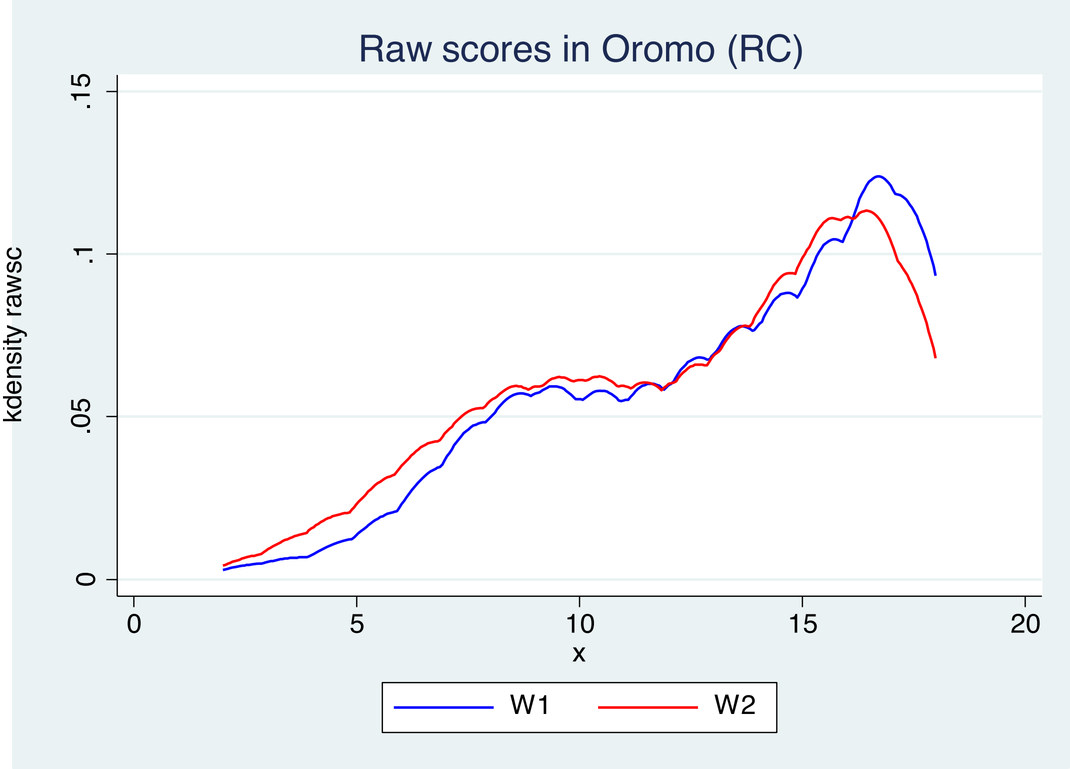


Figure A11. *Sidamo maths raw scores by wave*

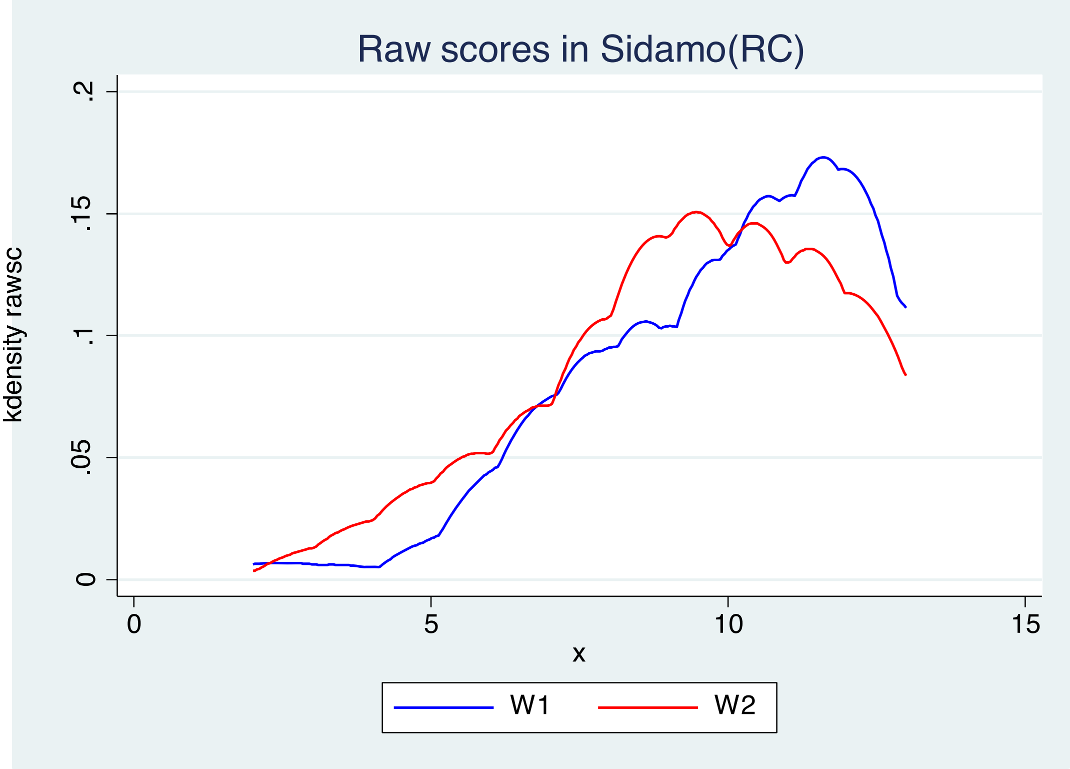


Figure A12. *Somali maths raw scores by wave*

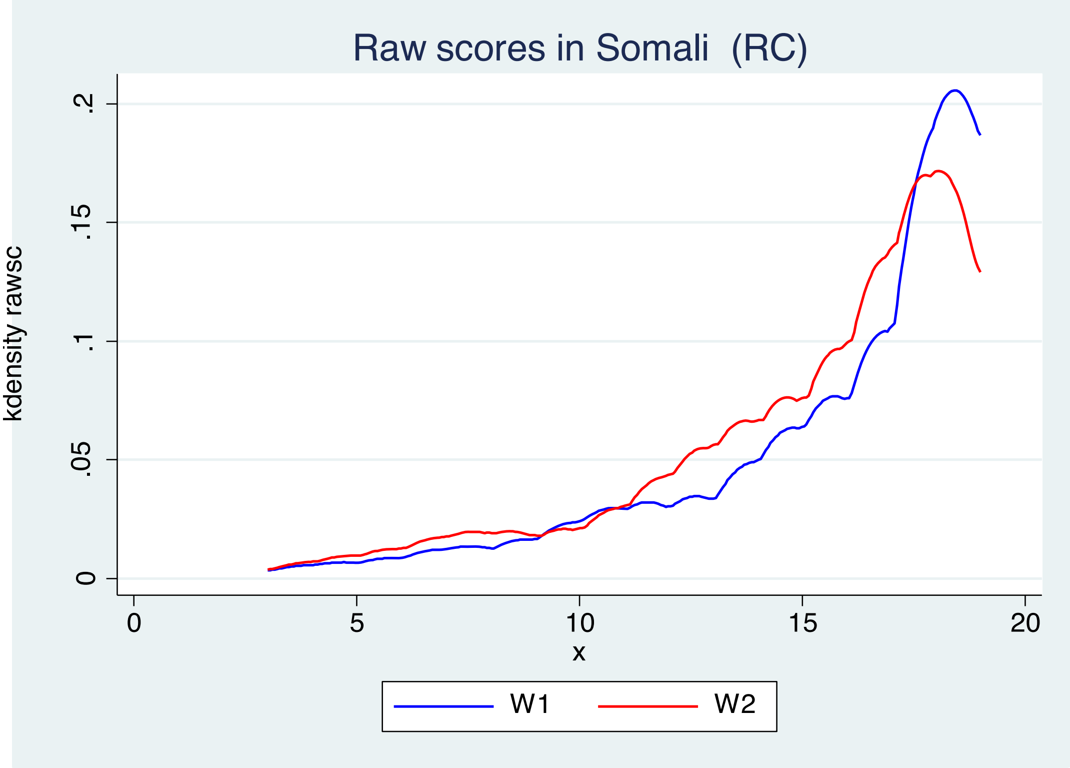


Figure A13. *Tigrinya maths raw scores by wave*

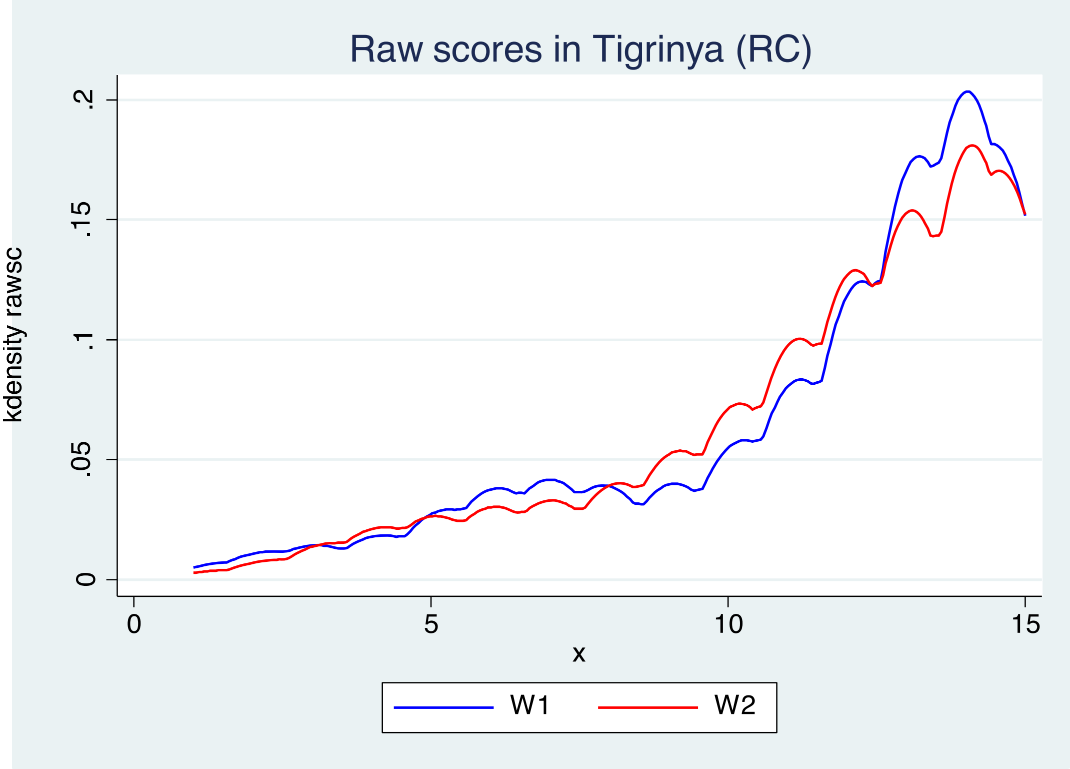
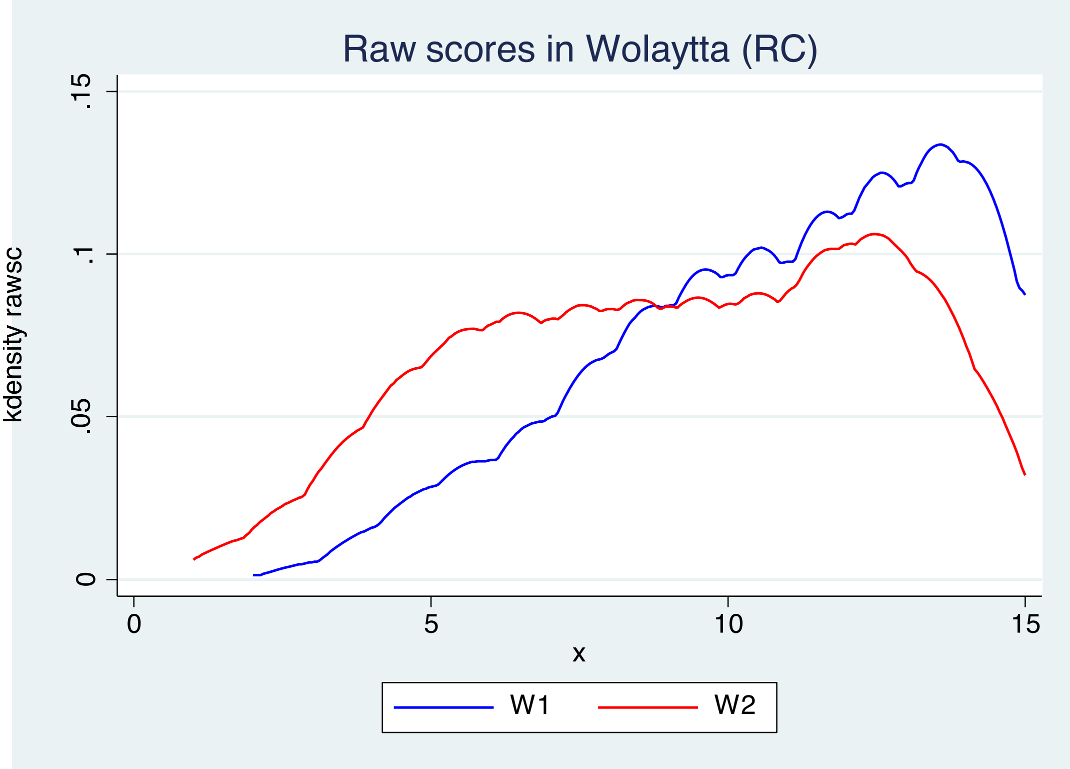


Figure A14. *Wolaytta maths raw scores by wave*

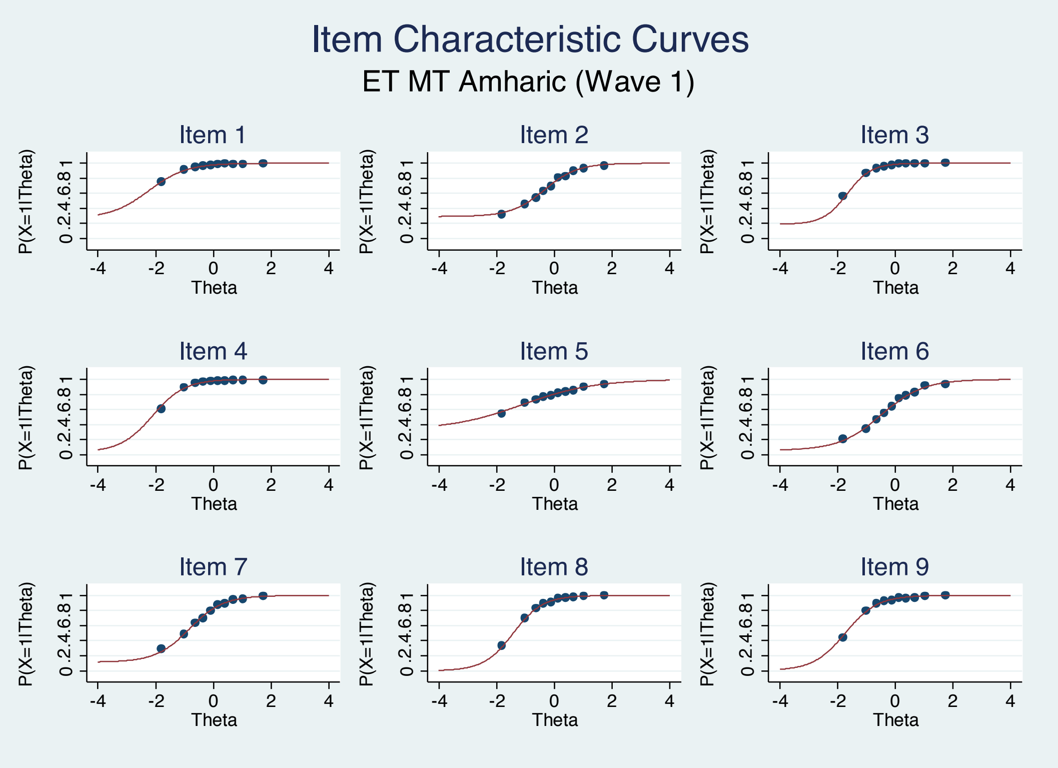


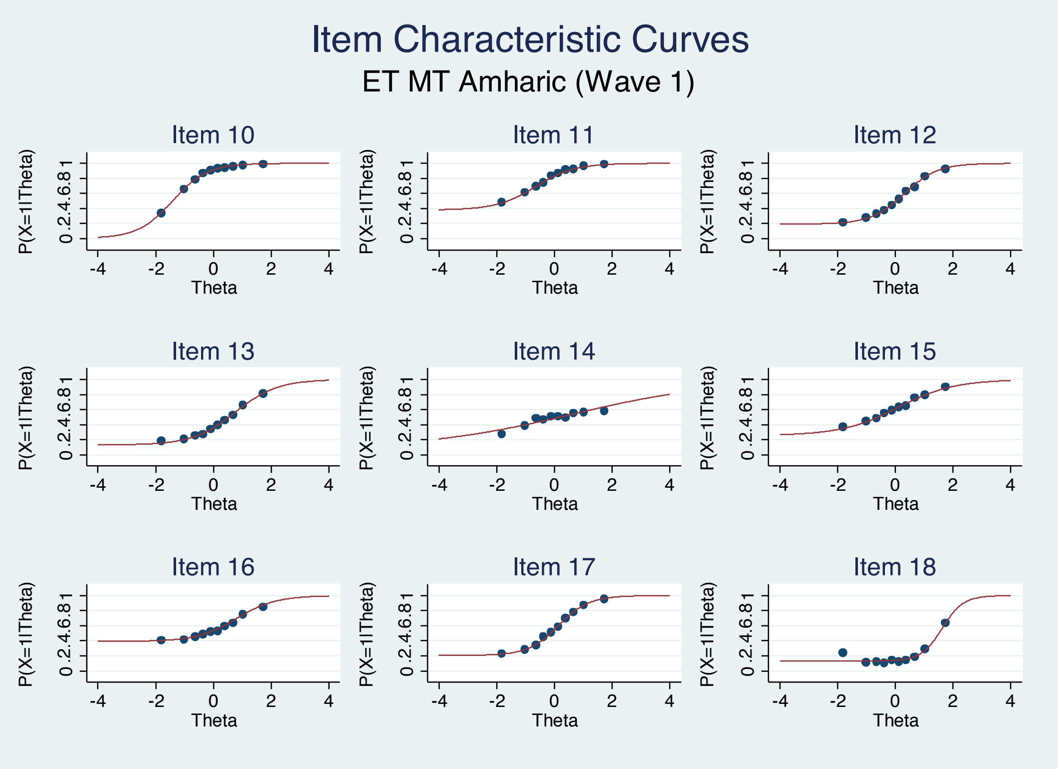
Appendix B: Item characteristic curves (ICCs) for maths and reading comprehension items in Waves 1, 2 and pooled sample, by language and area

Maths

Wave 1

Figure B1. *ICC for maths test for Amharic*




Item Characteristic Curves
ET MT Amharic (Wave 1)
Item 10
Item 11
Item 12
Item 13
Item 14
Item 15
Item 16
Item 17
Item 18
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

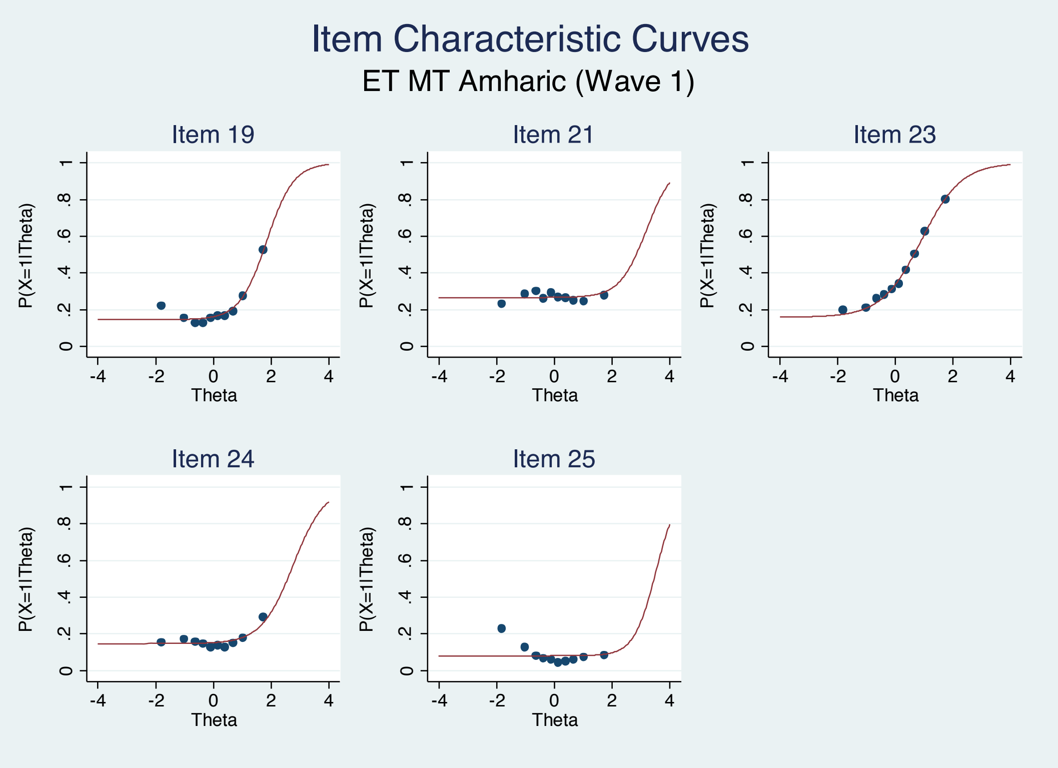

Item Characteristic Curves
ET MT Amharic (Wave 1)
Item 19
Item 21
Item 23
Item 24
Item 25
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B2. *ICC for maths test for Hadiyya*

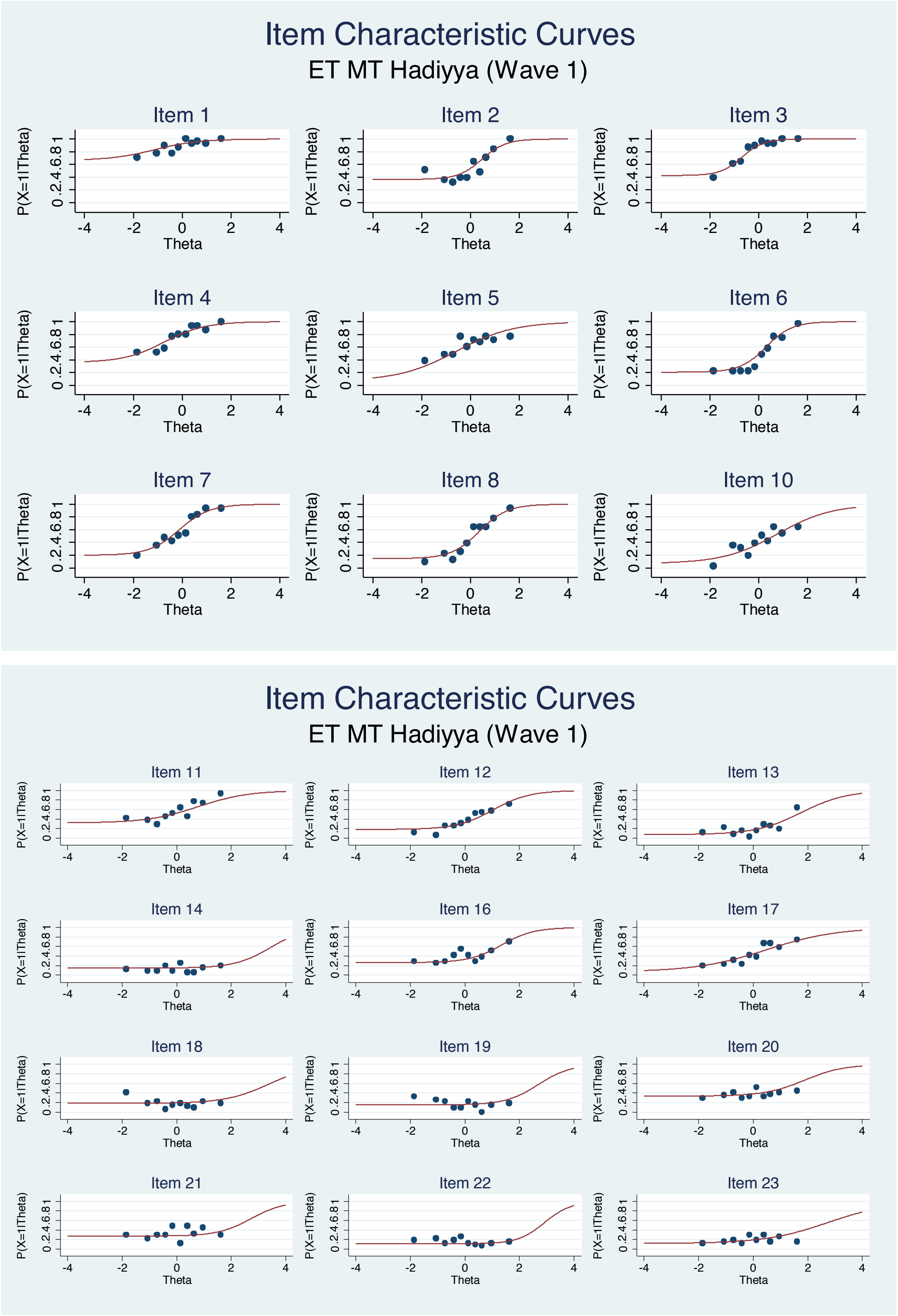


Figure B3. *ICC for maths test for Oromo*

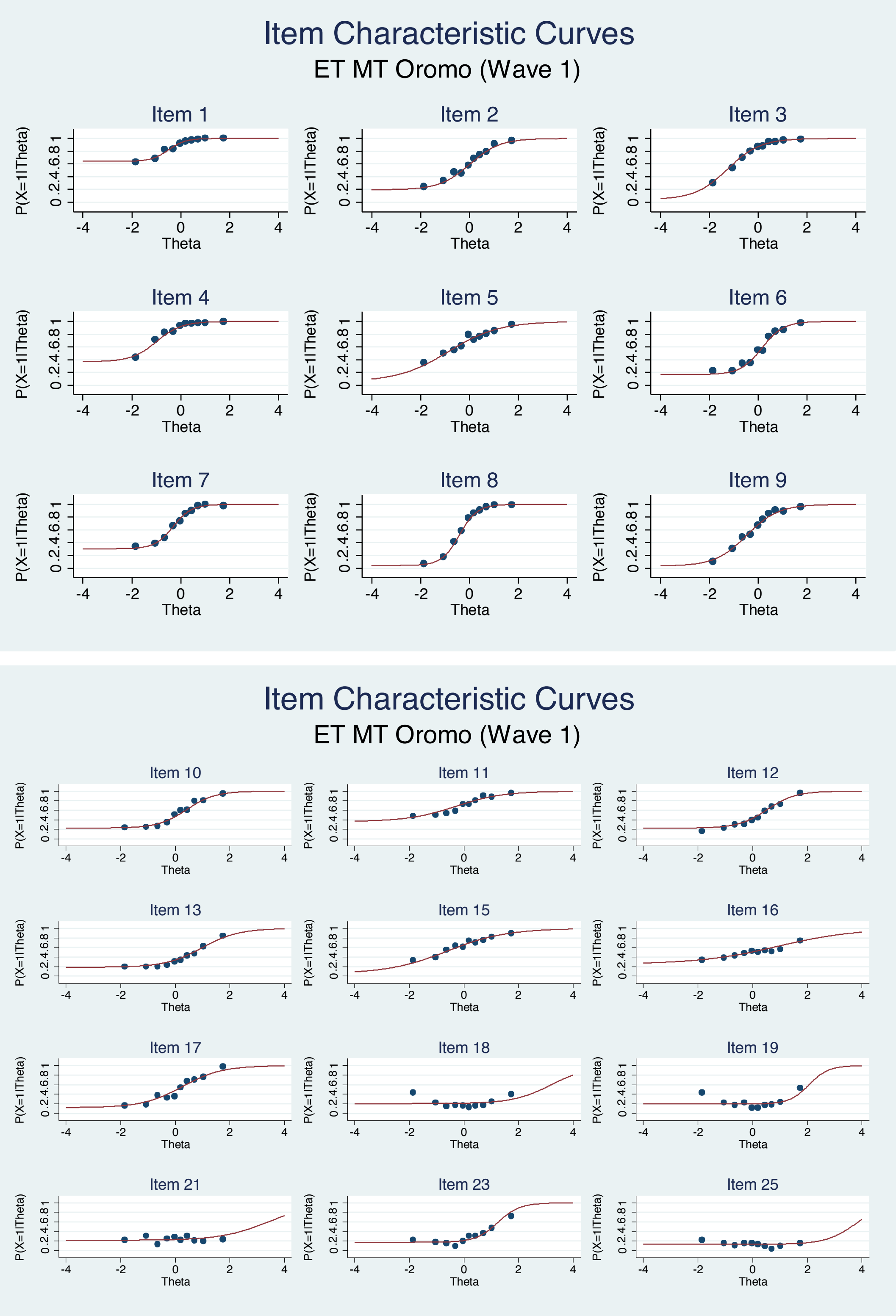
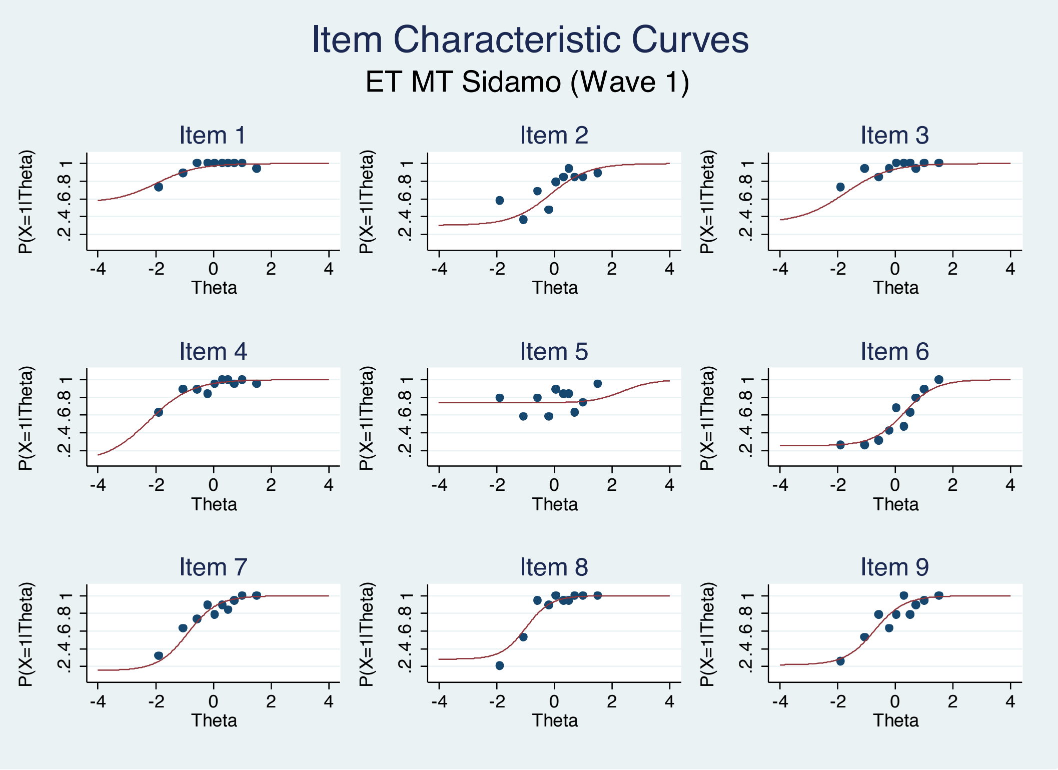
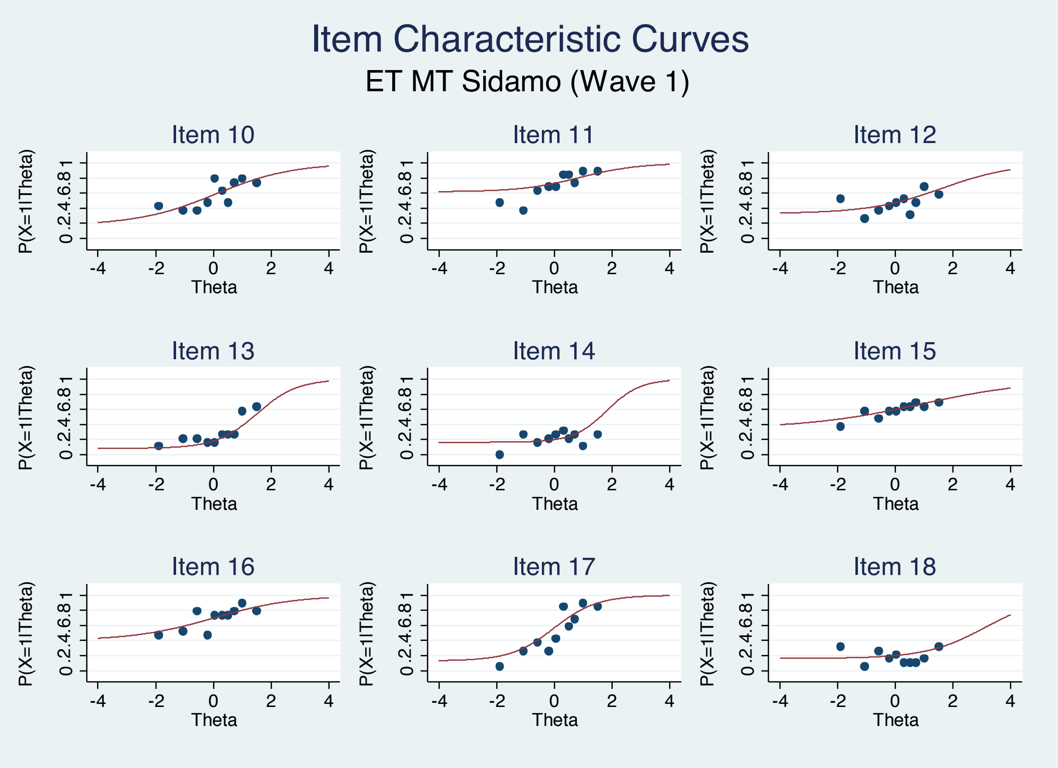


Figure B4. *ICC for maths test for Sidamo*





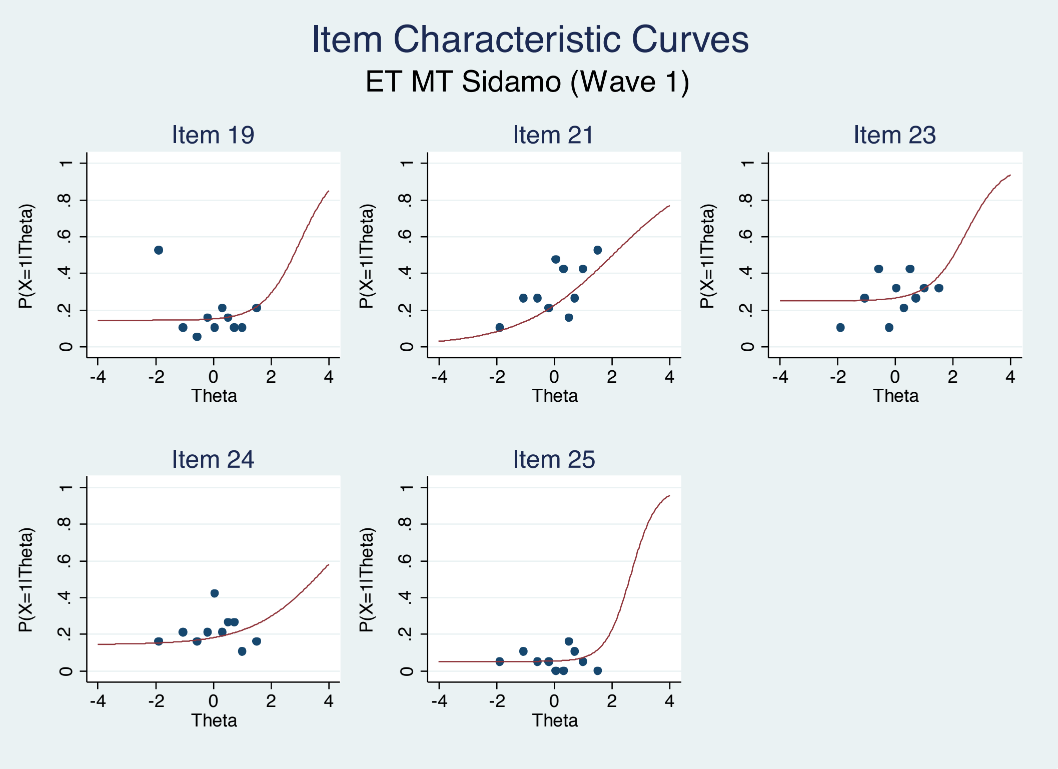
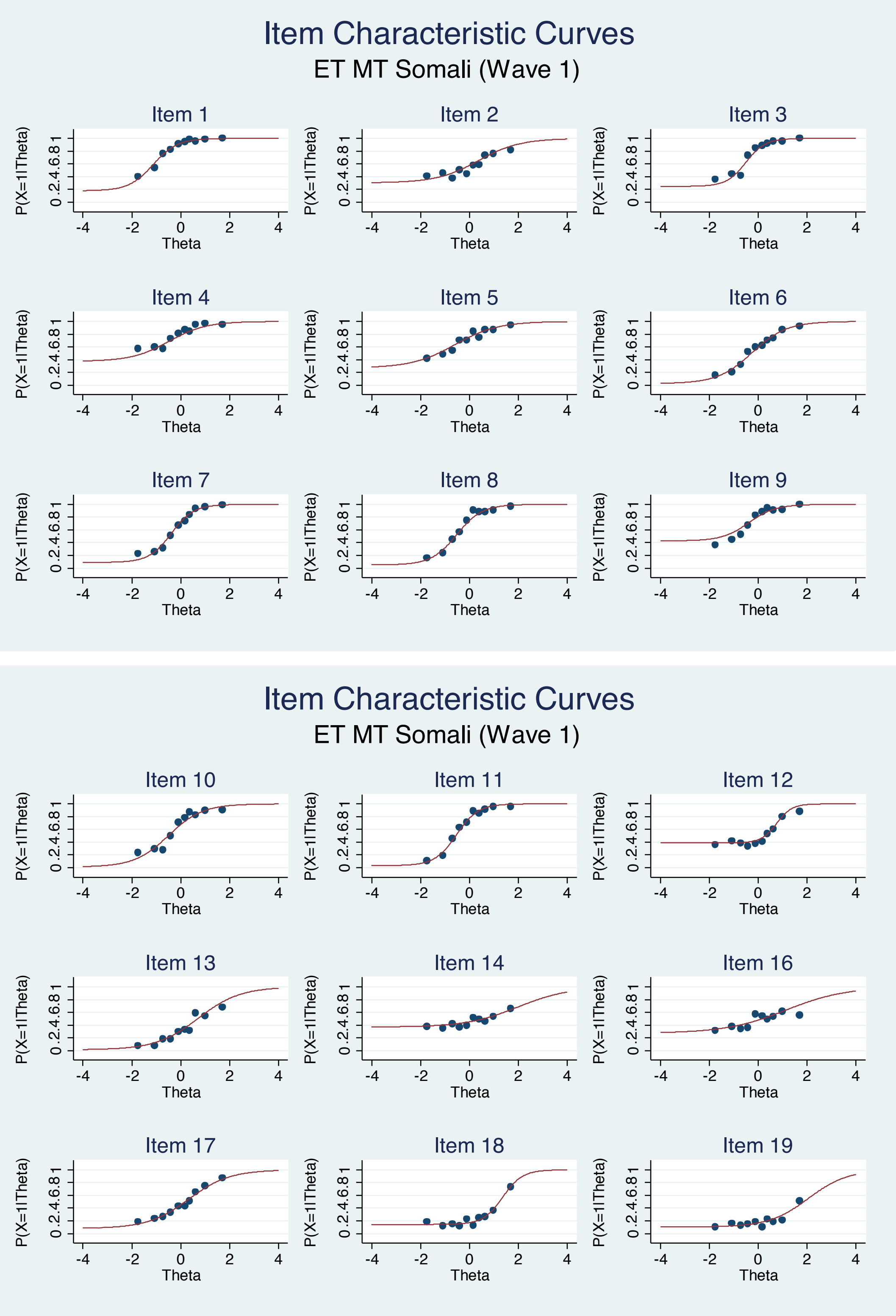

Item Characteristic Curves
ET MT Sidamo (Wave 1)
Item 19
Item 21
Item 23
Item 24
Item 25
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B5. *ICC for maths test for Somali*



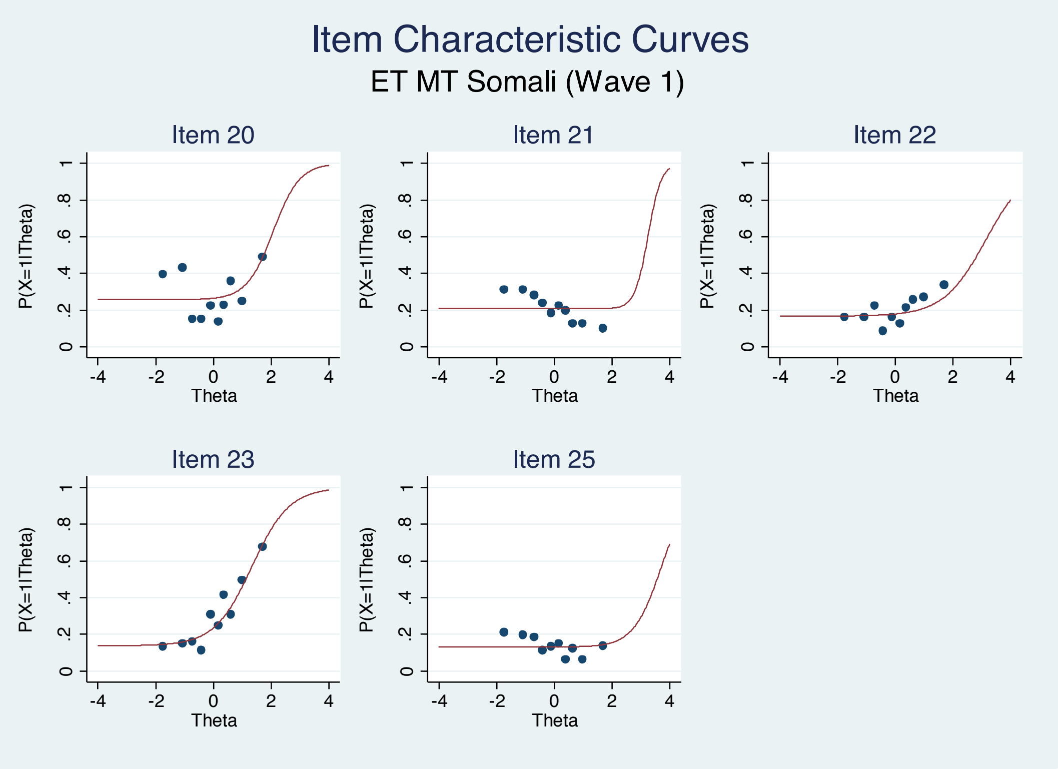

Item Characteristic Curves
ET MT Somali (Wave 1)
Item 20
Item 21
Item 22
Item 23
Item 25
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B6. *ICC for maths test for Tigrinya*

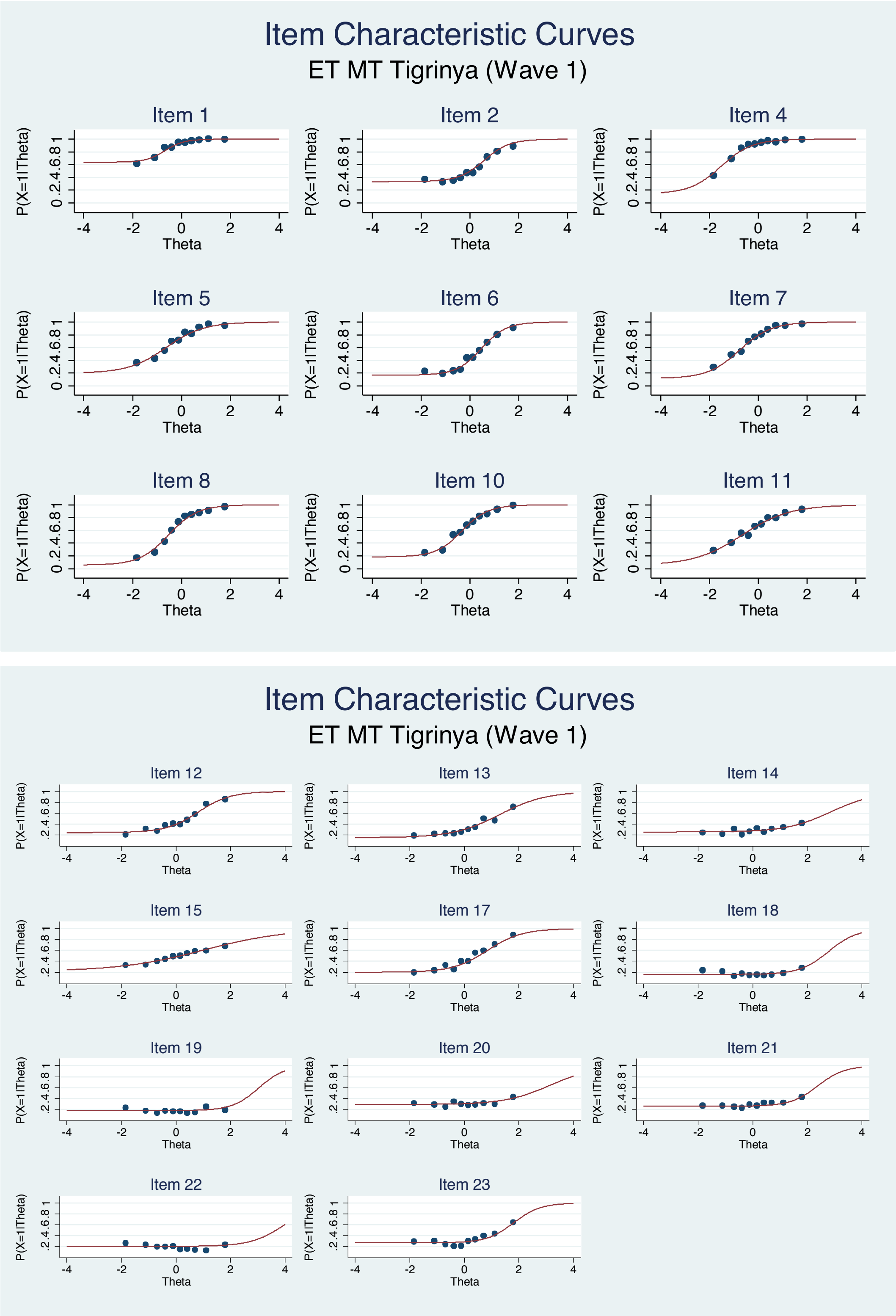
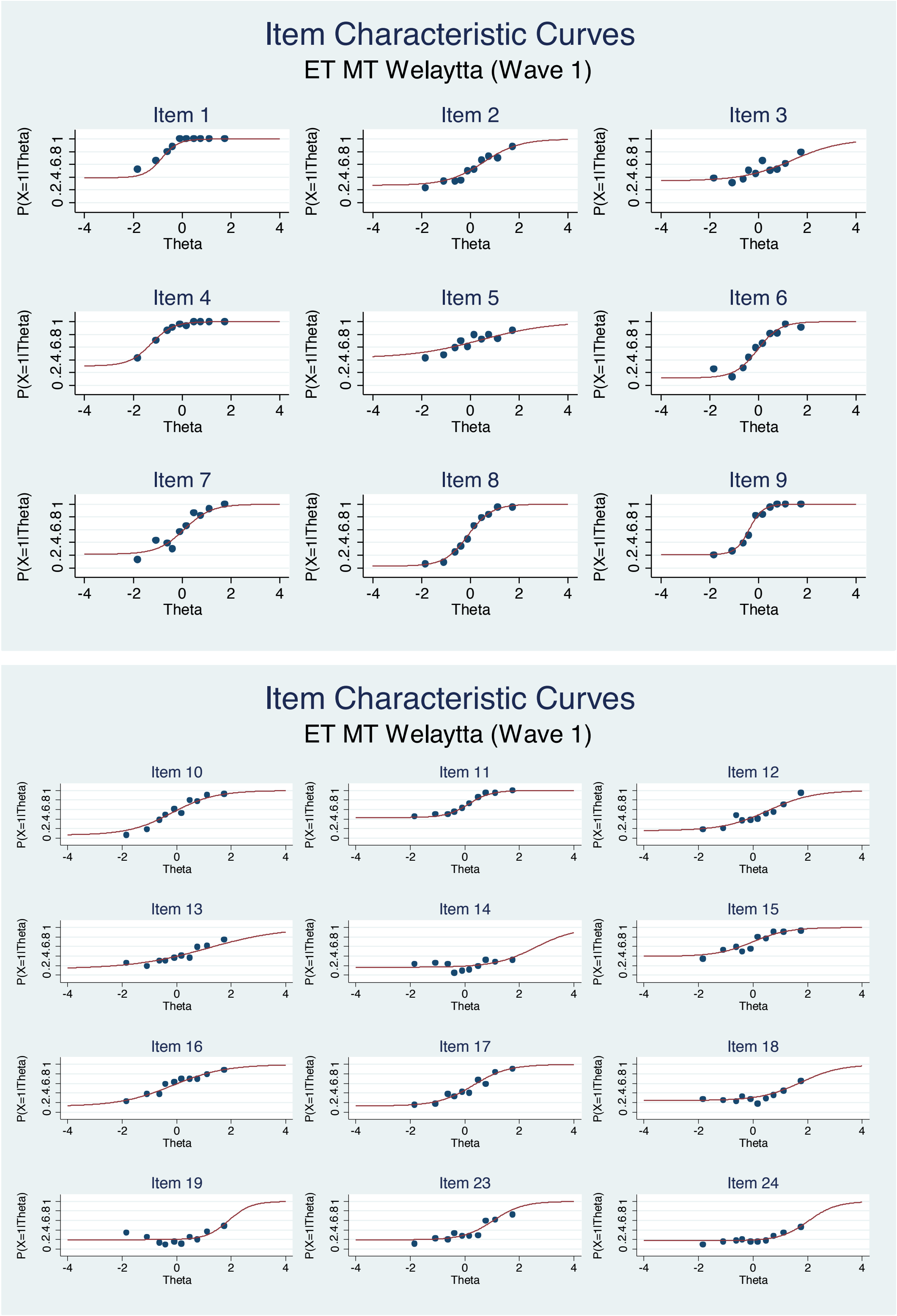
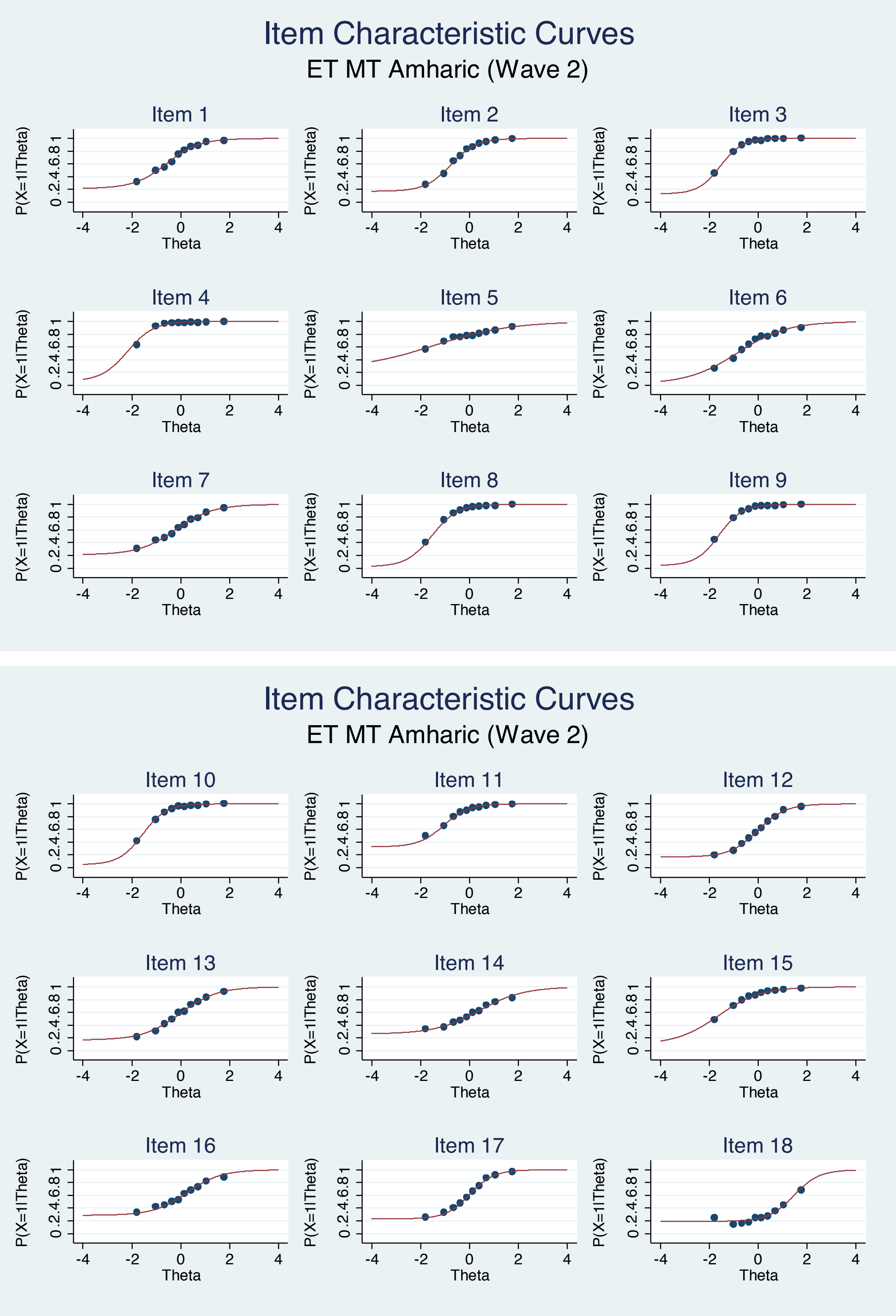


Figure B7. *ICC for maths test for Wolaytta*



Wave 2

Figure B8. *ICC for maths test for Amharic*



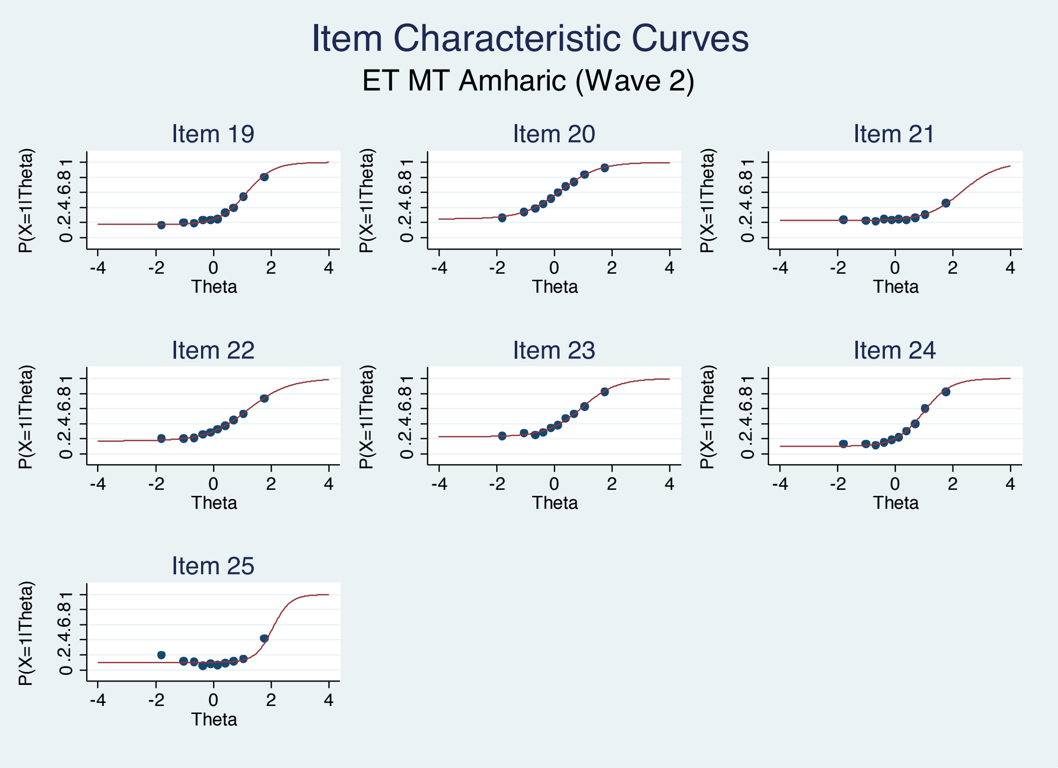
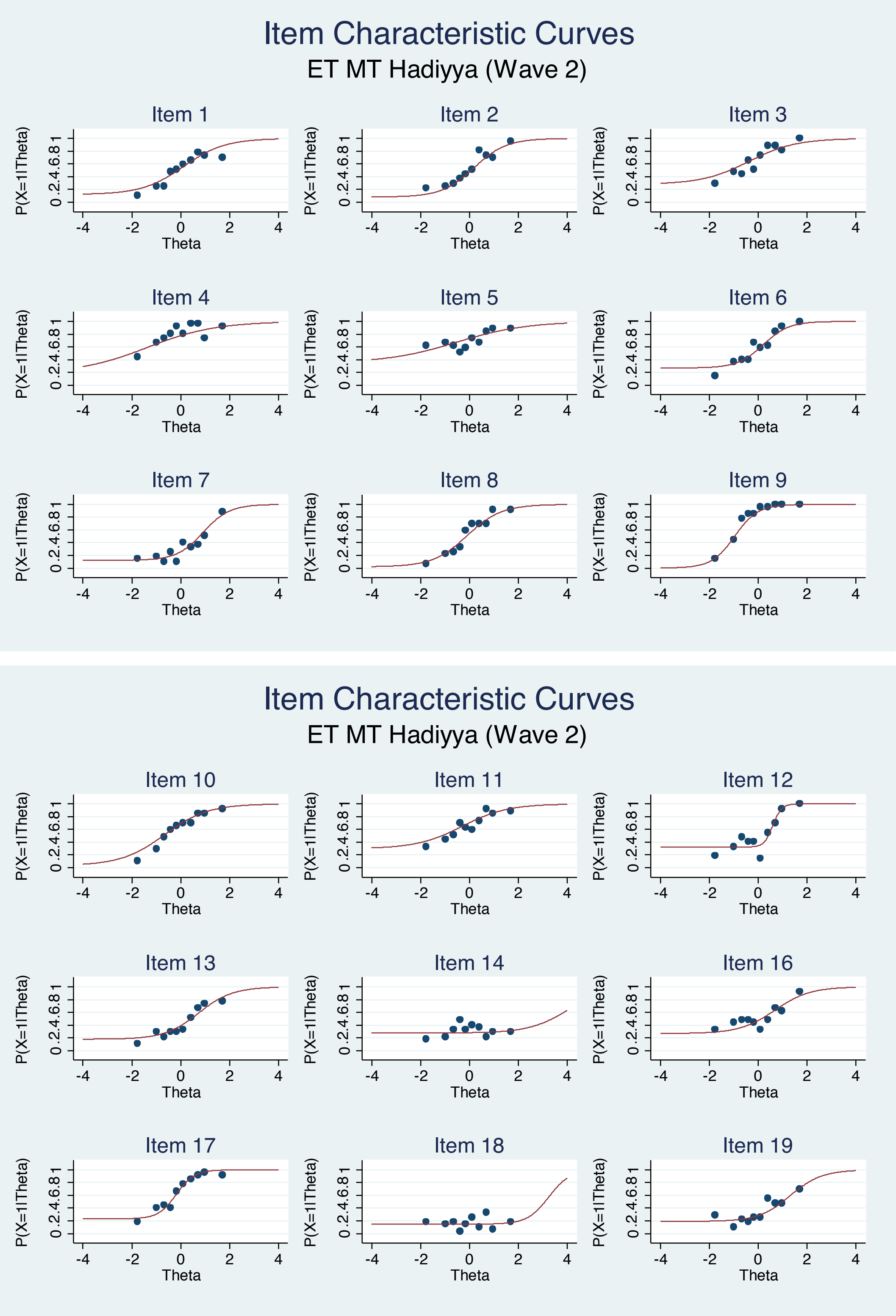

Item Characteristic Curves
ET MT Amharic (Wave 2)
Item 19
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B9. *ICC for maths test for Hadiyya*



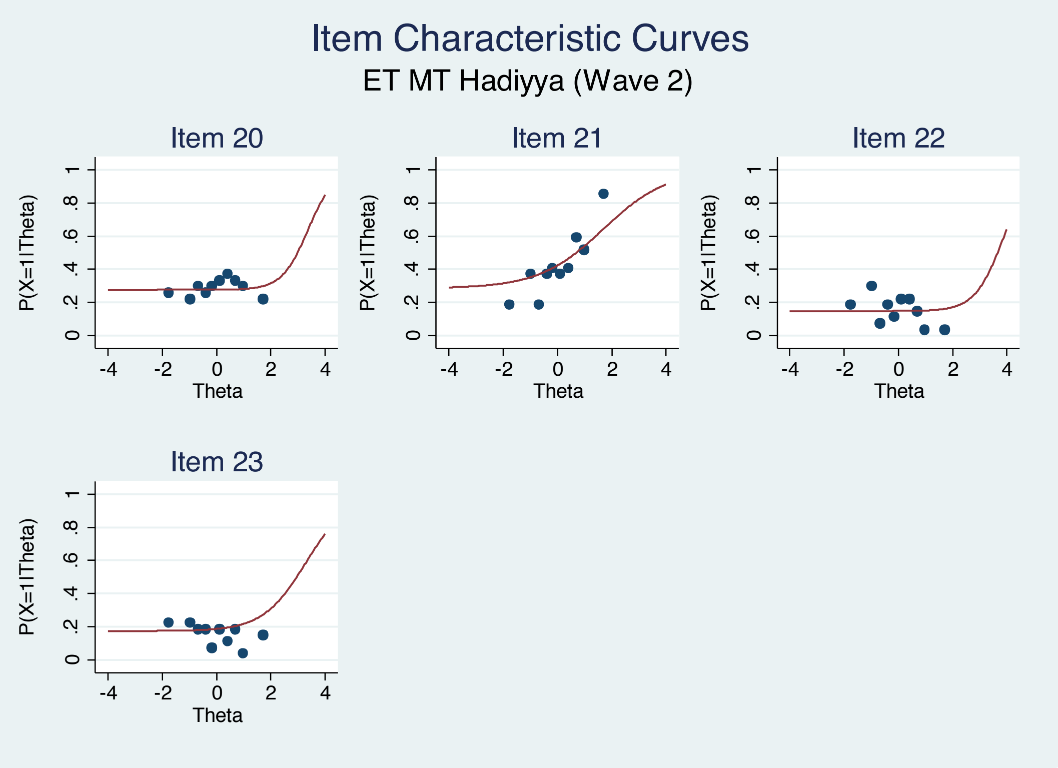

Item Characteristic Curves
ET MT Hadiyya (Wave 2)
Item 20
Item 21
Item 22
Item 23
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B10. *ICC for maths test for Oromo*

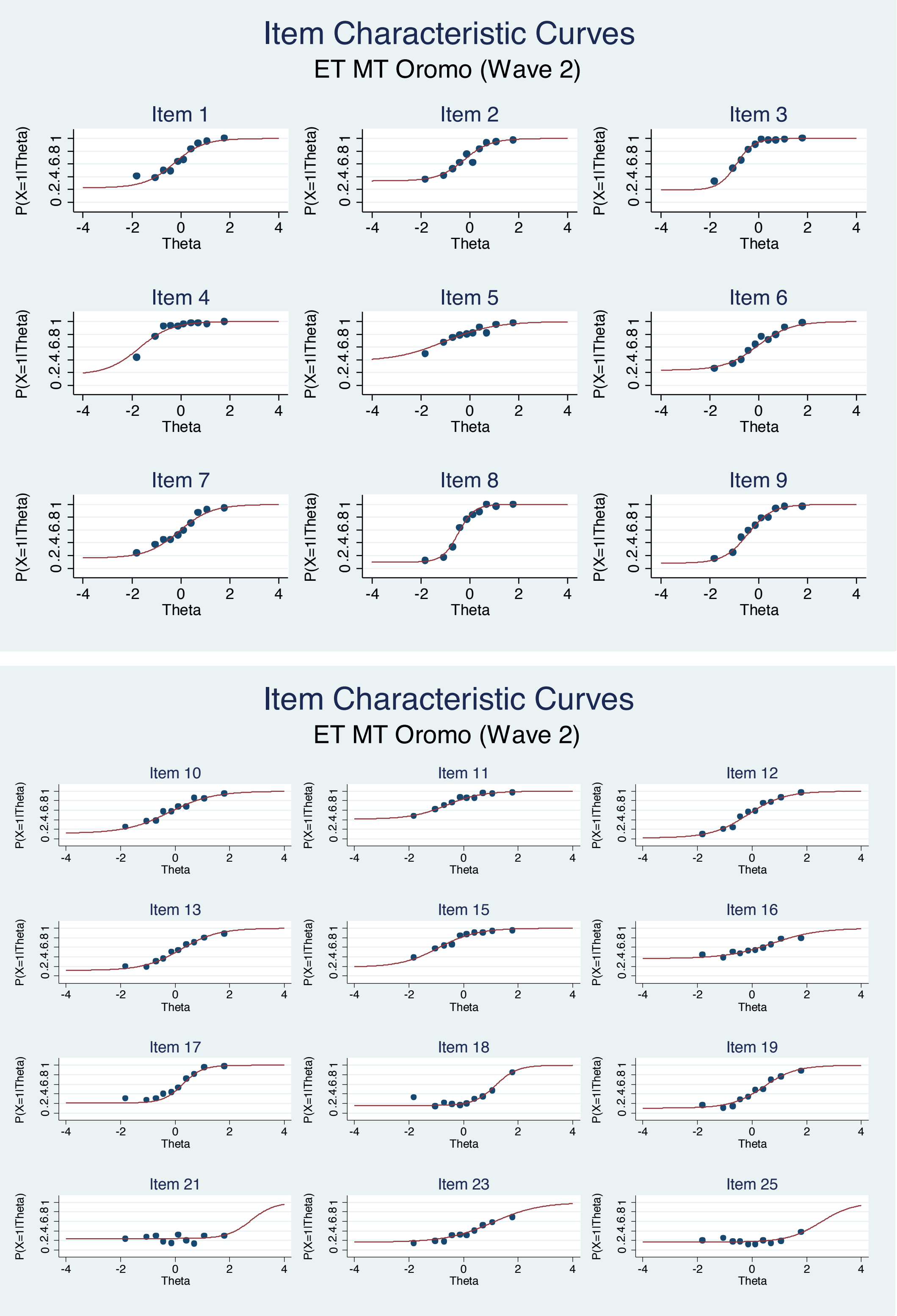
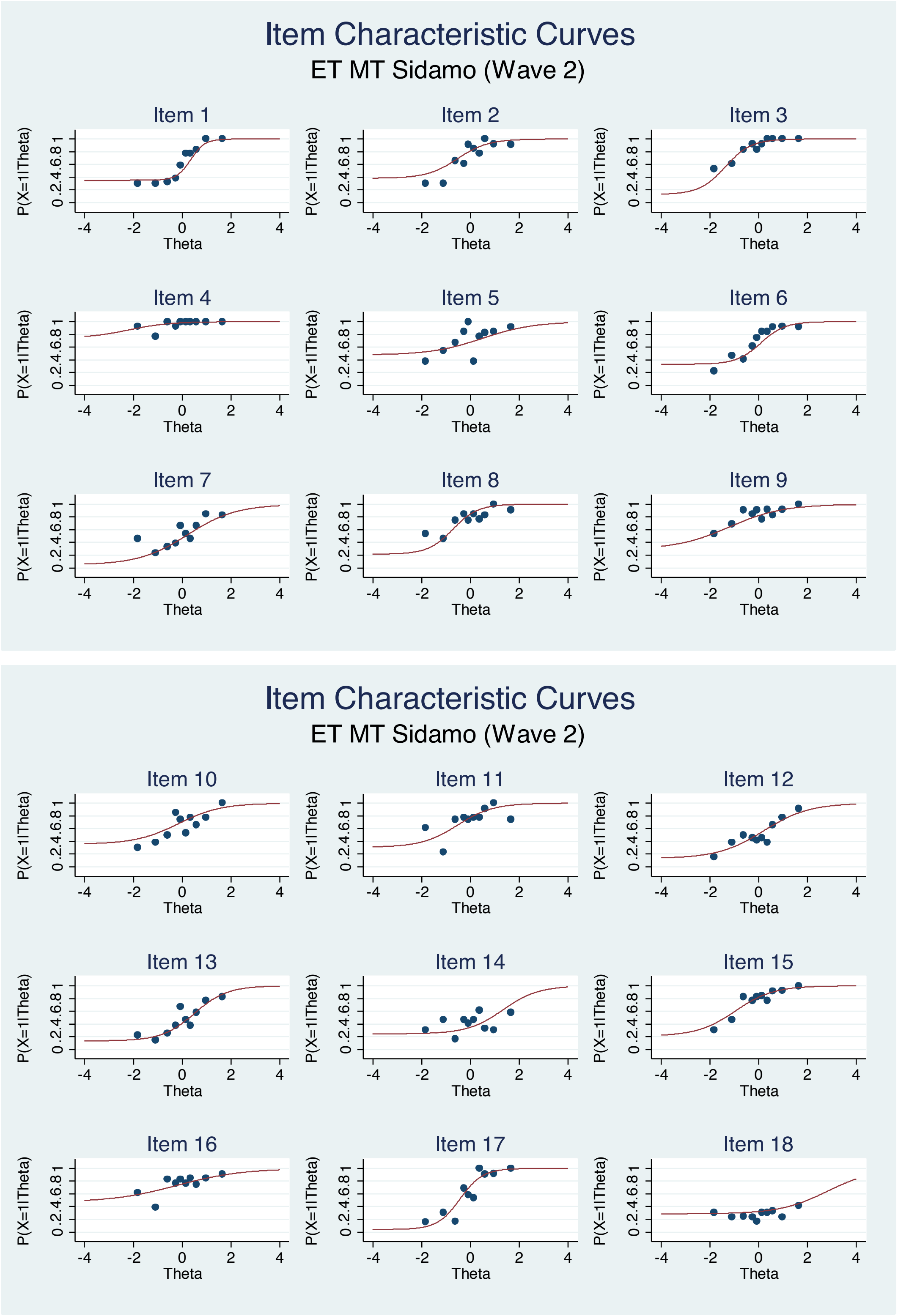


Figure B11. *ICC for maths test for Sidamo*



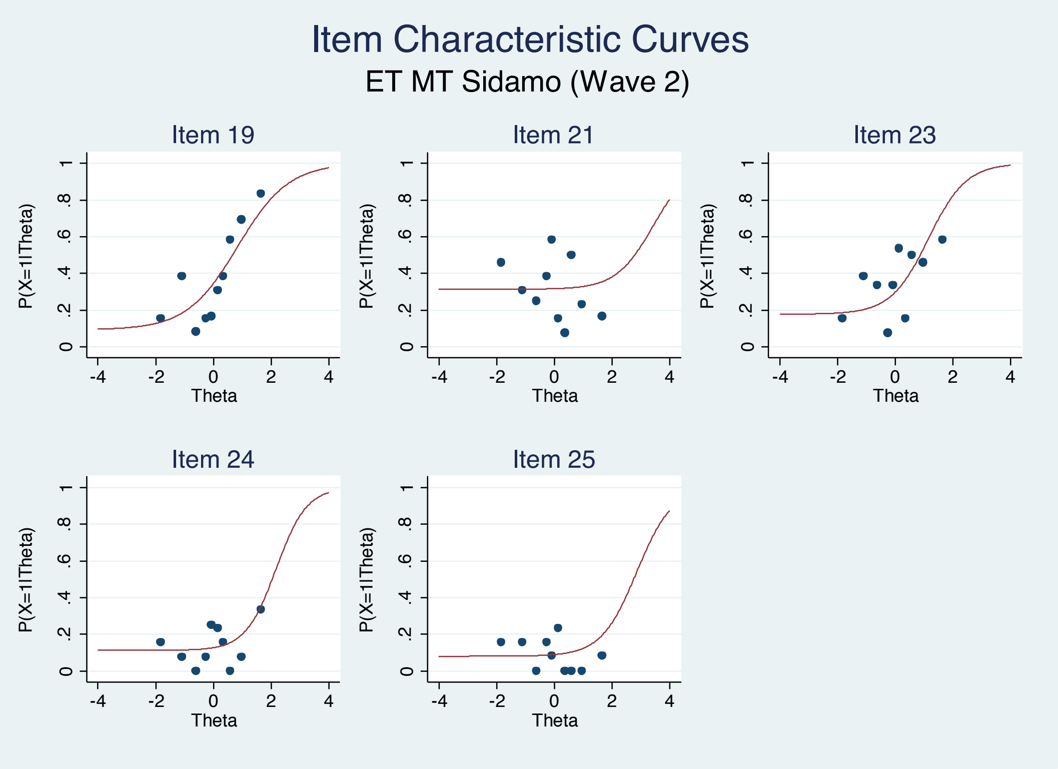
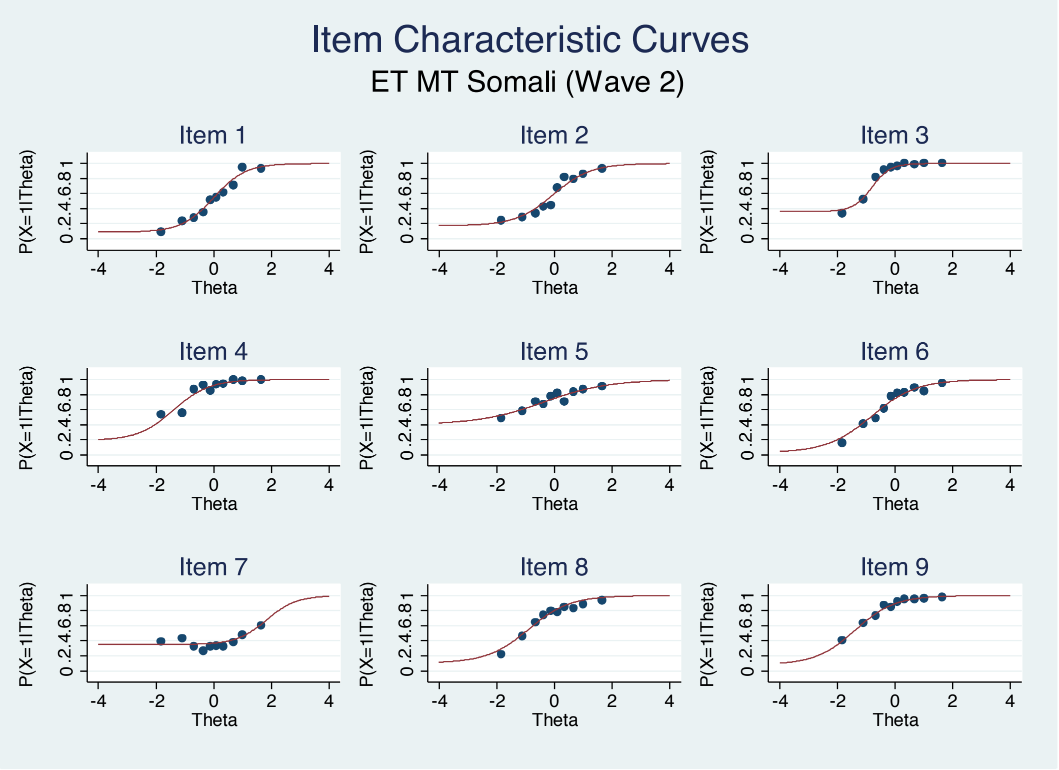
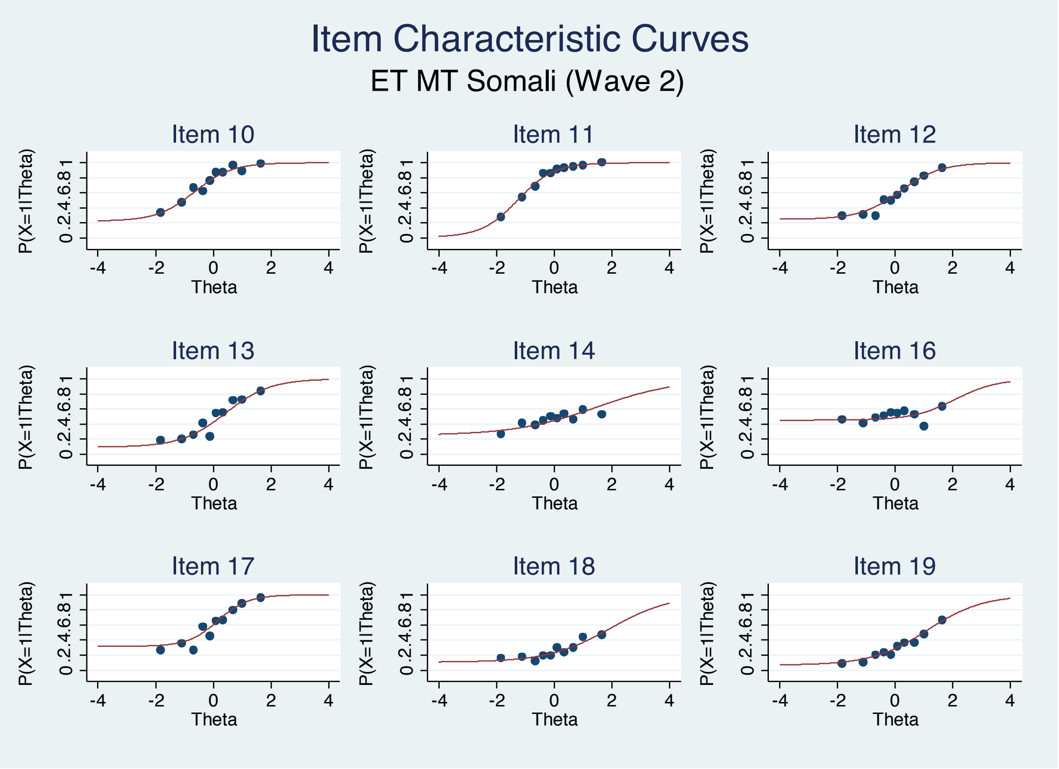

Item Characteristic Curves
ET MT Sidamo (Wave 2)
Item 19
Item 21
Item 23
Item 24
Item 25
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B12. *ICC for maths test for Somali*





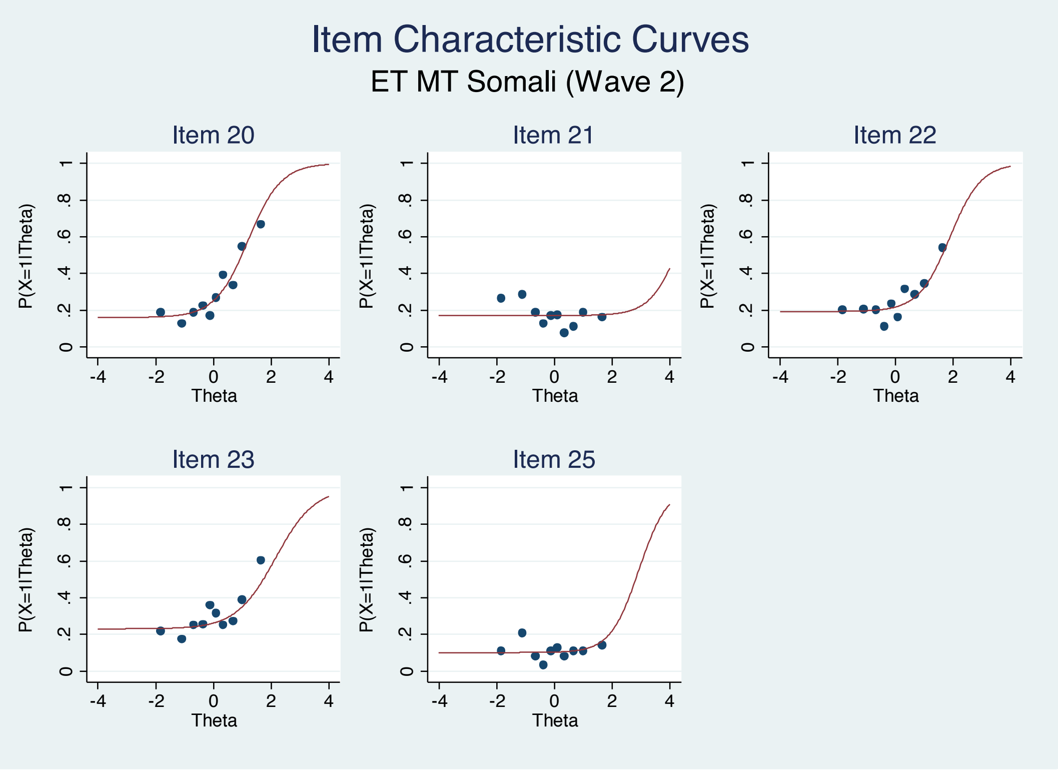
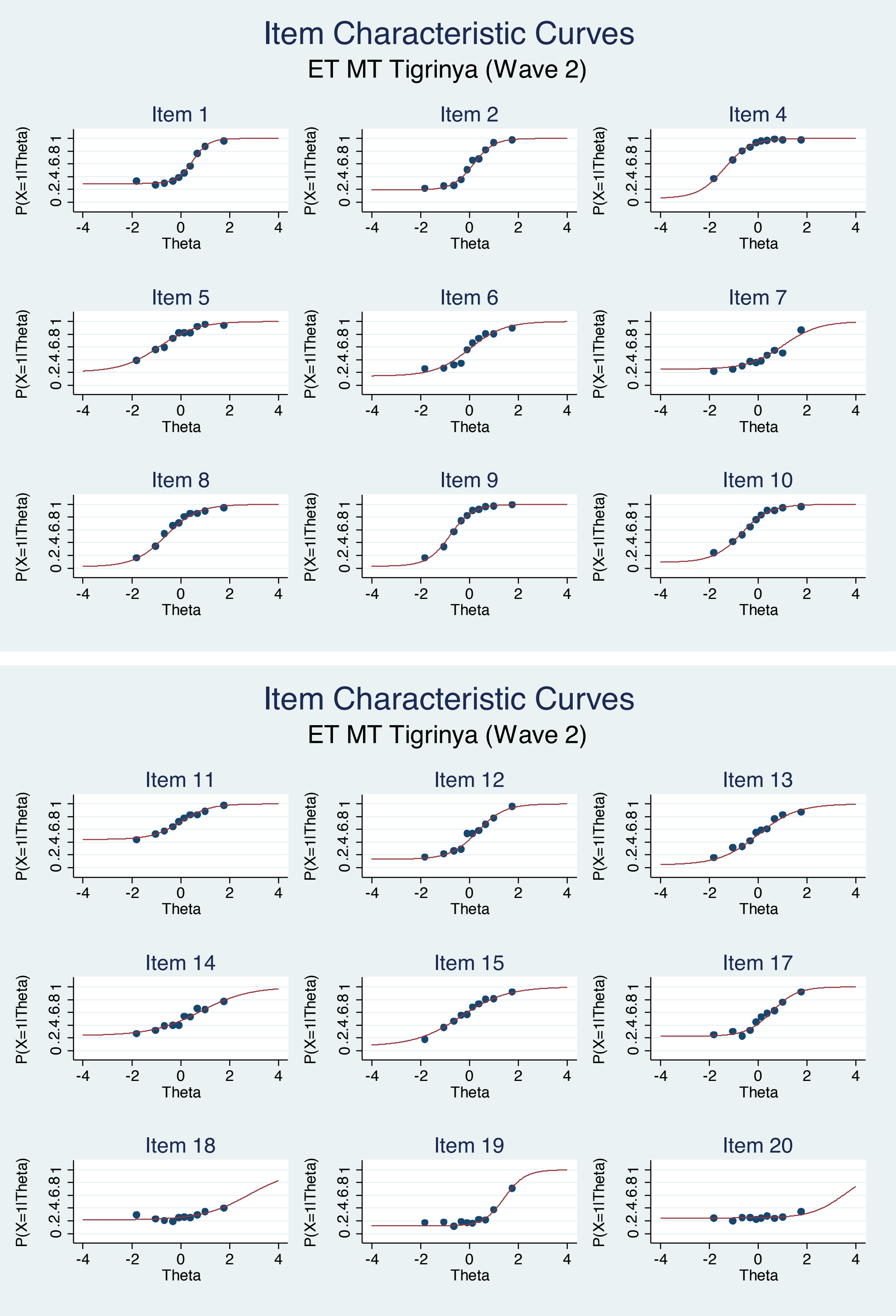

Item Characteristic Curves
ET MT Somali (Wave 2)
Item 20
Item 21
Item 22
Item 23
Item 25
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B13. *ICC for maths test for Tigrinya*



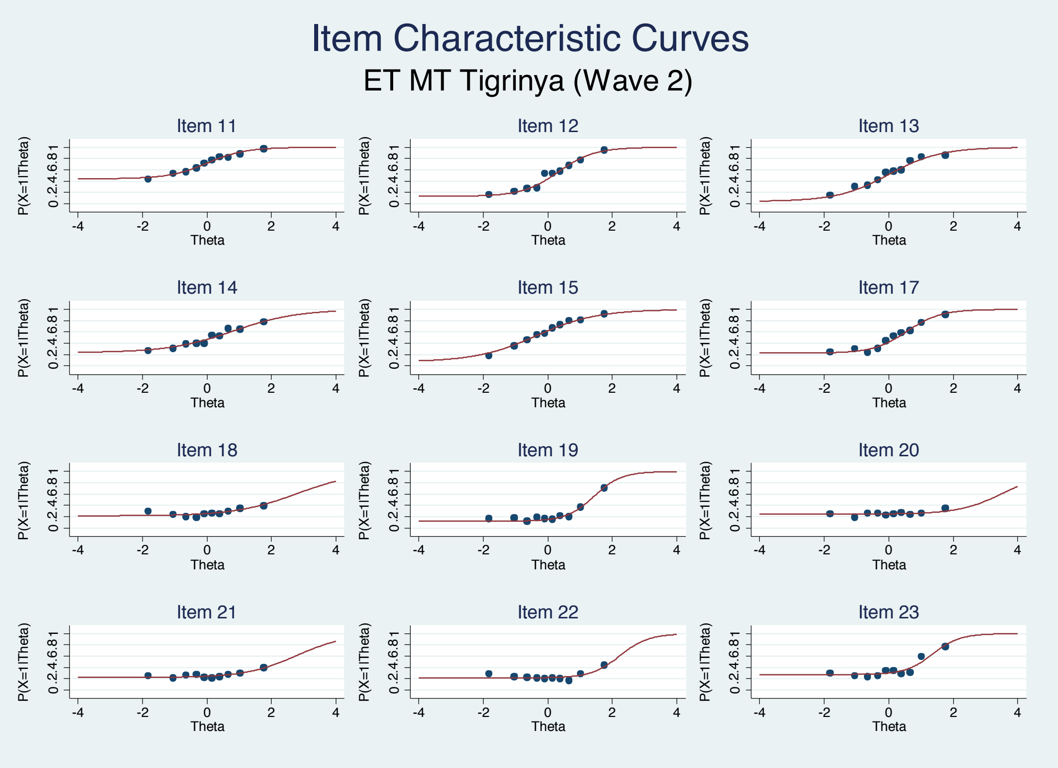
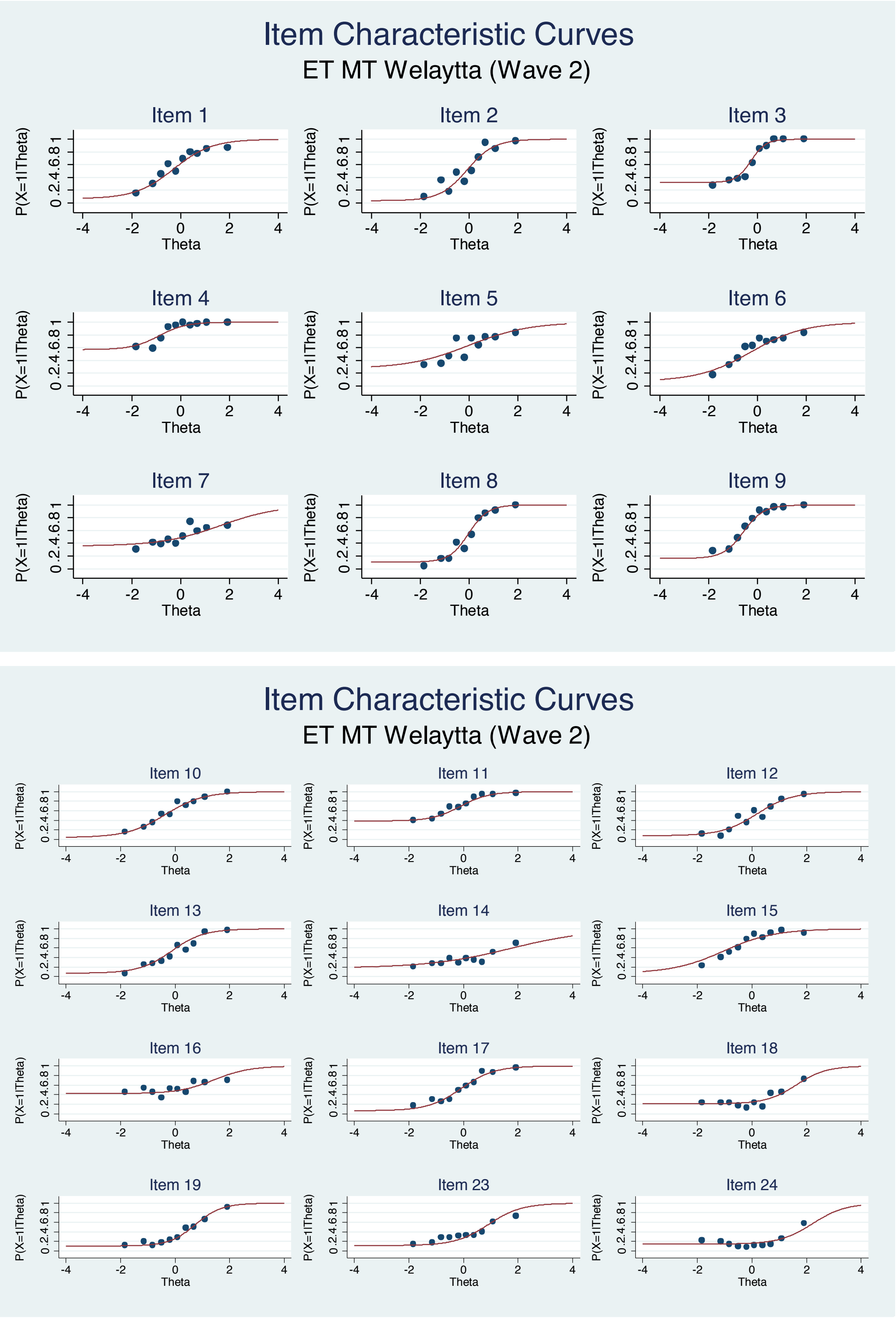
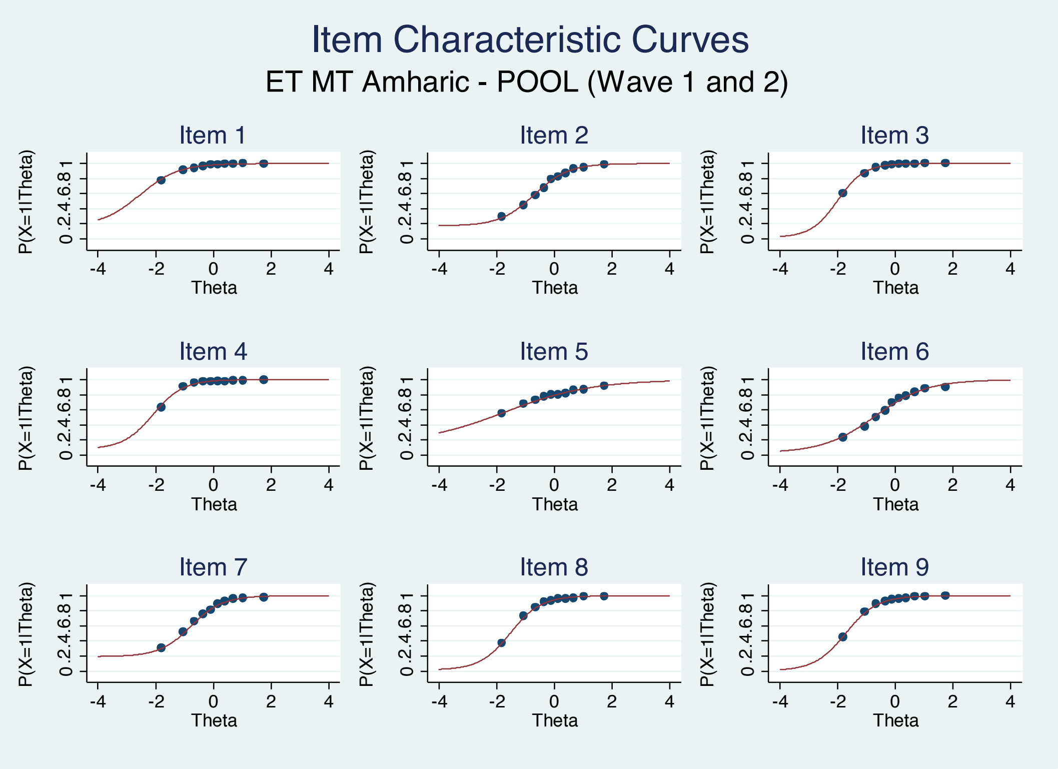

Item Characteristic Curves
ET MT Tigrinya (Wave 2)
Item 11
Item 12
Item 13
Item 14
Item 15
Item 17
Item 18
Item 19
Item 20
Item 21
Item 22
Item 23
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4
-2
0
2
4
Theta

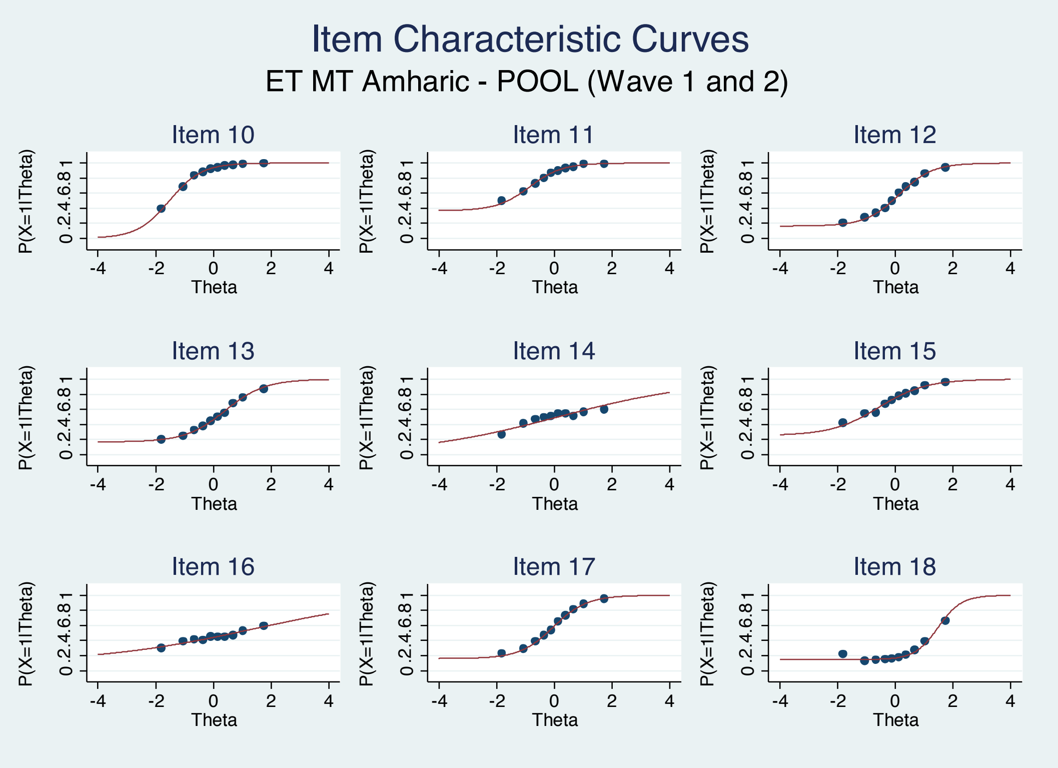
Figure B14. *ICC for maths test for Wolaytta*

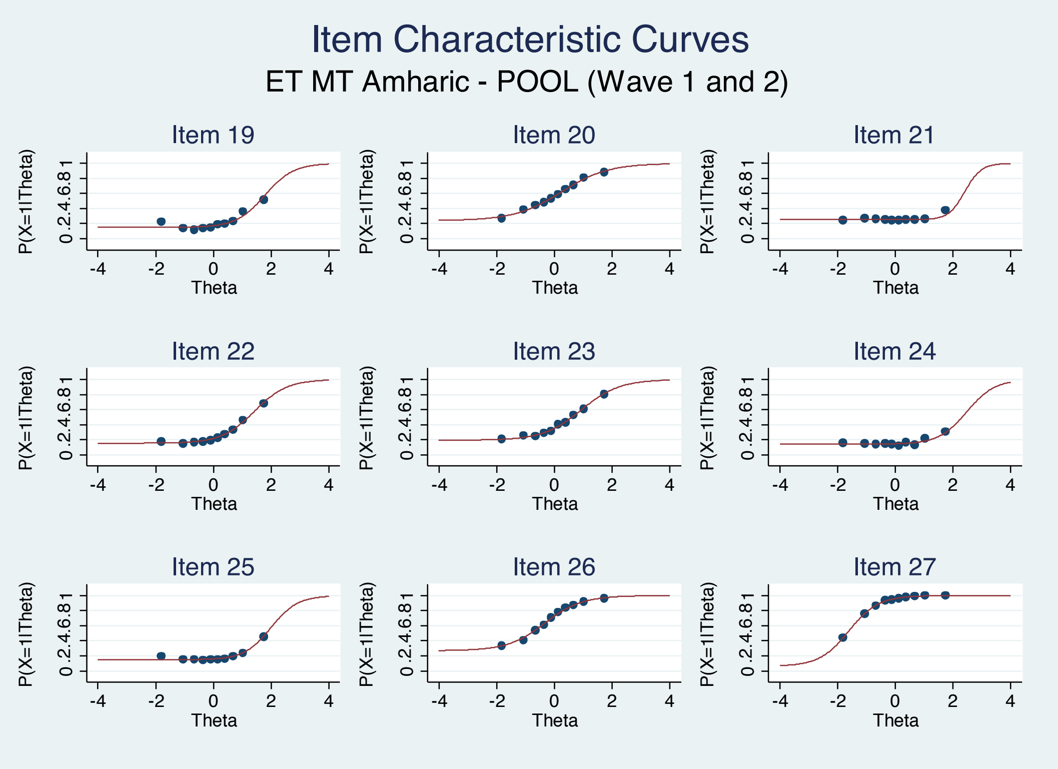


Pooled sample (Waves 1 and 2)

Figure B15. *ICC for maths test for Amharic*






Item Characteristic Curves
ET MT Amharic - POOL (Wave 1 and 2)
Item 19
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
Item 26
Item 27
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

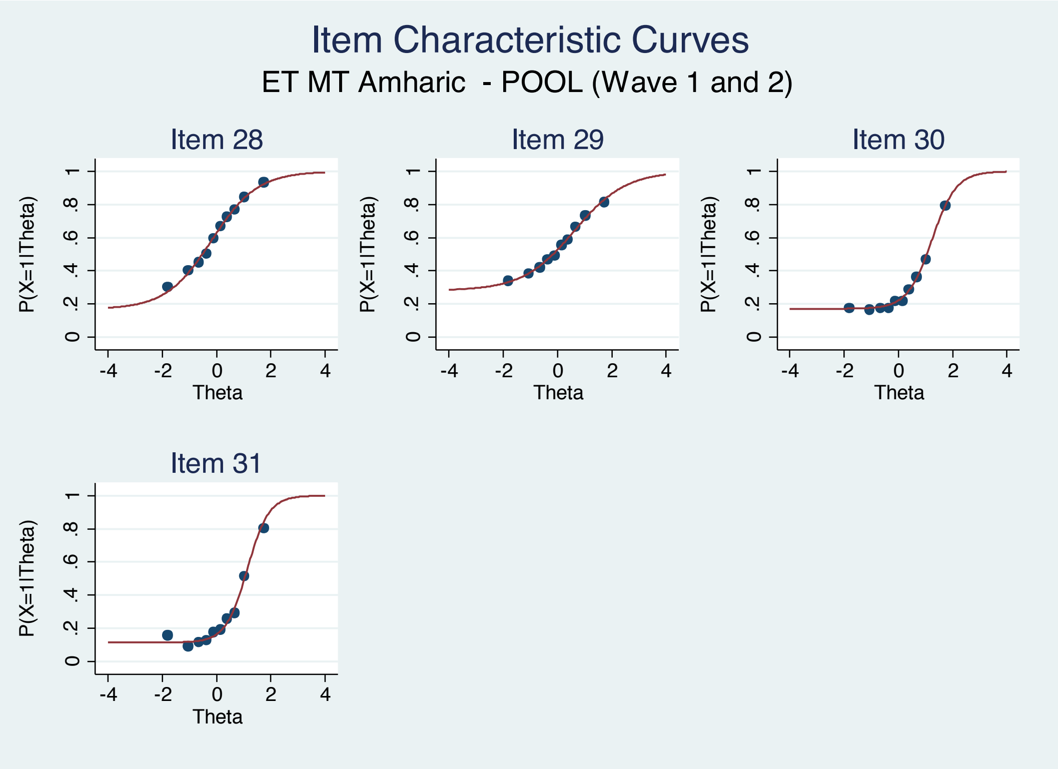
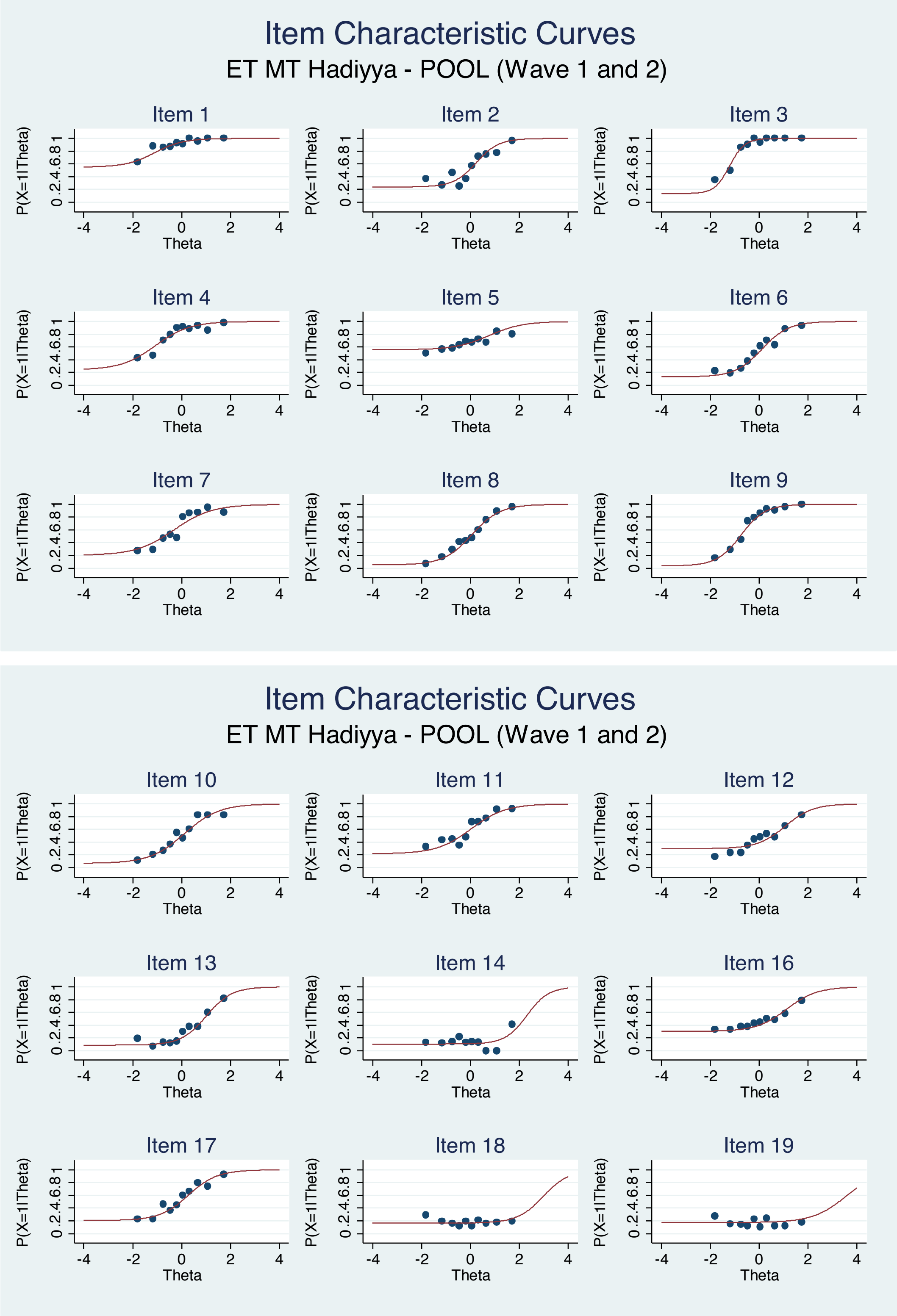

Item Characteristic Curves
ET MT Amharic - POOL (Wave 1 and 2)
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B16. *ICC for maths test for Hadiyya*



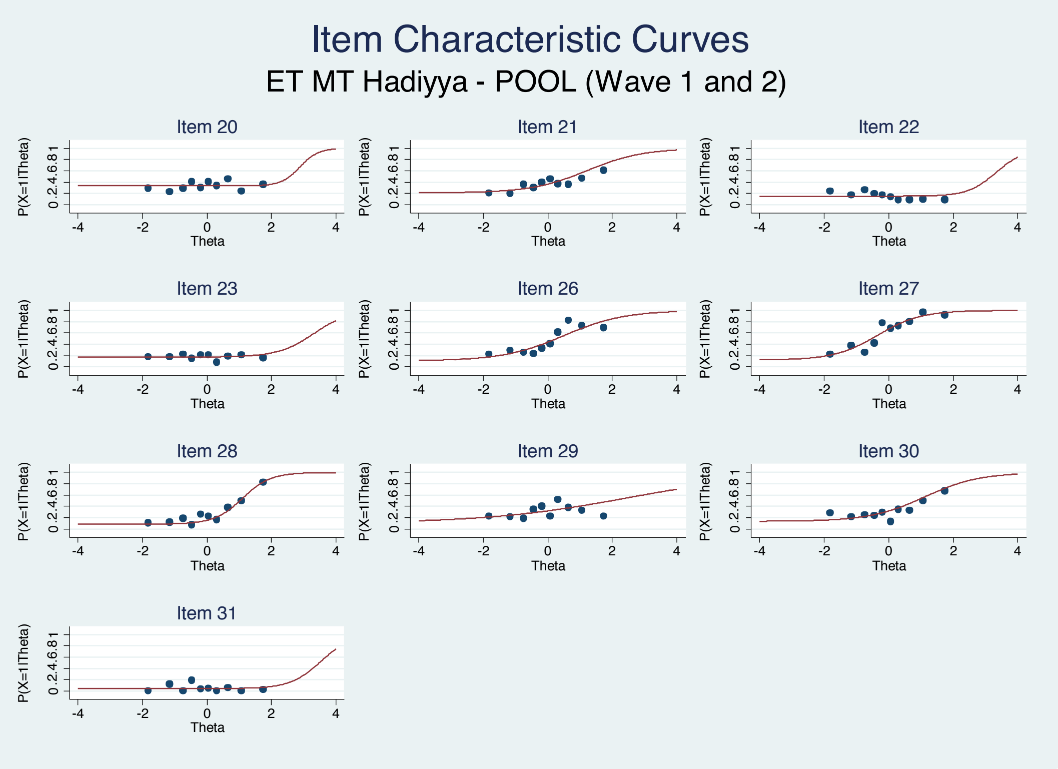
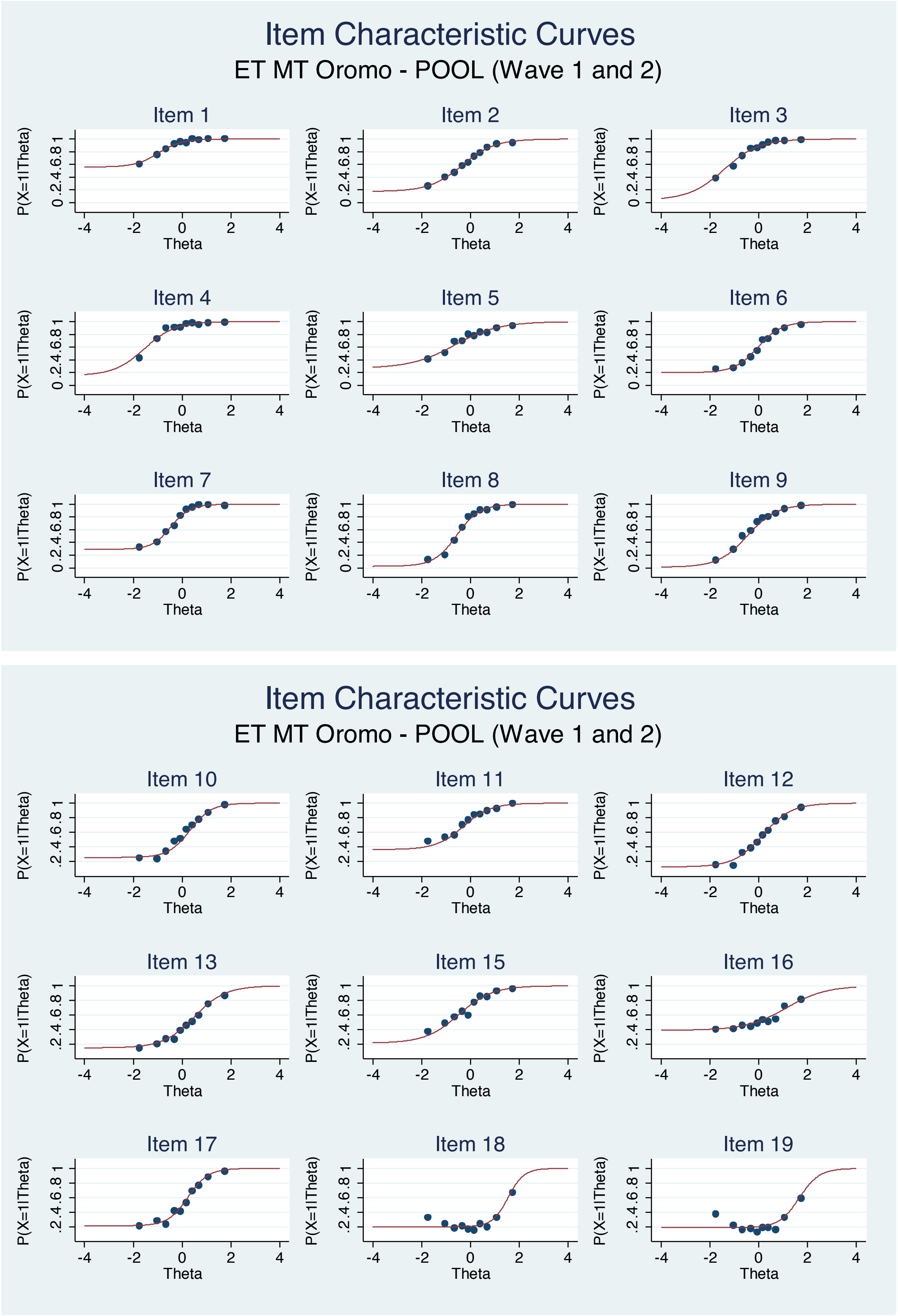

Item Characteristic Curves
ET MT Hadiyya - POOL (Wave 1 and 2)
Item 20
Item 21
Item 22
Item 23
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B17. *ICC for maths test for Oromo*



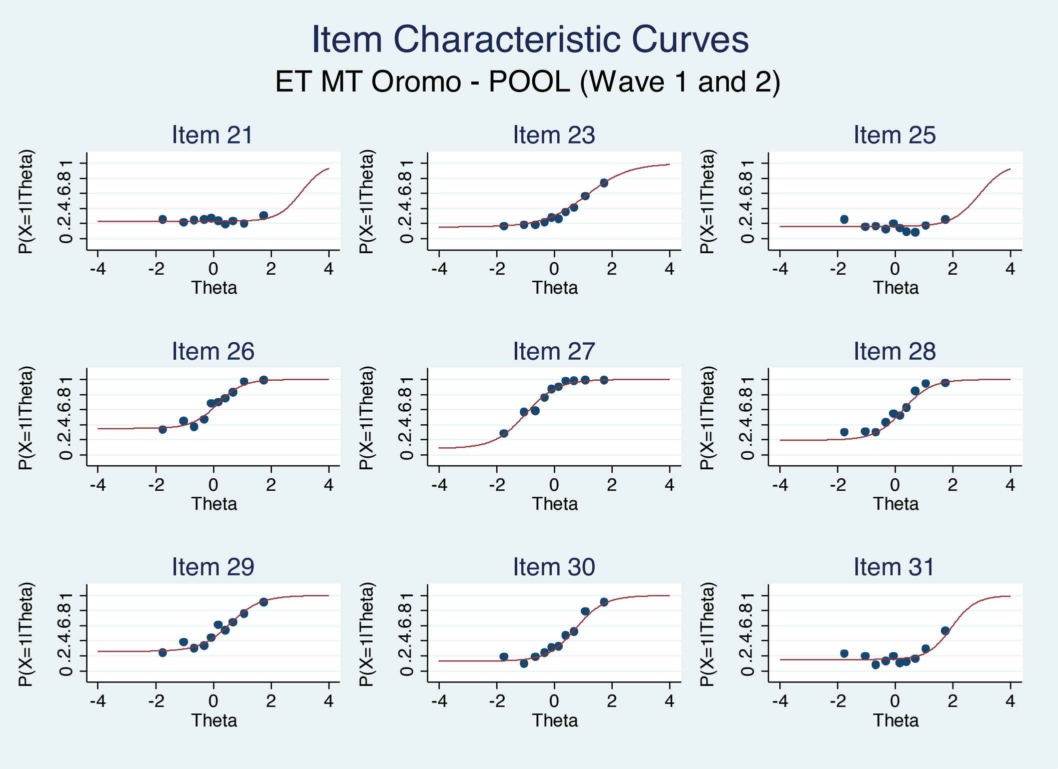
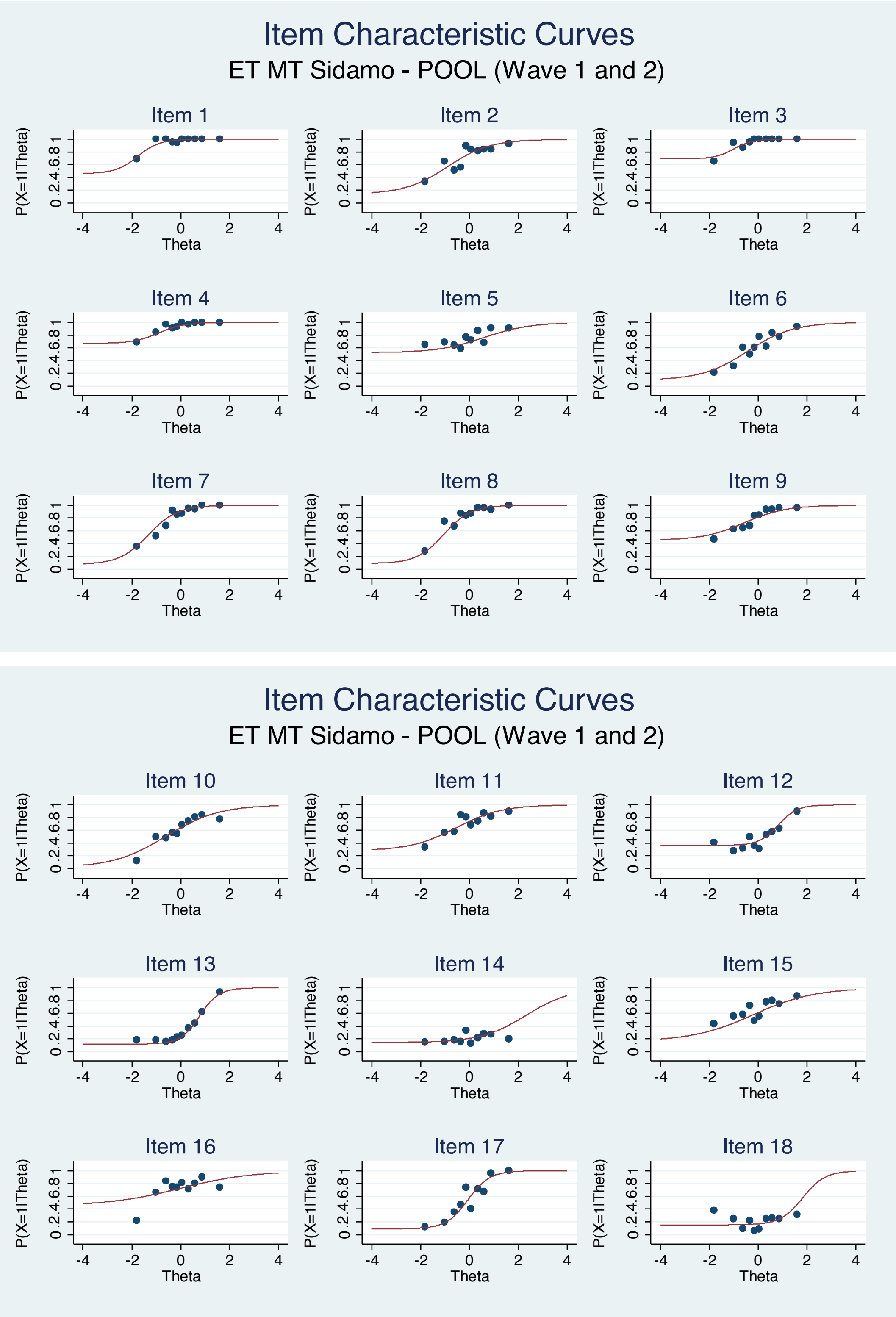

Item Characteristic Curves
ET MT Oromo - POOL (Wave 1 and 2)
Item 21
Item 23
Item 25
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B18. *ICC for maths test for Sidamo*



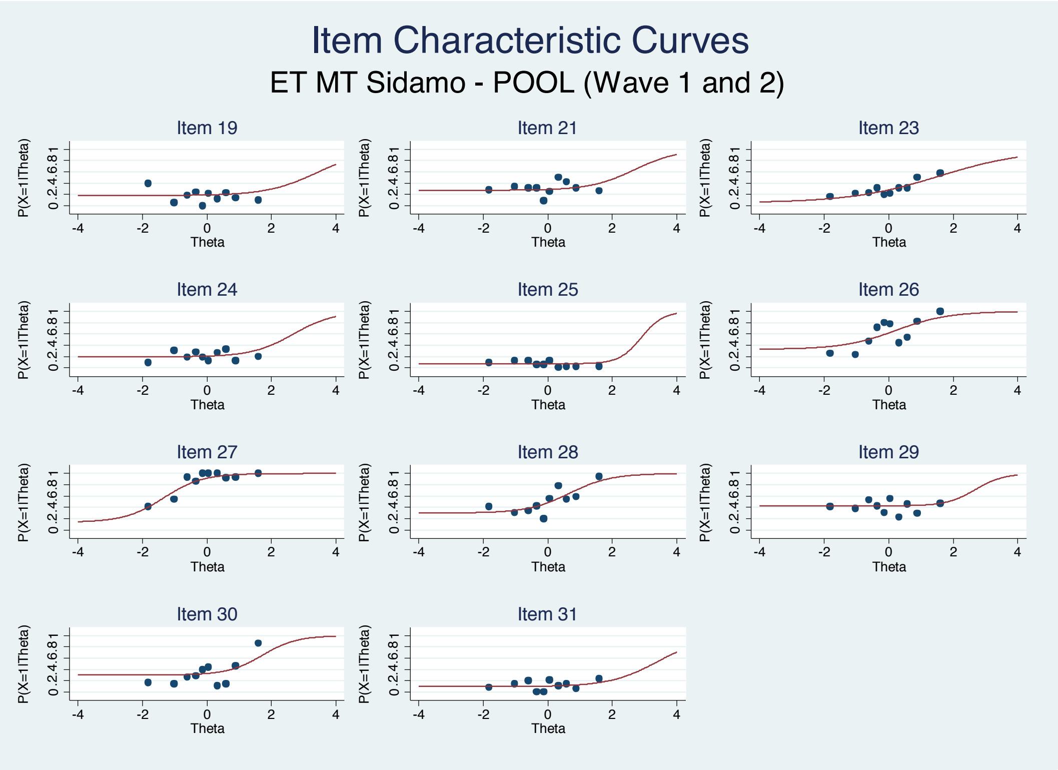
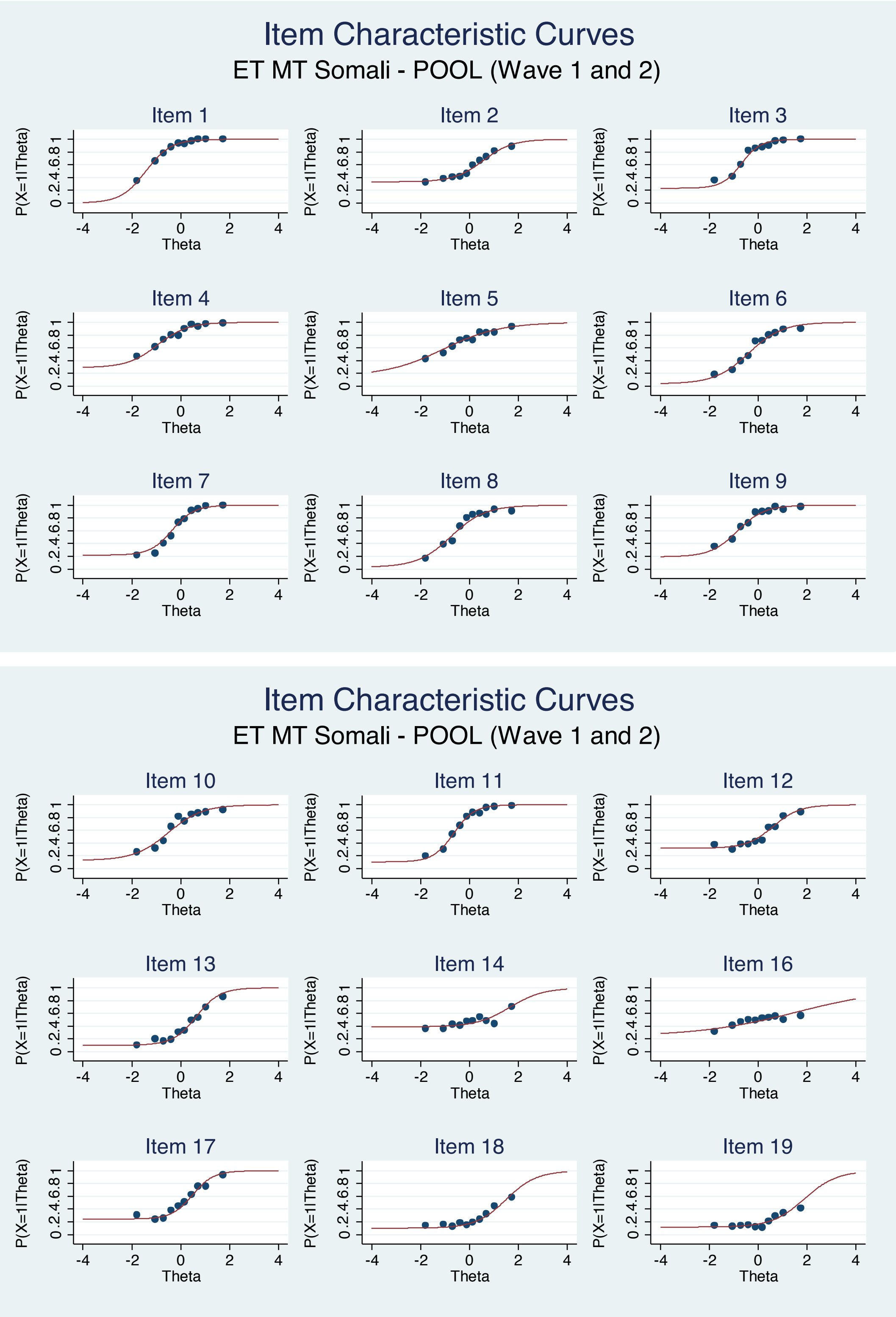

Item Characteristic Curves
ET MT Sidamo - POOL (Wave 1 and 2)
Item 19
Item 21
Item 23
Item 24
Item 25
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B19. *ICC for maths test for Somali*



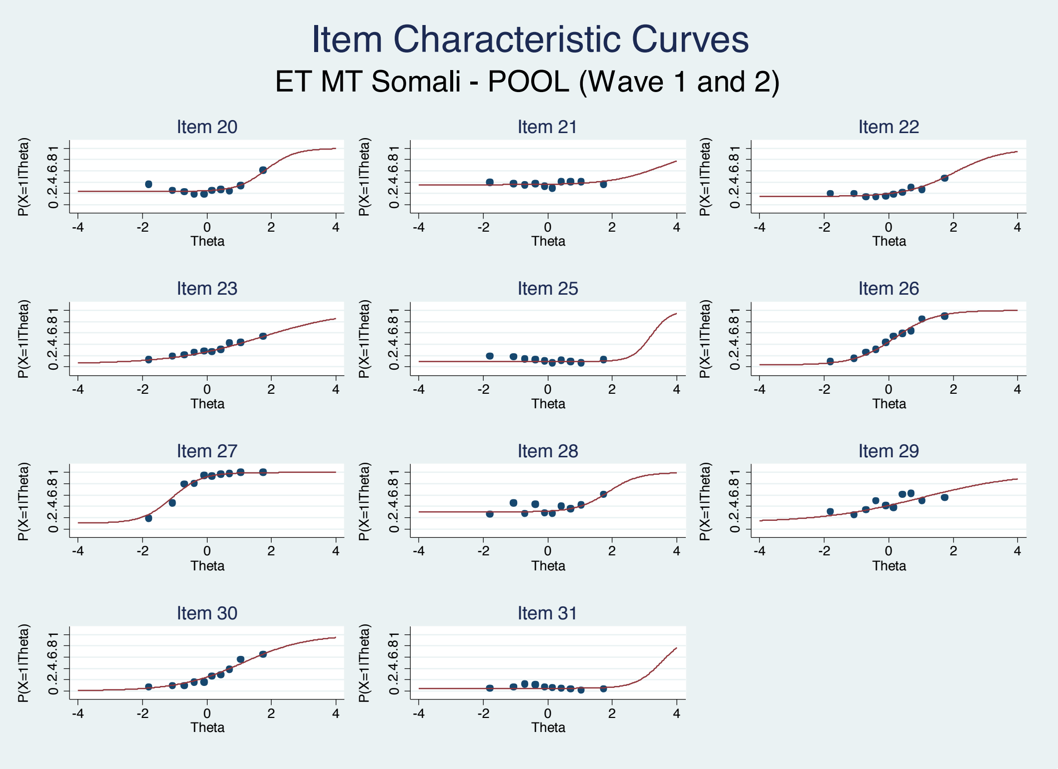
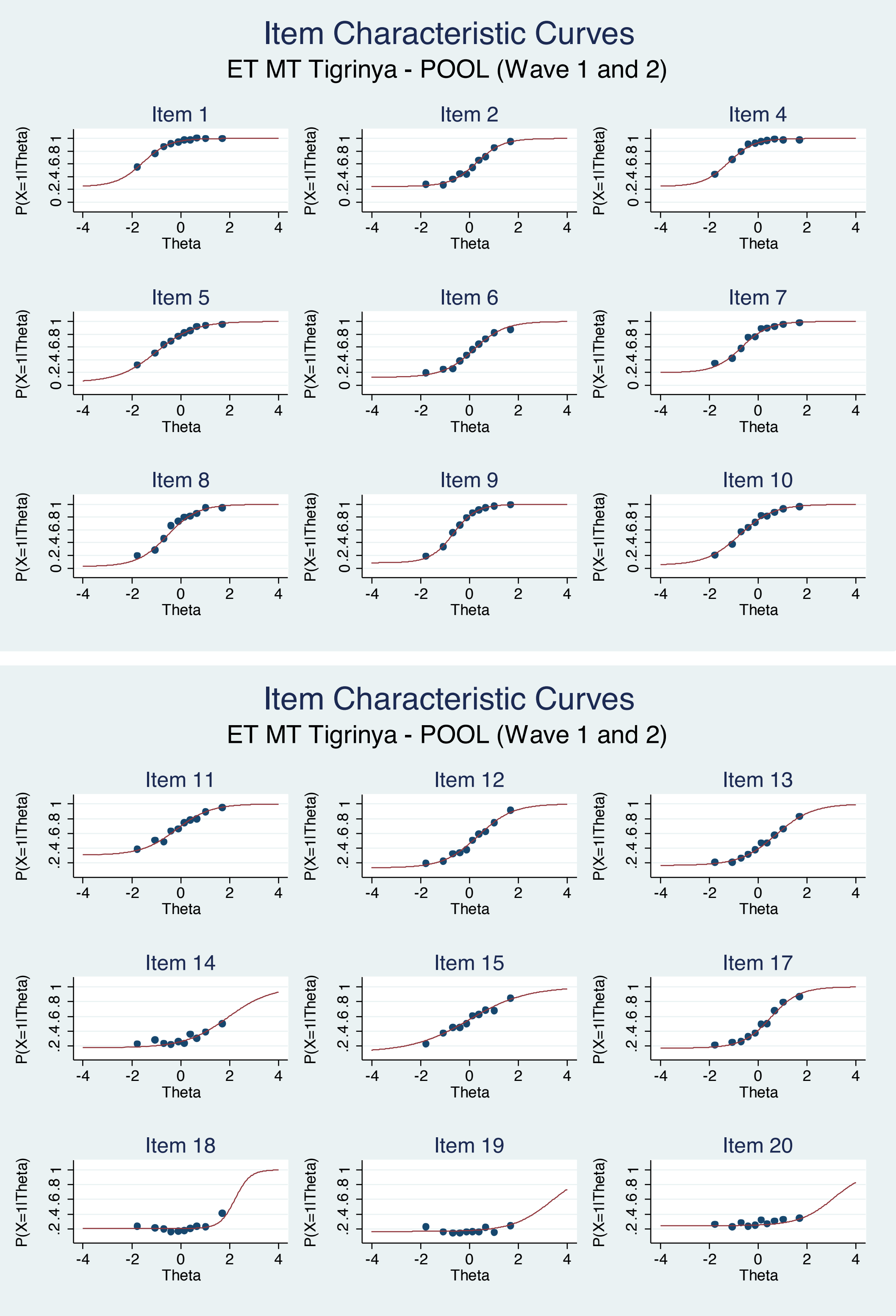

Item Characteristic Curves
ET MT Somali - POOL (Wave 1 and 2)
Item 20
Item 21
Item 22
Item 23
Item 25
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B20. *ICC for maths test for Tigrinya*



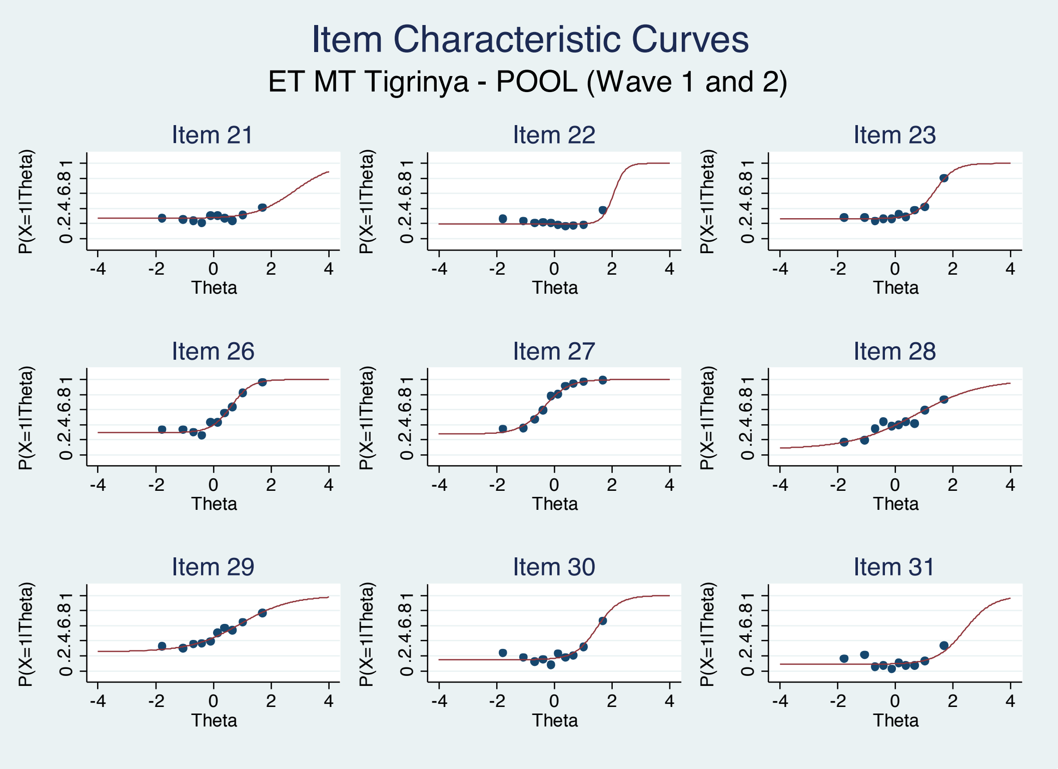
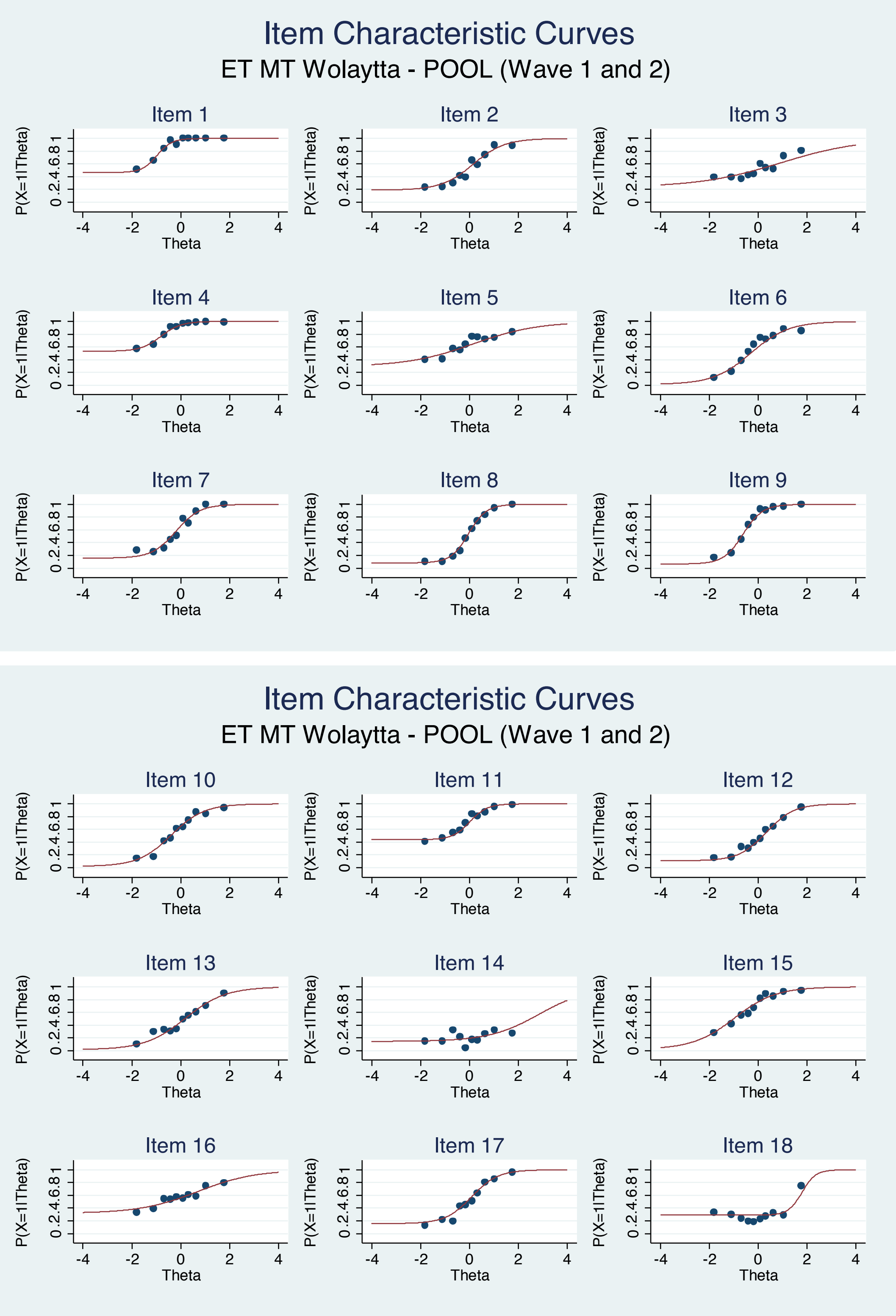
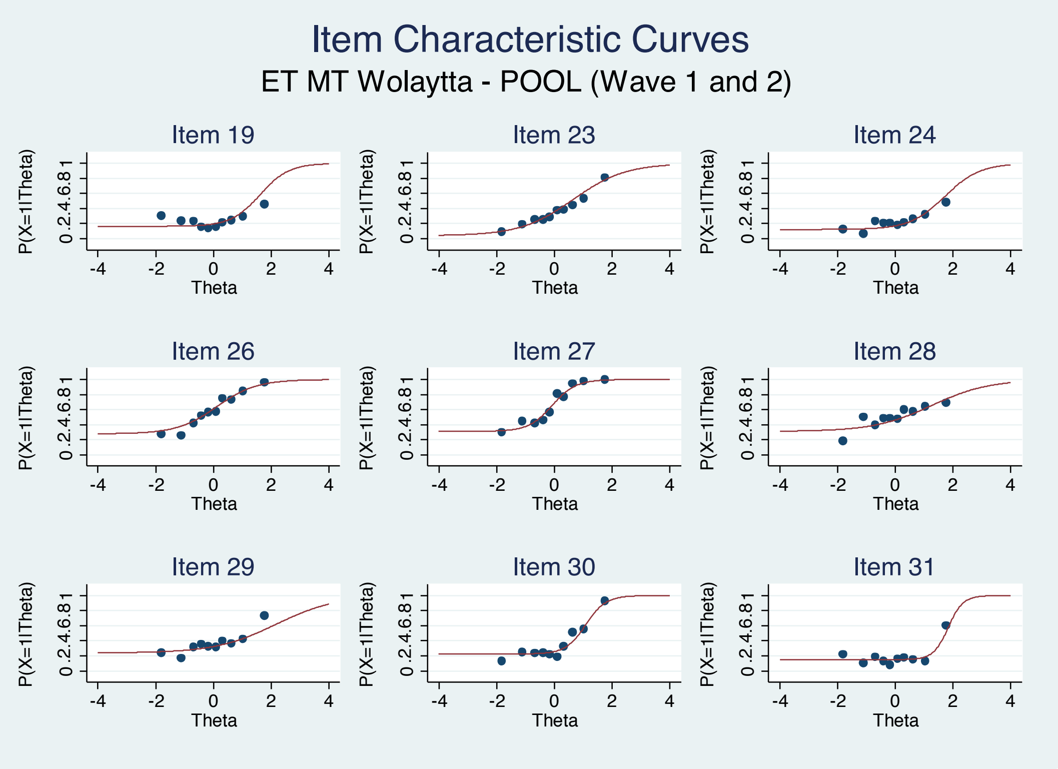

Item Characteristic Curves
ET MT Tigrinya - POOL (Wave 1 and 2)
Item 21
Item 22
Item 23
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B21. *ICC for maths test for Wolaytta*

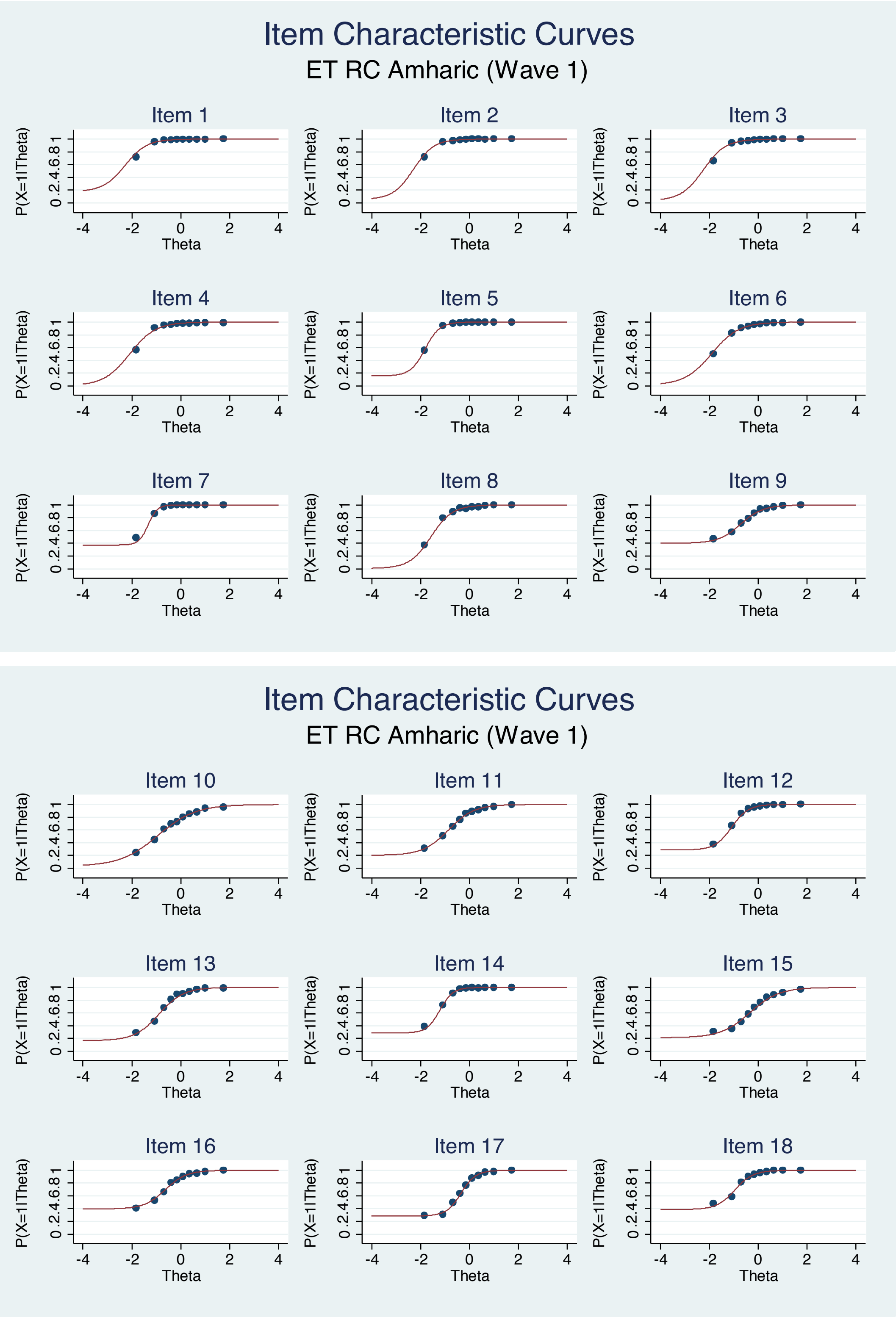



Item Characteristic Curves
ET MT Wolaytta - POOL (Wave 1 and 2)
Item 19
Item 23
Item 24
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Reading comprehension

Wave 1

Figure B22. *ICC for reading comprehension test for Amharic*



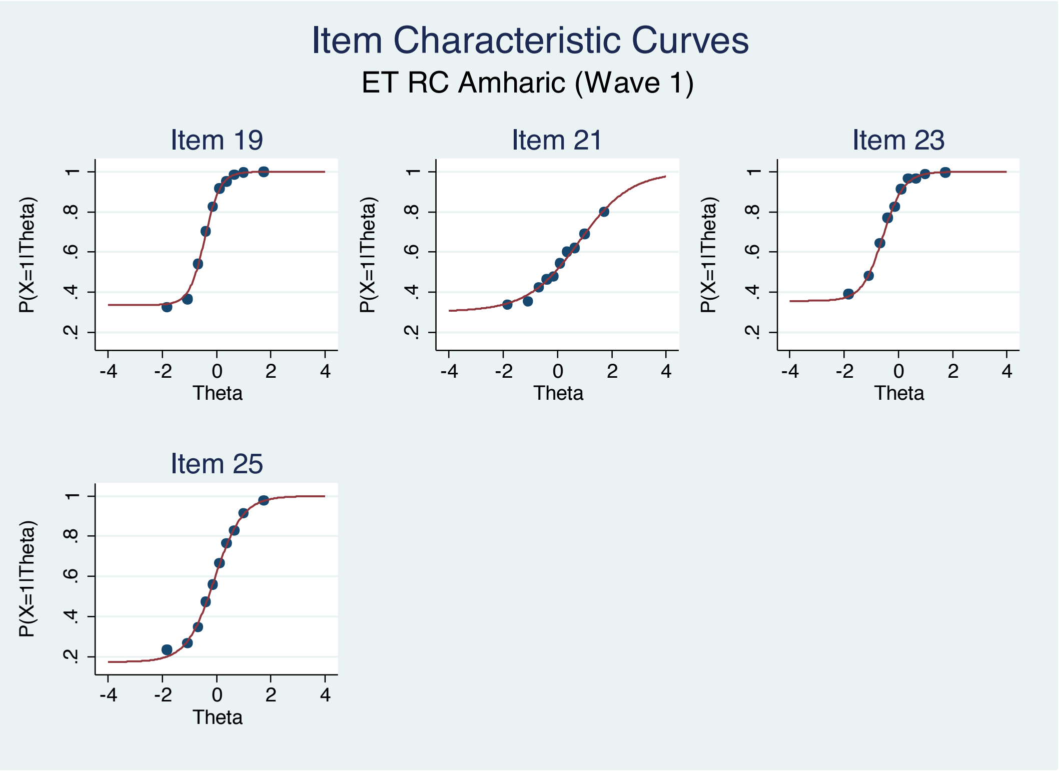
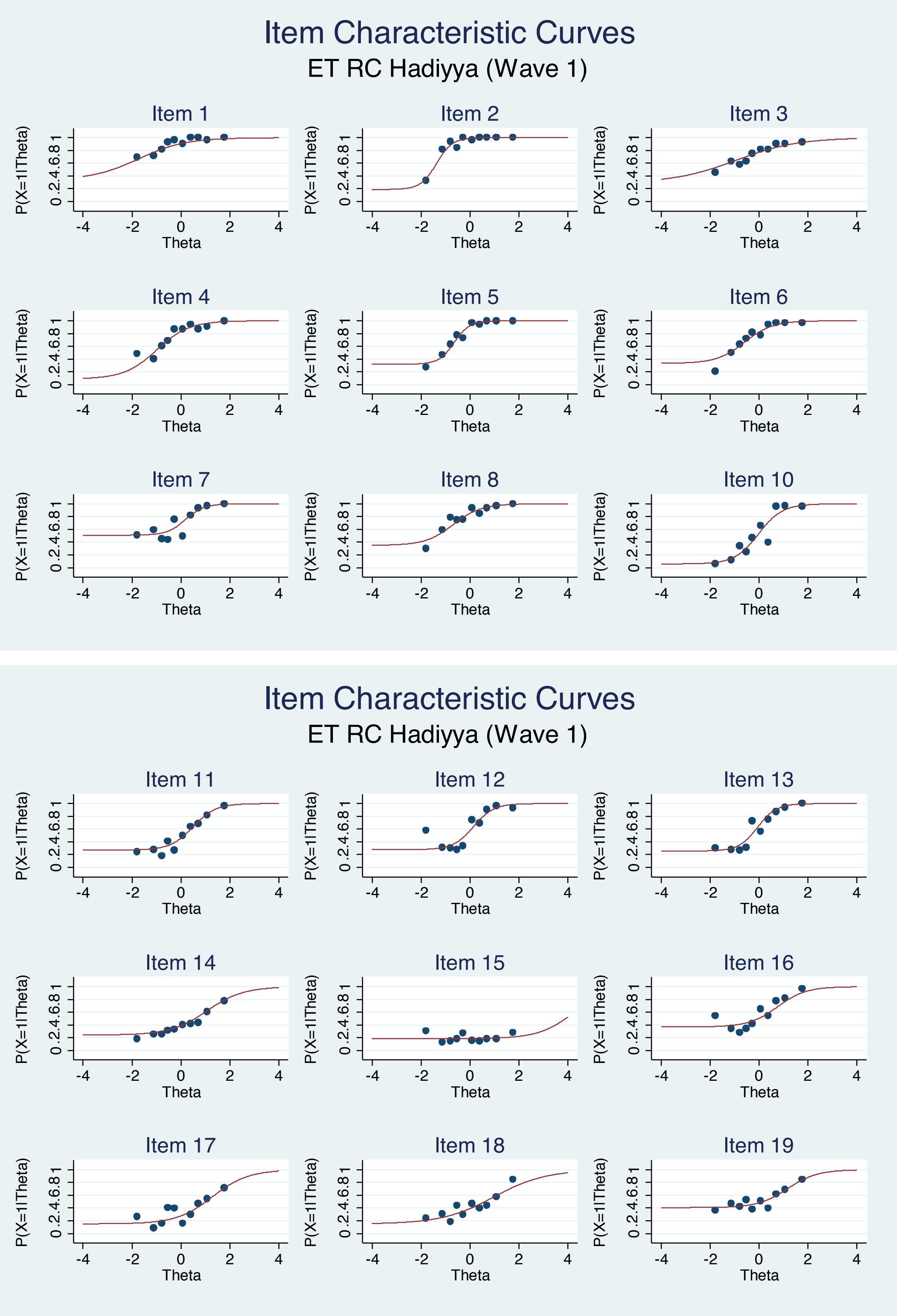

Item Characteristic Curves
ET RC Amharic (Wave 1)
Item 19
Item 21
Item 23
Item 25
P(X=1|Theta)
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B23. *ICC for reading comprehension test for Hadiyya*



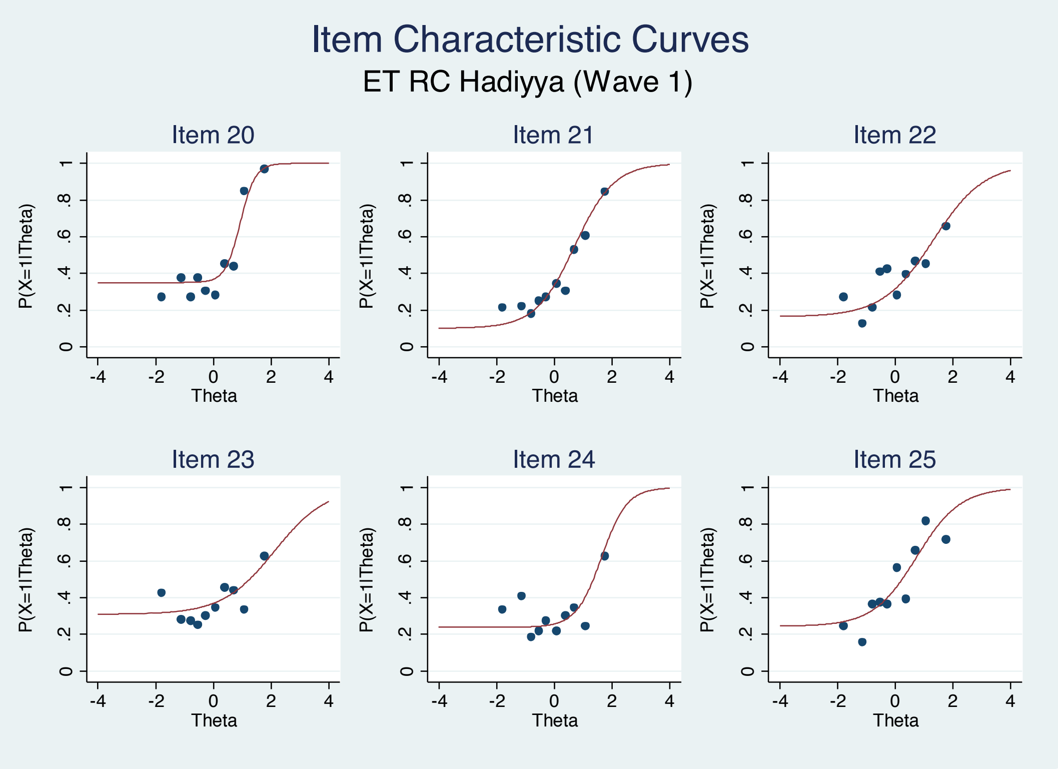
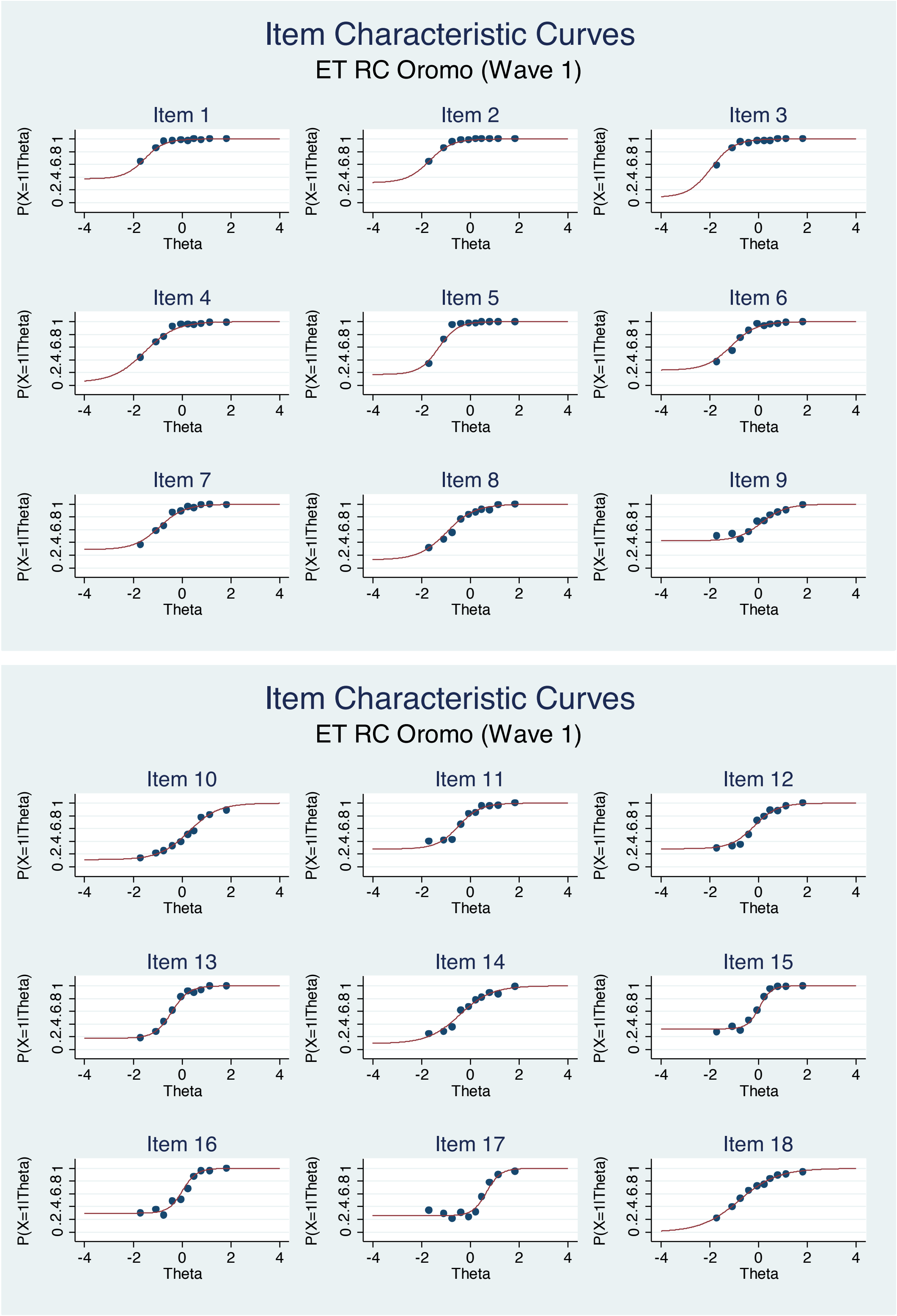

Item Characteristic Curves
ET RC Hadiyya (Wave 1)
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B24. *ICC for reading comprehension test for Oromo*



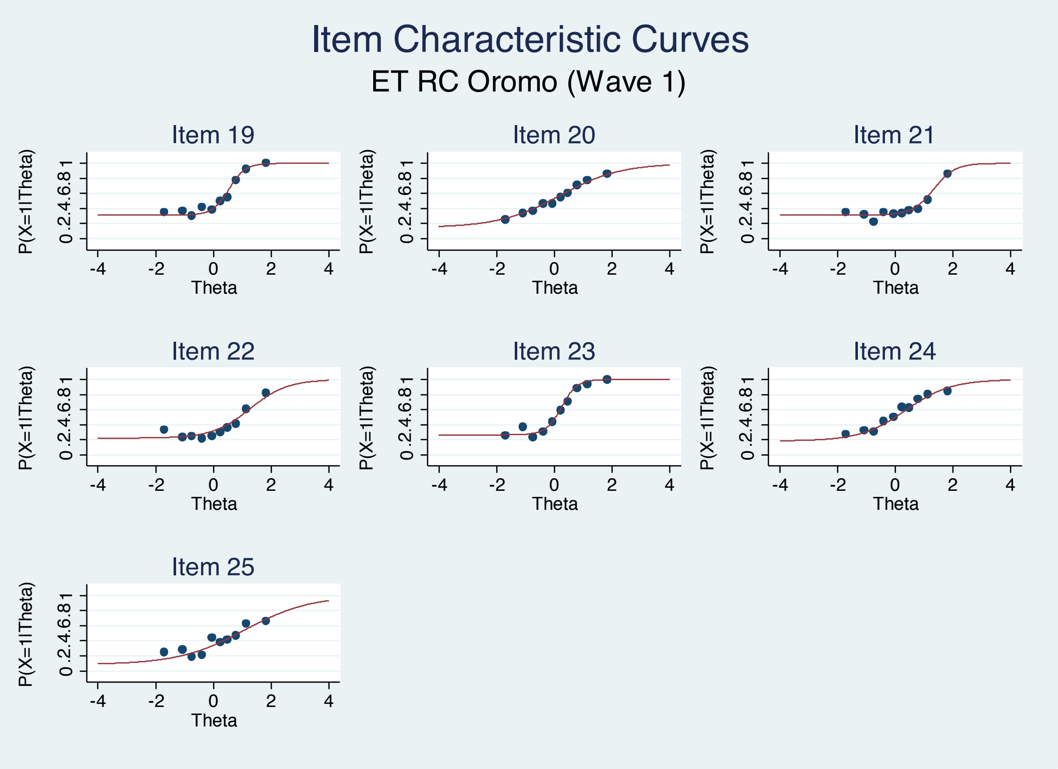
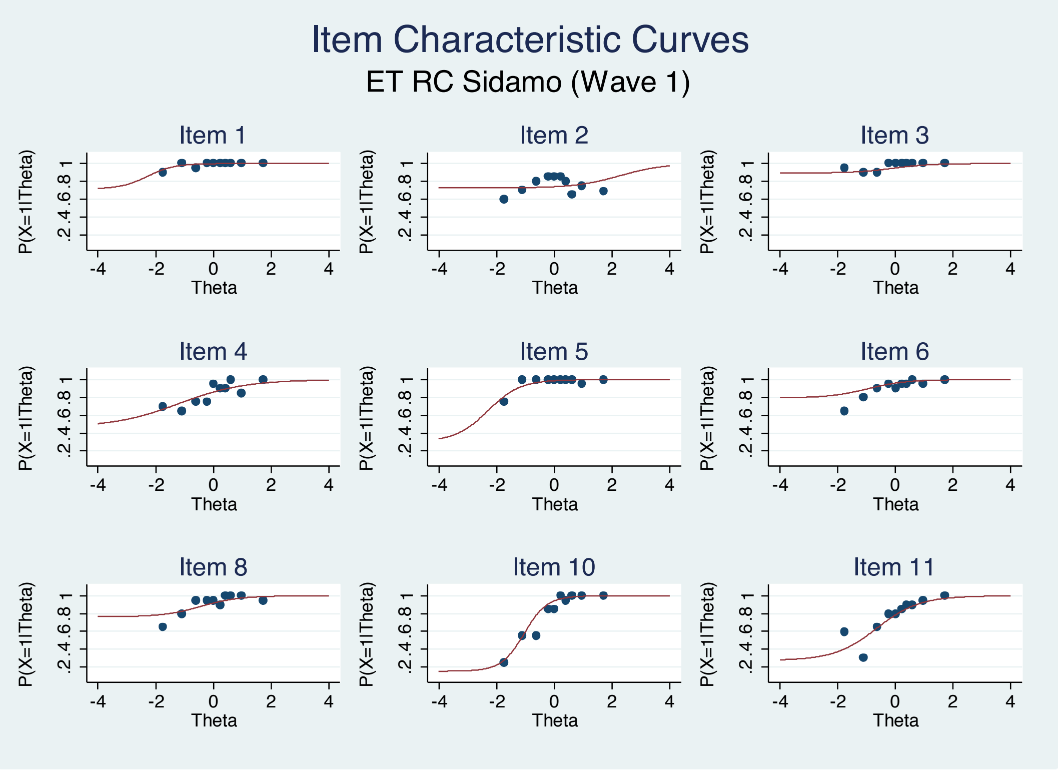
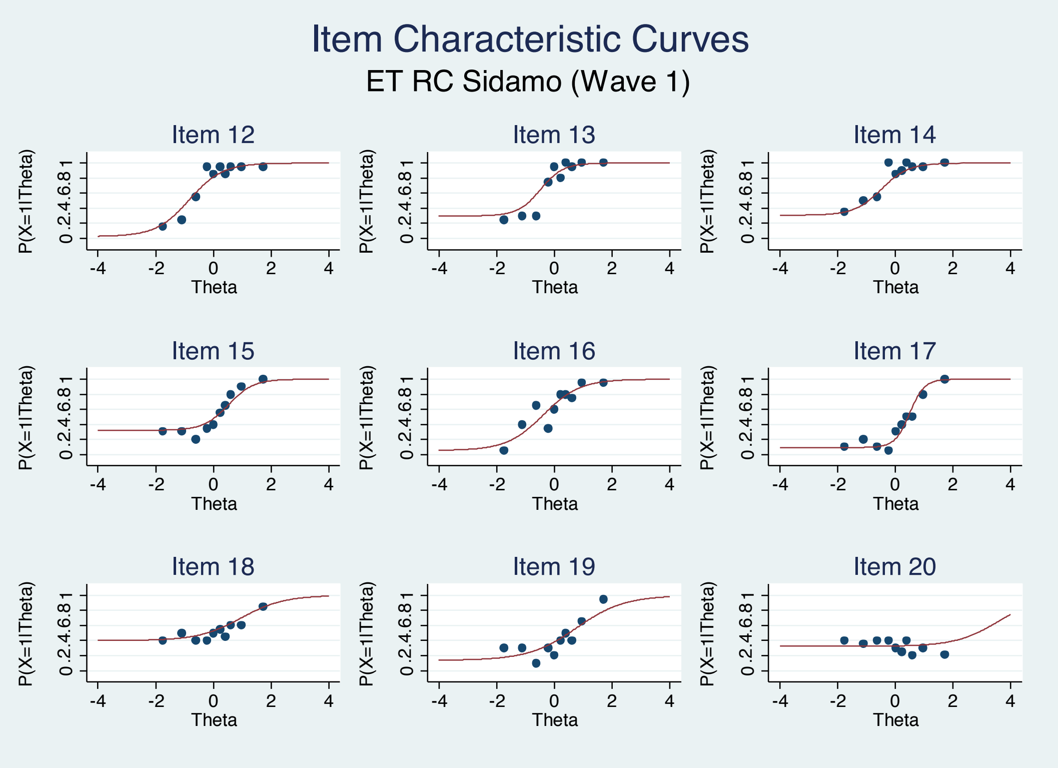

Item Characteristic Curves
ET RC Oromo (Wave 1)
Item 19
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B25. *ICC for reading comprehension test for Sidamo*





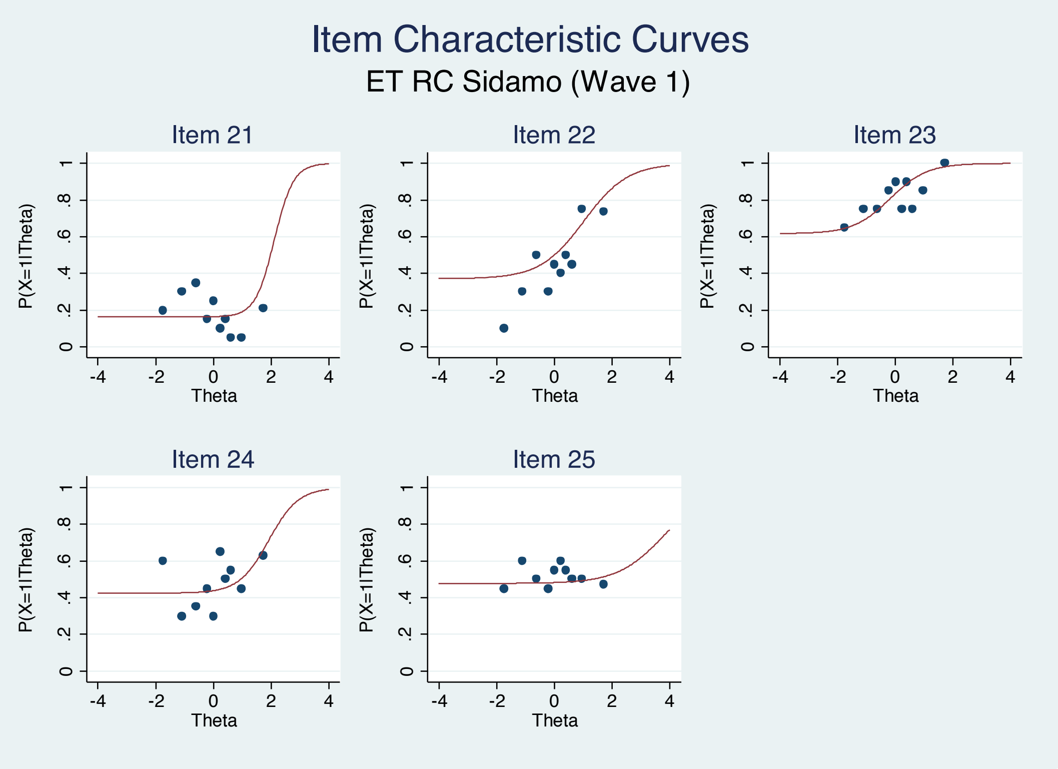
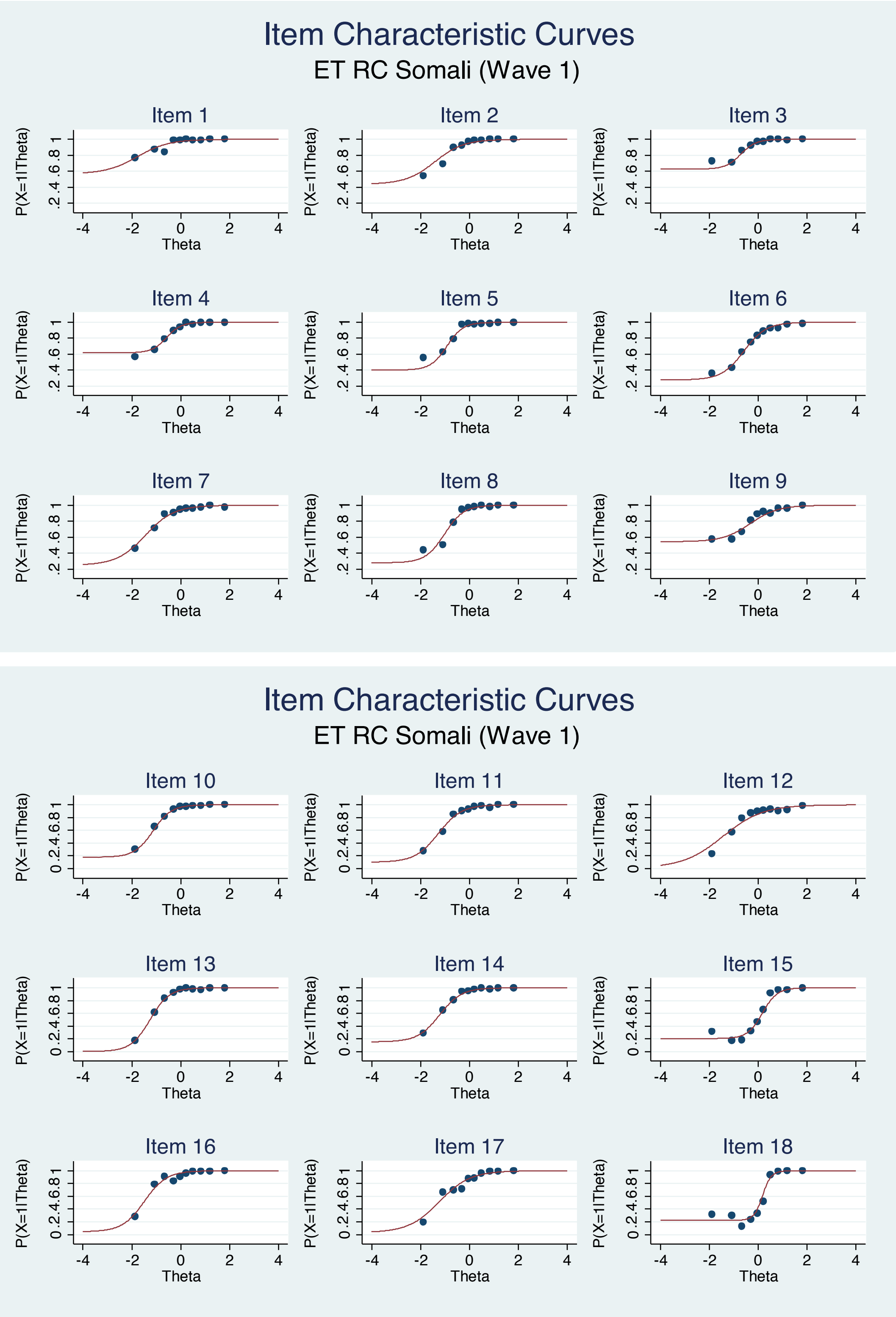

Item Characteristic Curves
ET RC Sidamo (Wave 1)
Item 21
Item 22
Item 23
Item 24
Item 25
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B26. *ICC for reading comprehension test for Somali*



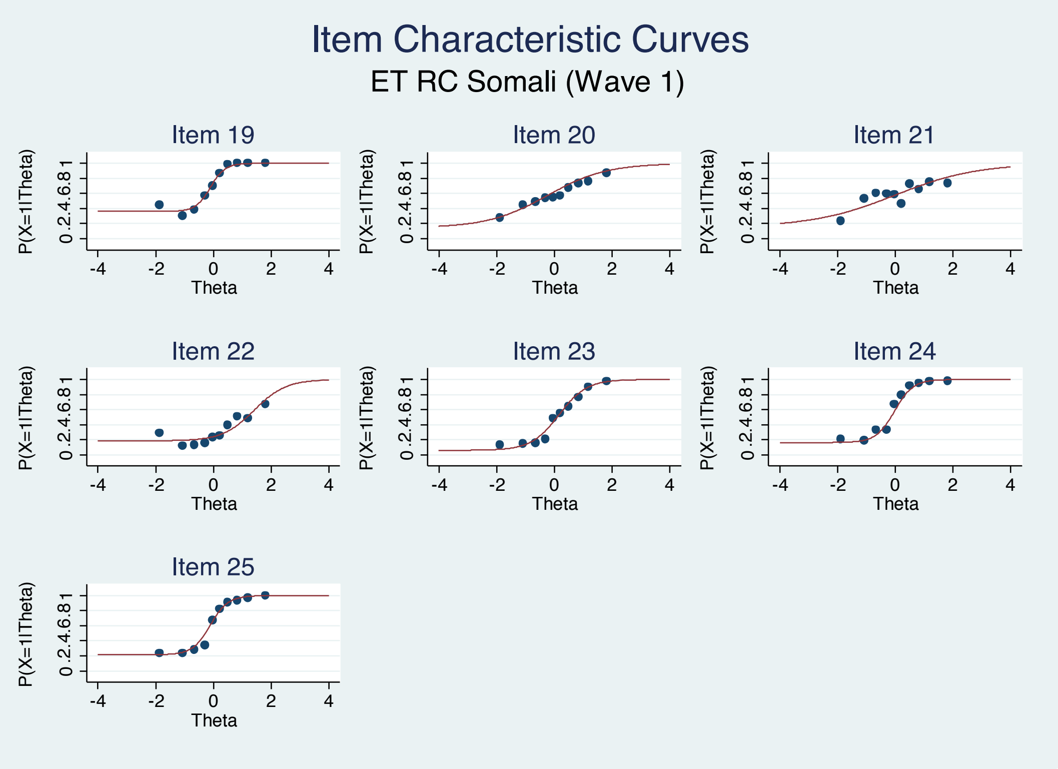
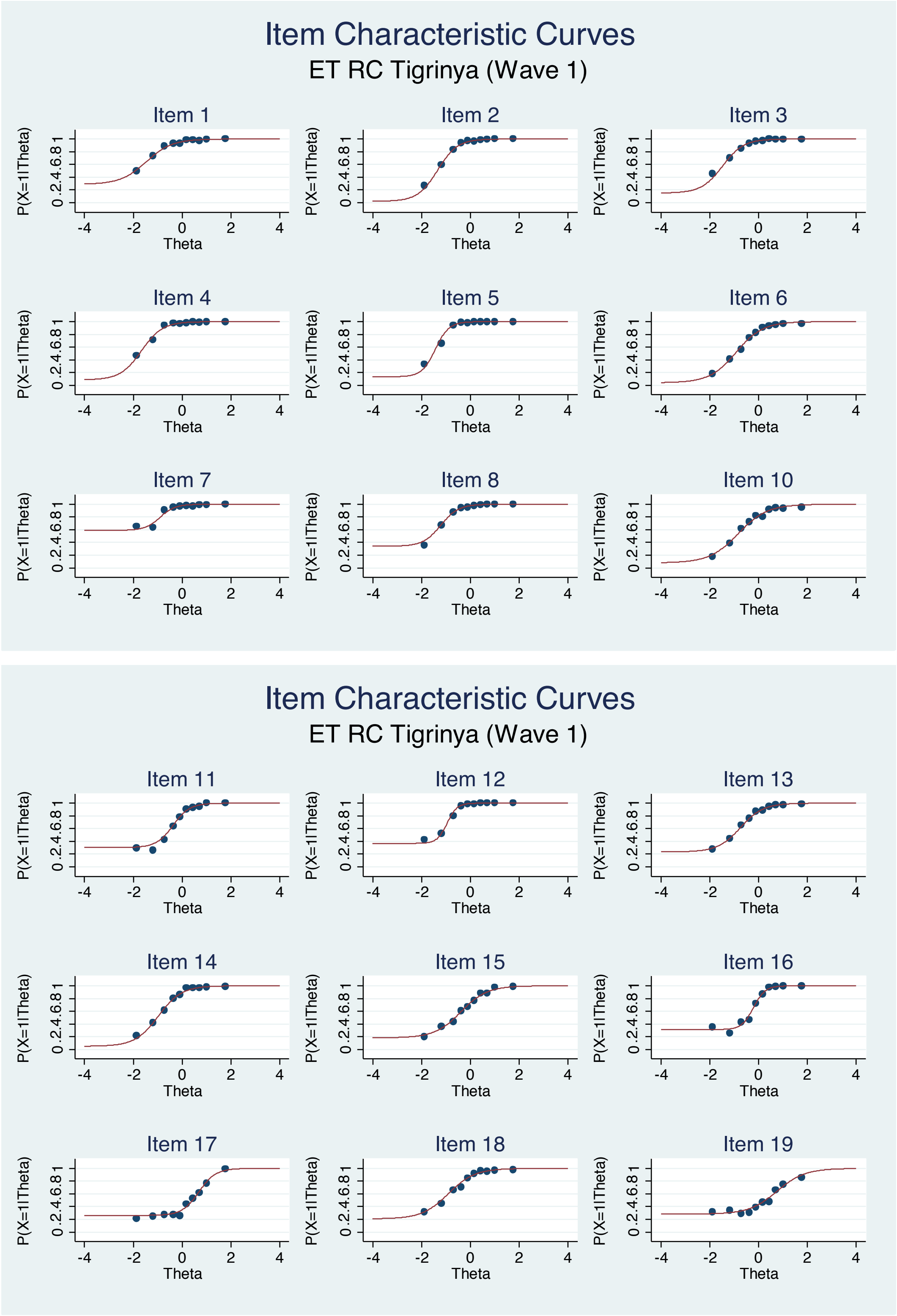

Item Characteristic Curves
ET RC Somali (Wave 1)
Item 19
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4
-2
0
2
4
Theta

Figure B27. *ICC for reading comprehension test for Tigrinya*



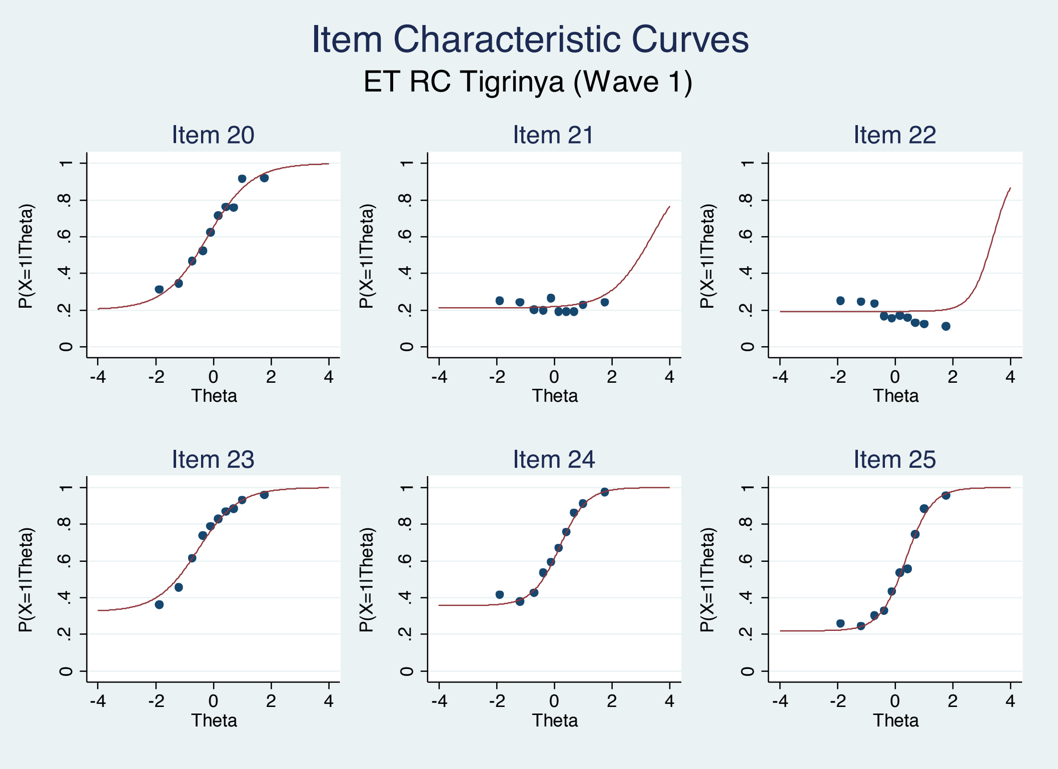
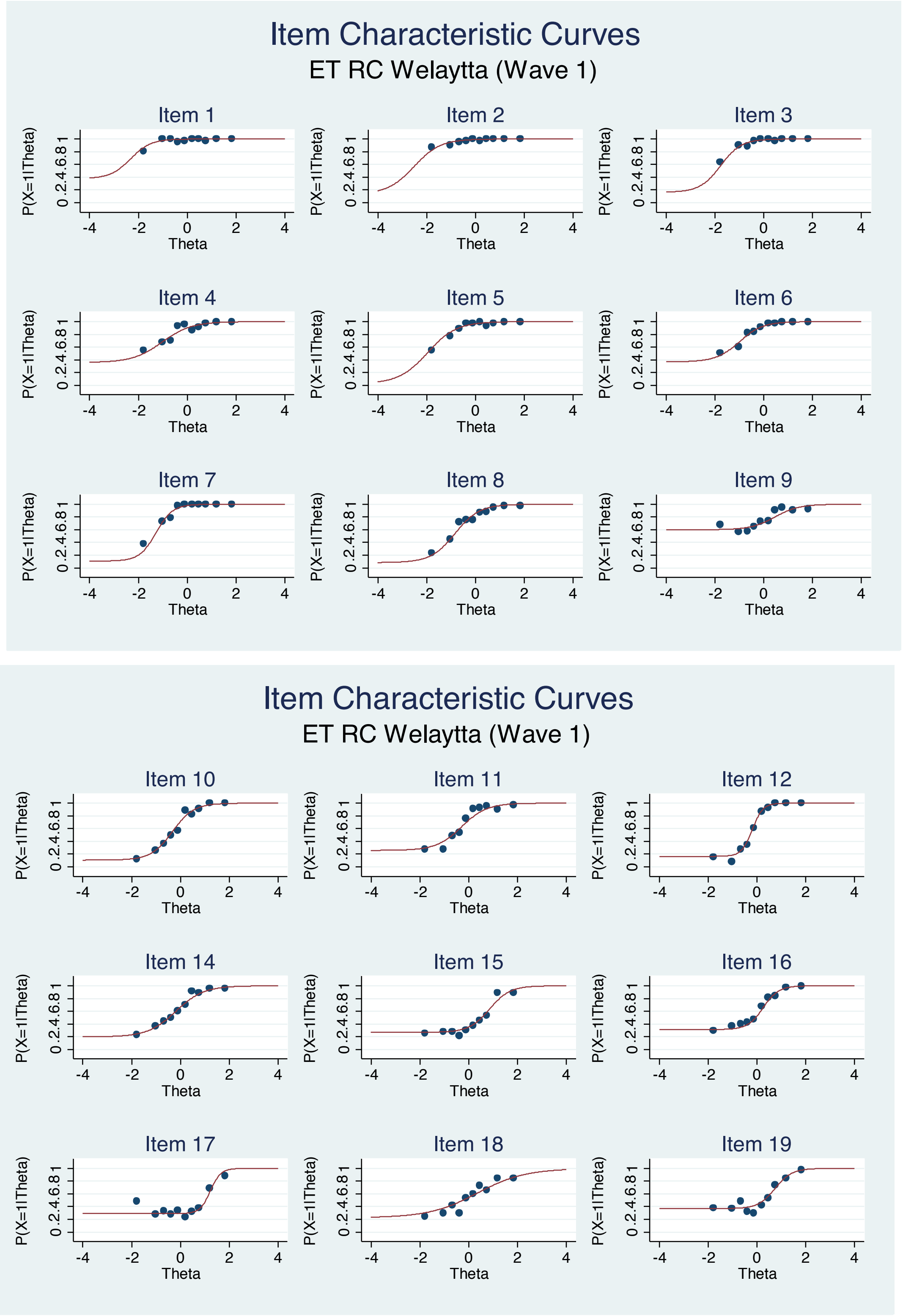
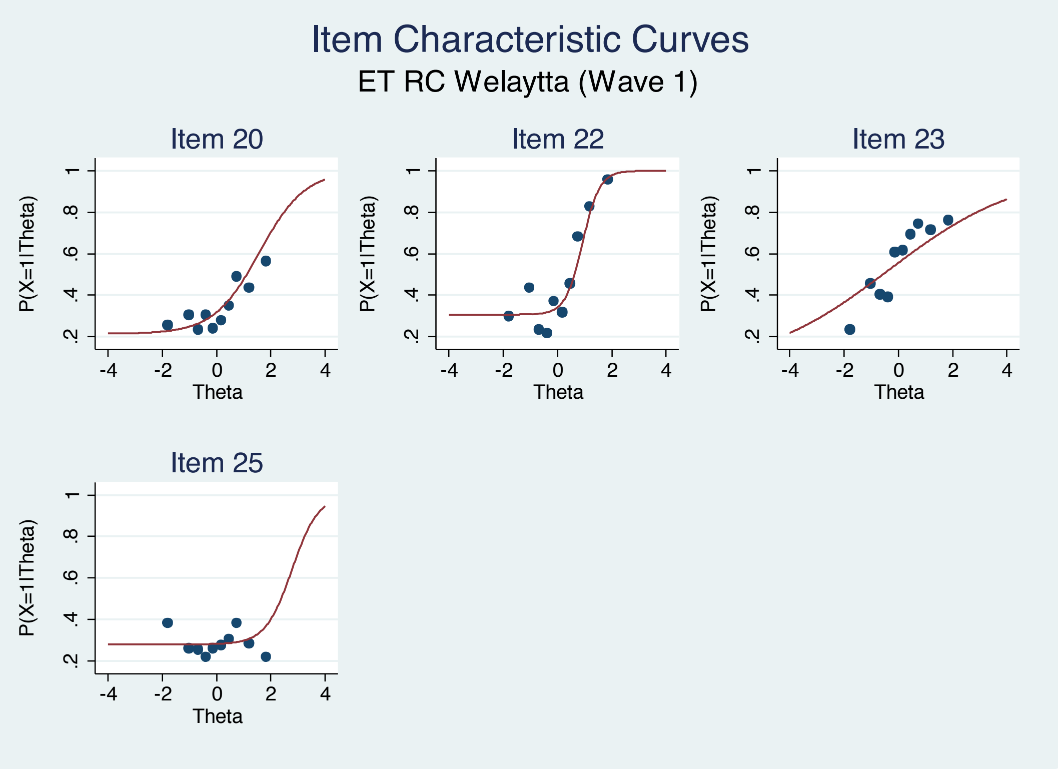

Item Characteristic Curves
ET RC Tigrinya (Wave 1)
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B28. *ICC for reading comprehension test for Wolaytta*




Item Characteristic Curves
ET RC Welaytta (Wave 1)
Item 20
Item 22
Item 23
Item 25
P(X=1|Theta)
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Wave 2

Figure B29. *ICC for reading comprehension test for Amharic*

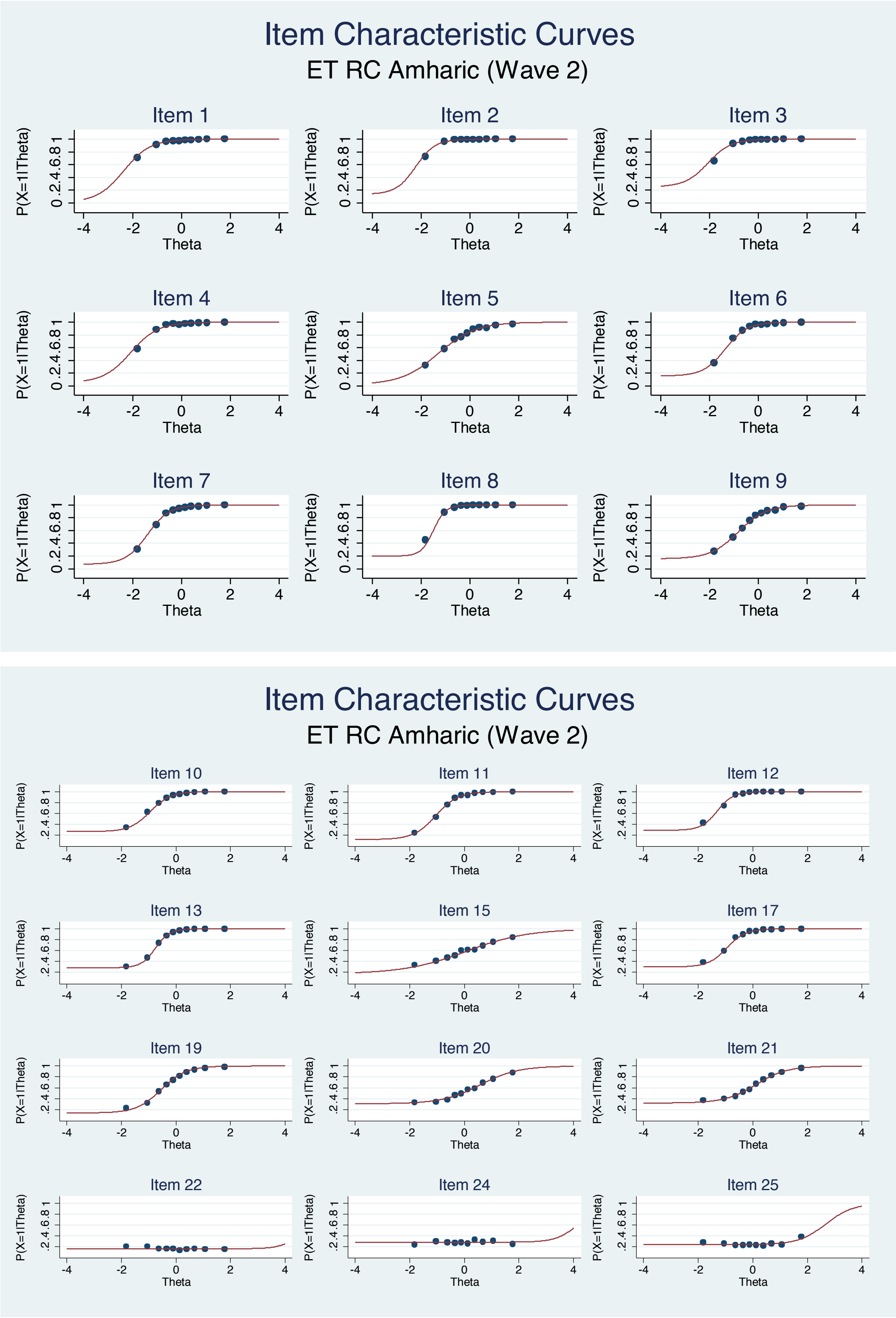
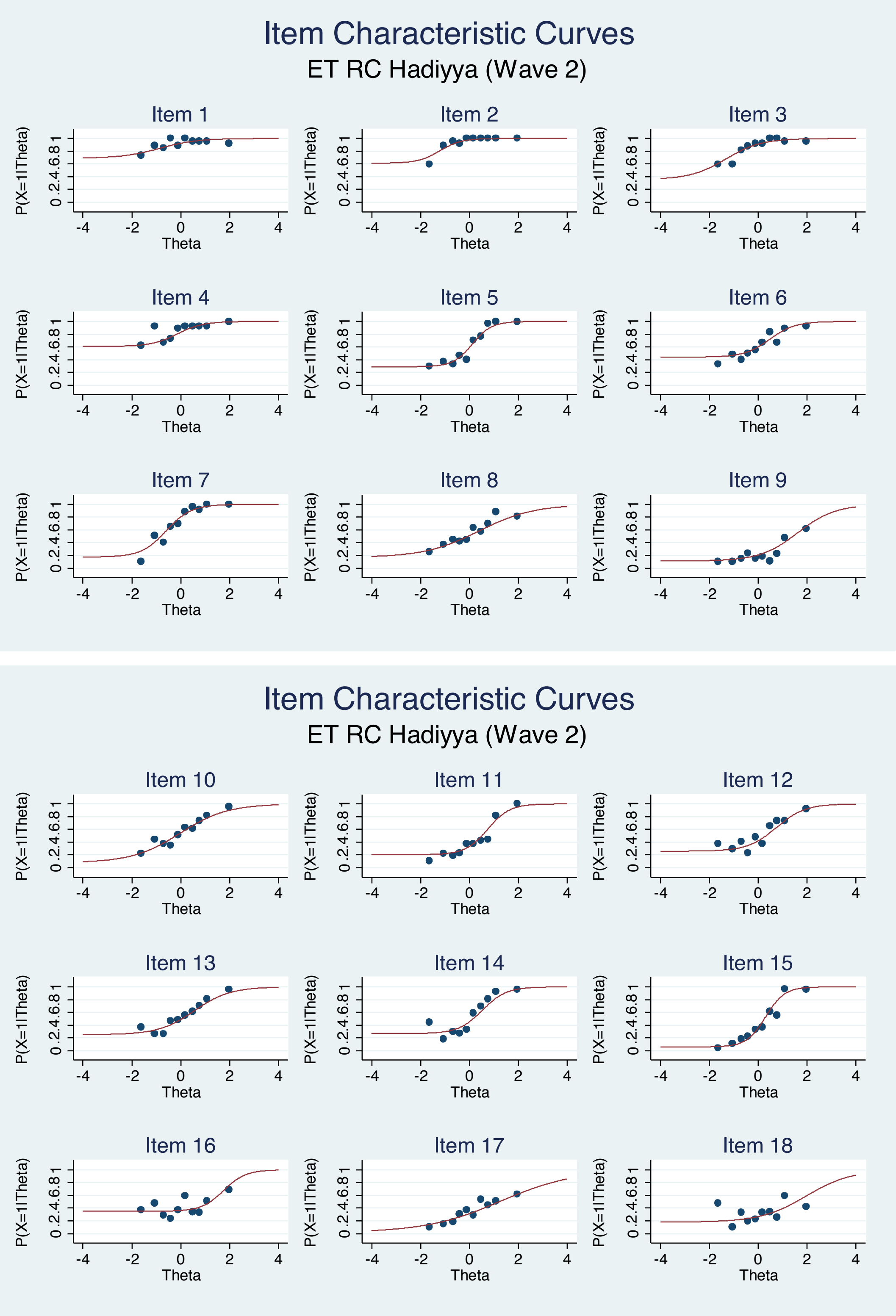


Figure B30. *ICC for reading comprehension test for Hadiyya*



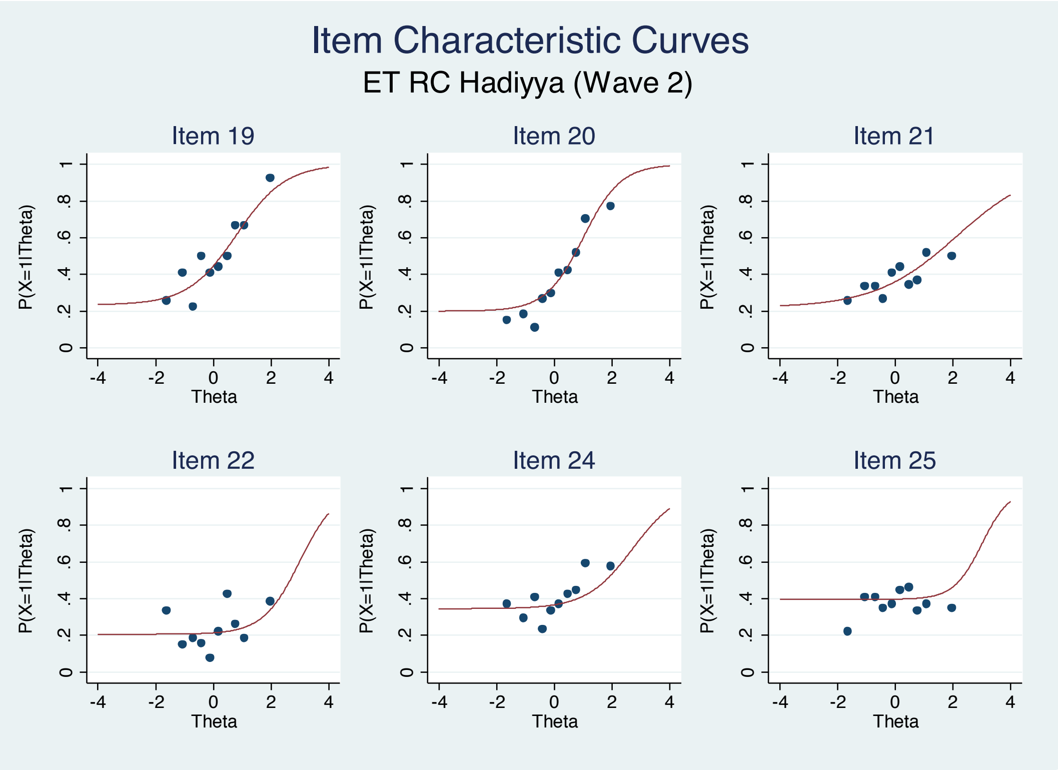
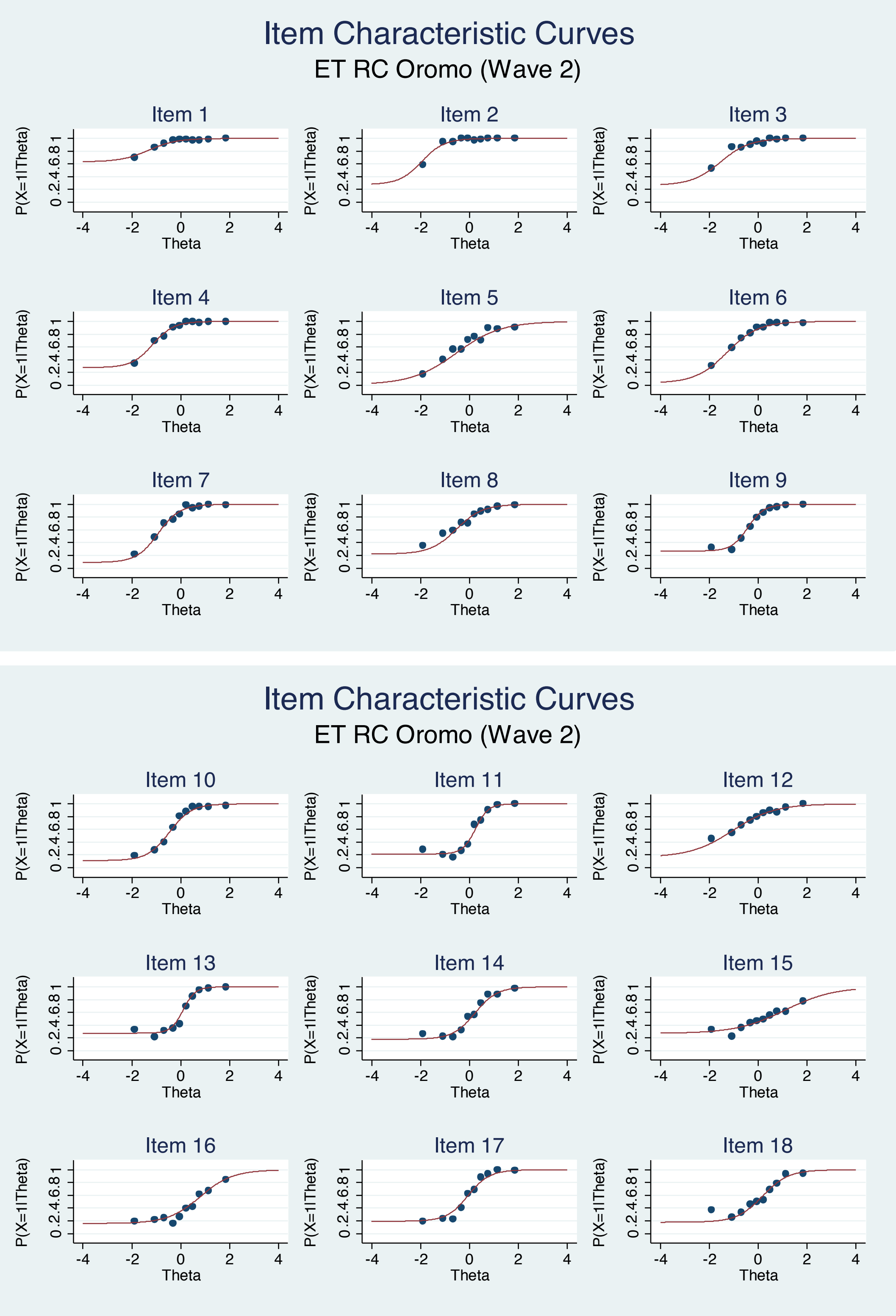

Item Characteristic Curves
ET RC Hadiyya (Wave 2)
Item 19
Item 20
Item 21
Item 22
Item 24
Item 25
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B31. *ICC for reading comprehension test for Oromo*



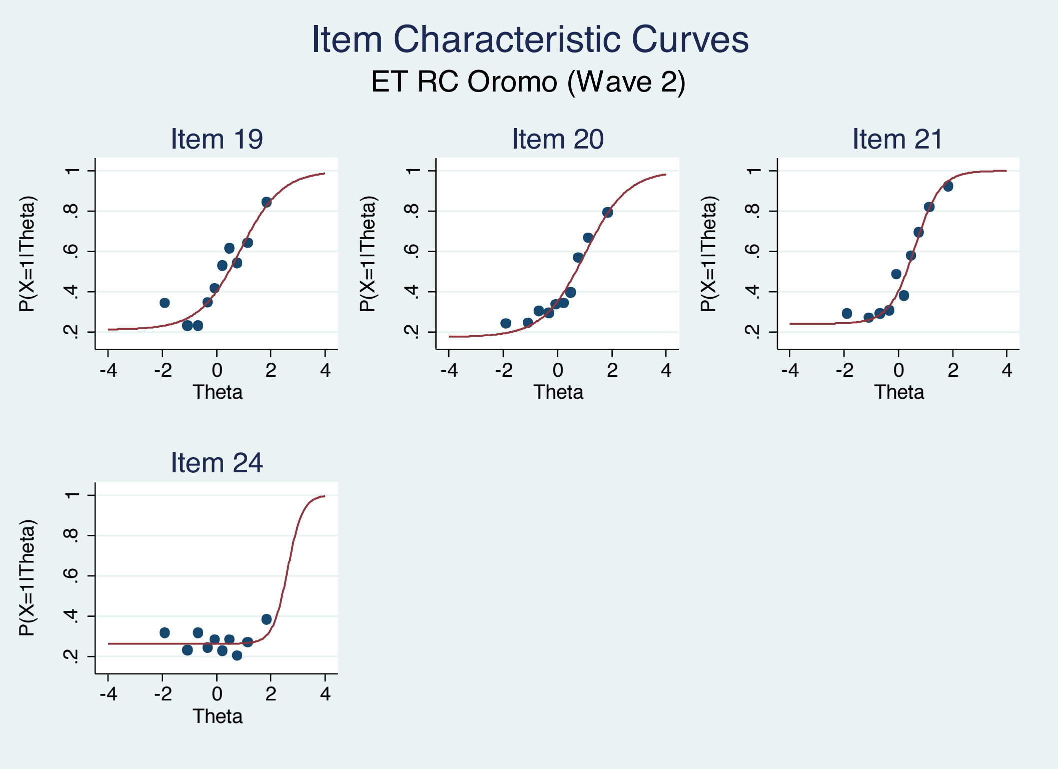
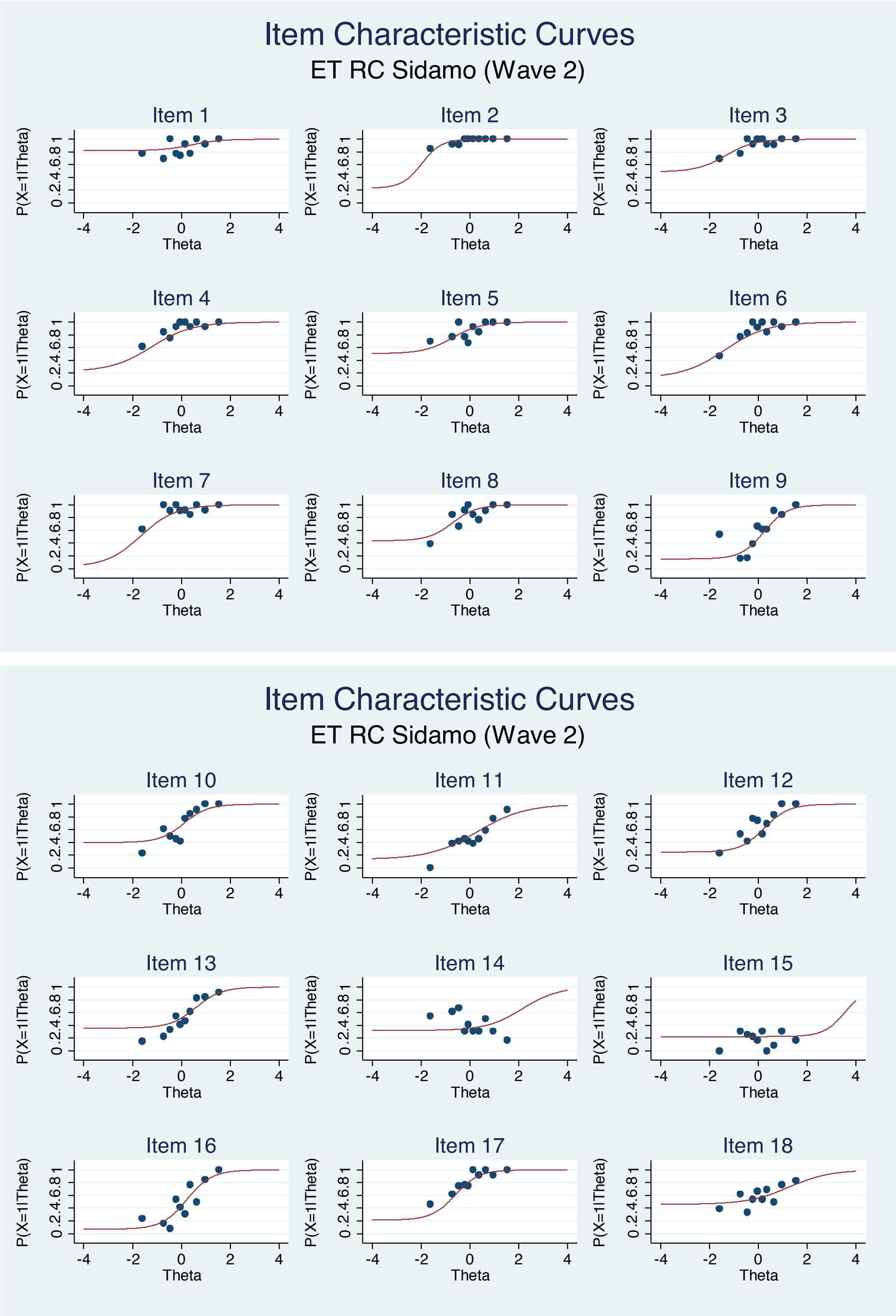

Item Characteristic Curves
ET RC Oromo (Wave 2)
Item 19
Item 20
Item 21
Item 24
P(X=1|Theta)
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B32. *ICC for reading comprehension test for Sidamo*



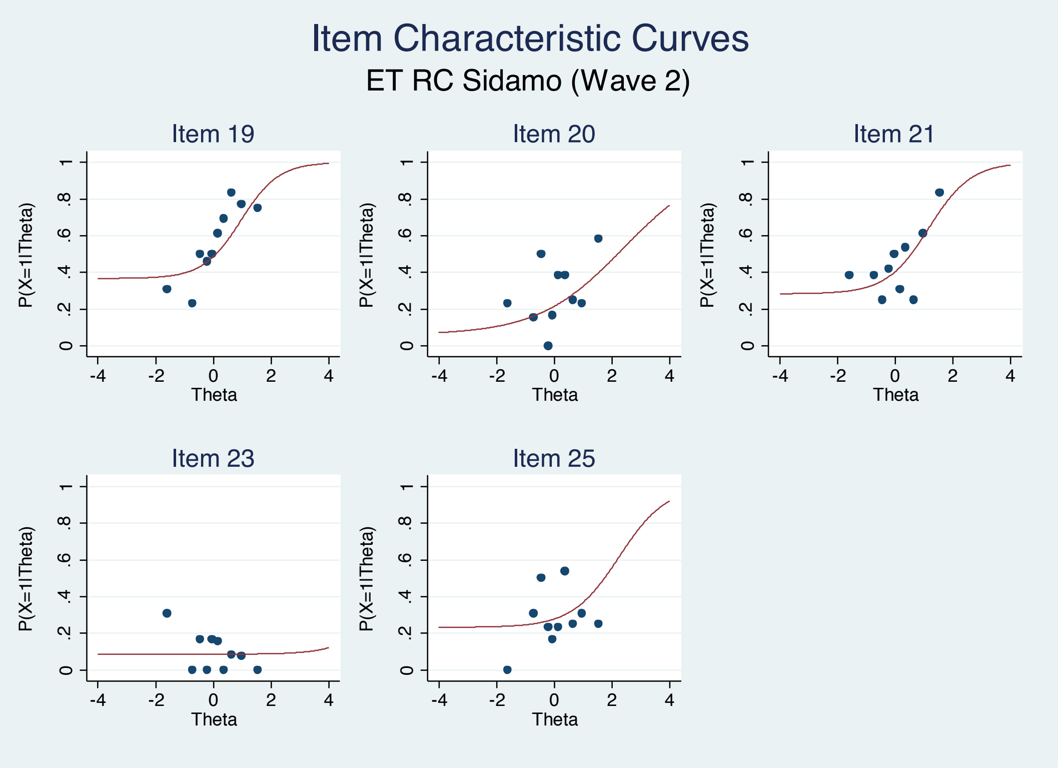
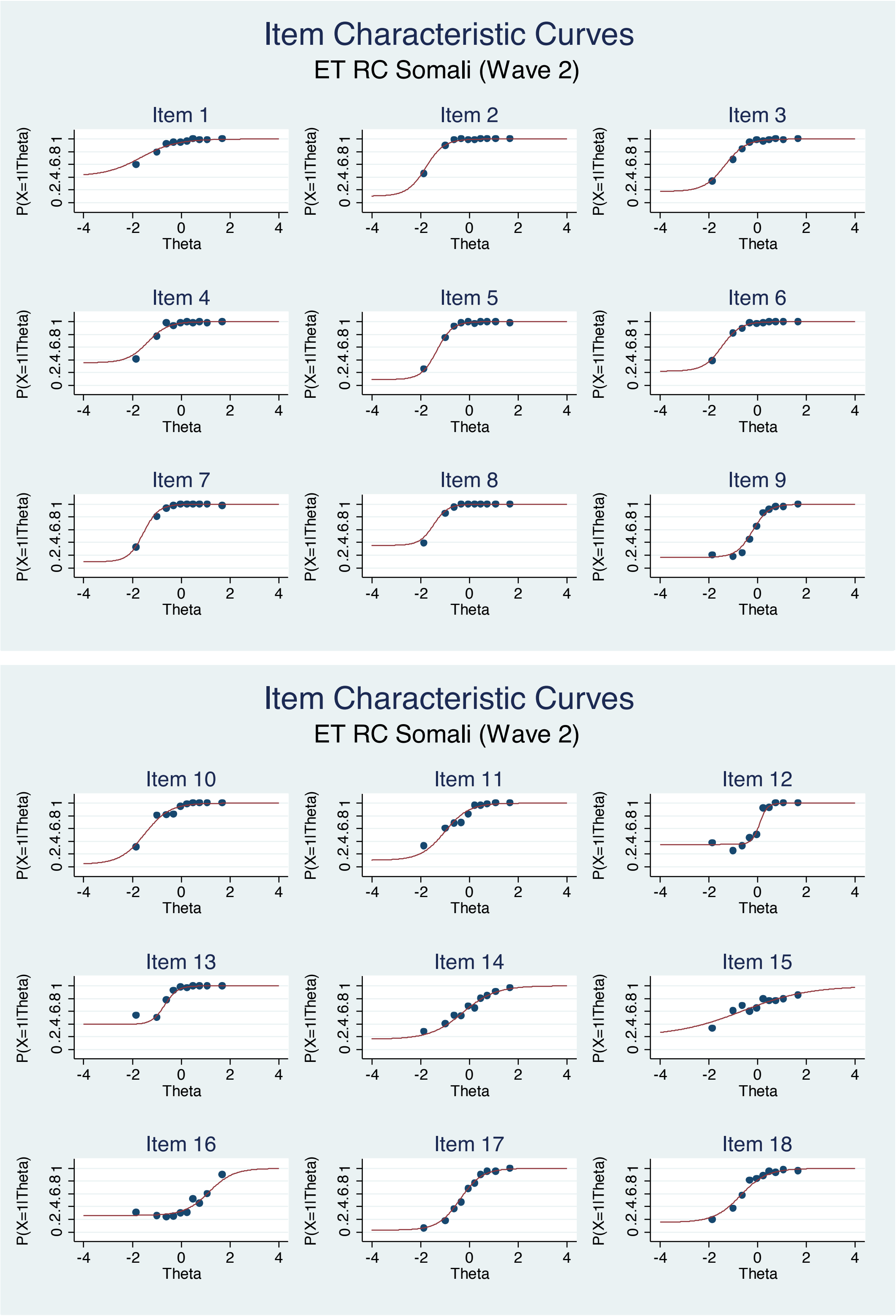

Item Characteristic Curves
ET RC Sidamo (Wave 2)
Item 19
Item 20
Item 21
Item 23
Item 25
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B33. *ICC for reading comprehension test for Somali*



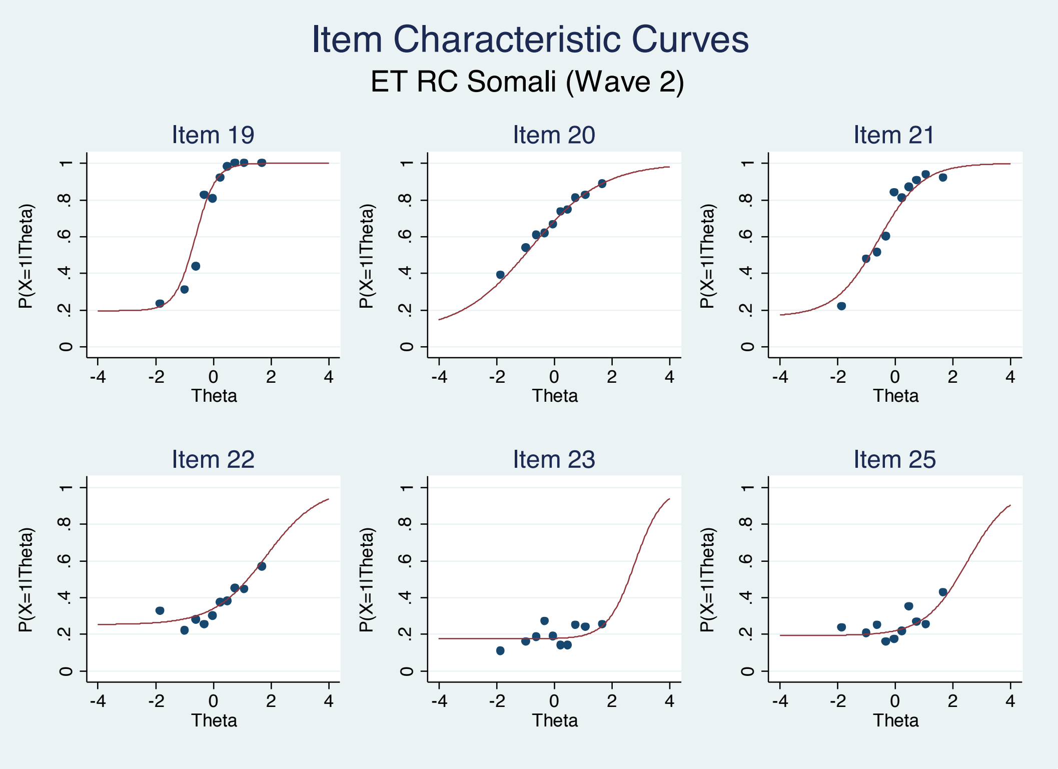
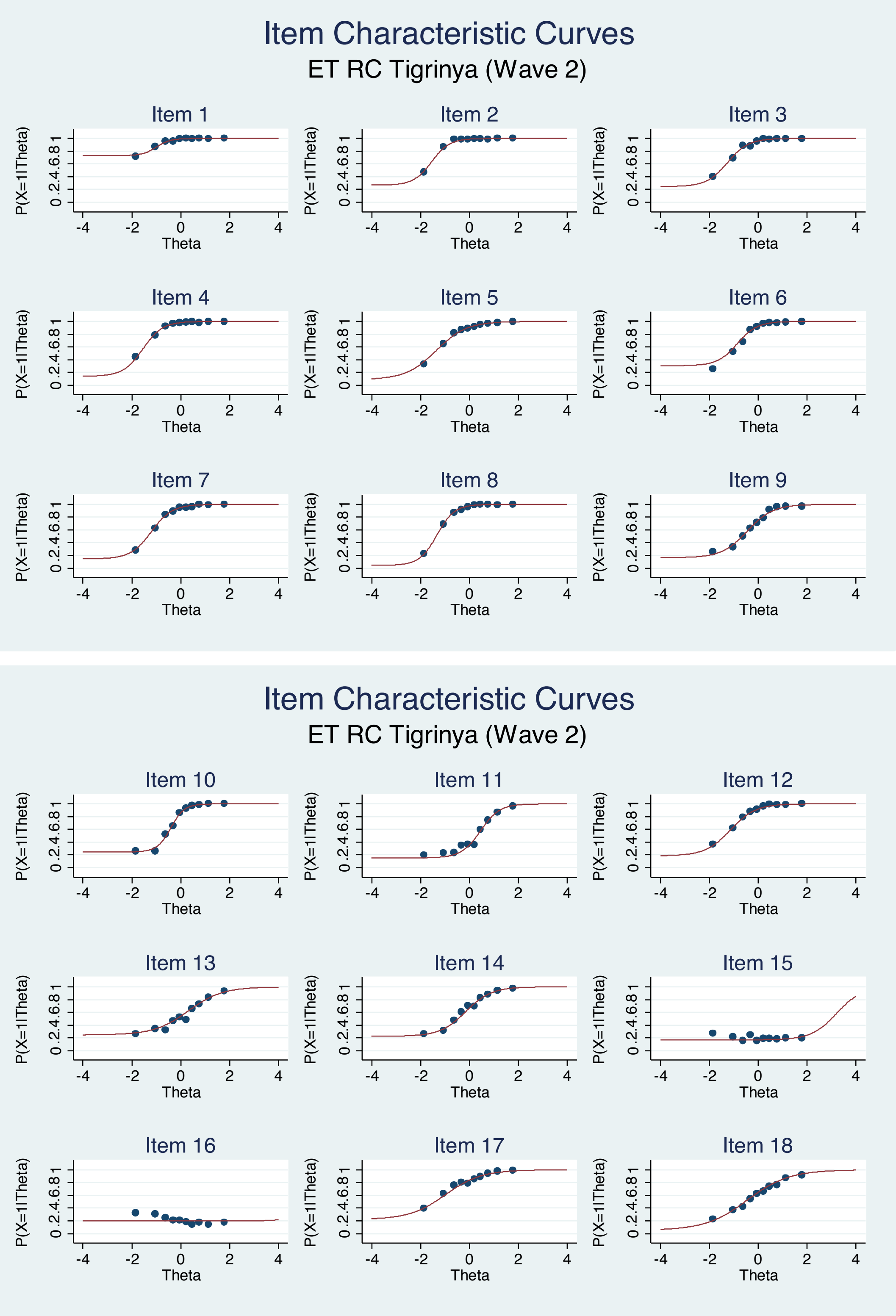

Item Characteristic Curves
ET RC Somali (Wave 2)
Item 19
Item 20
Item 21
Item 22
Item 23
Item 25
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Figure B34. *ICC for reading comprehension test for Tigrinya*



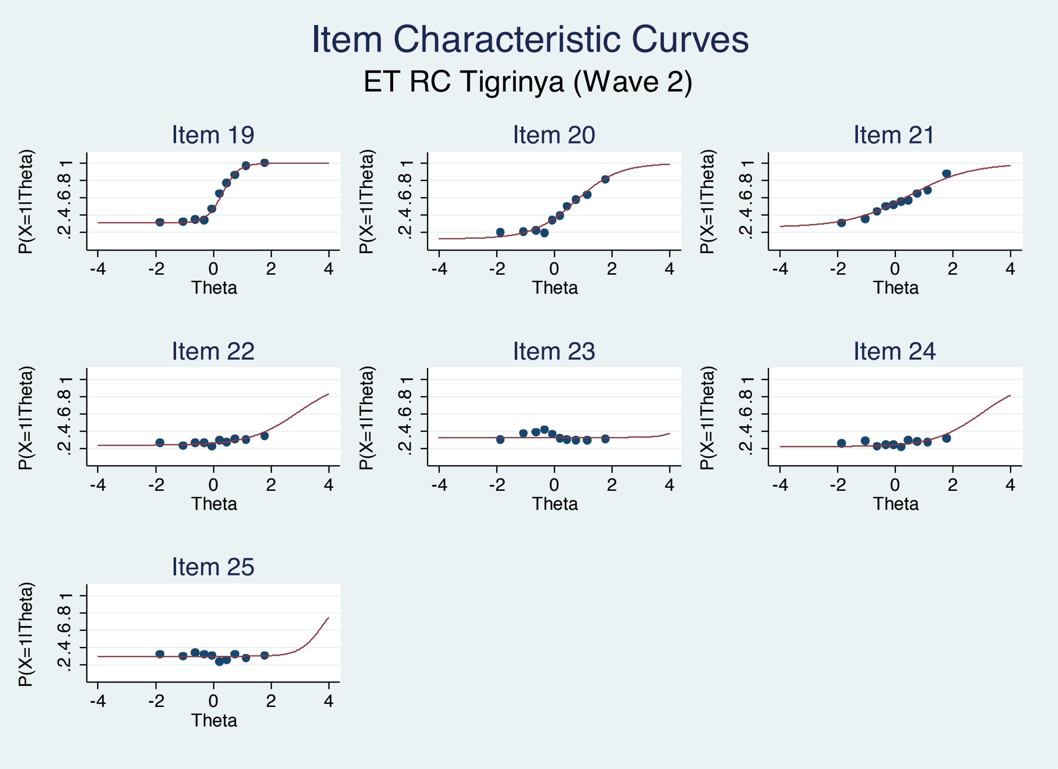
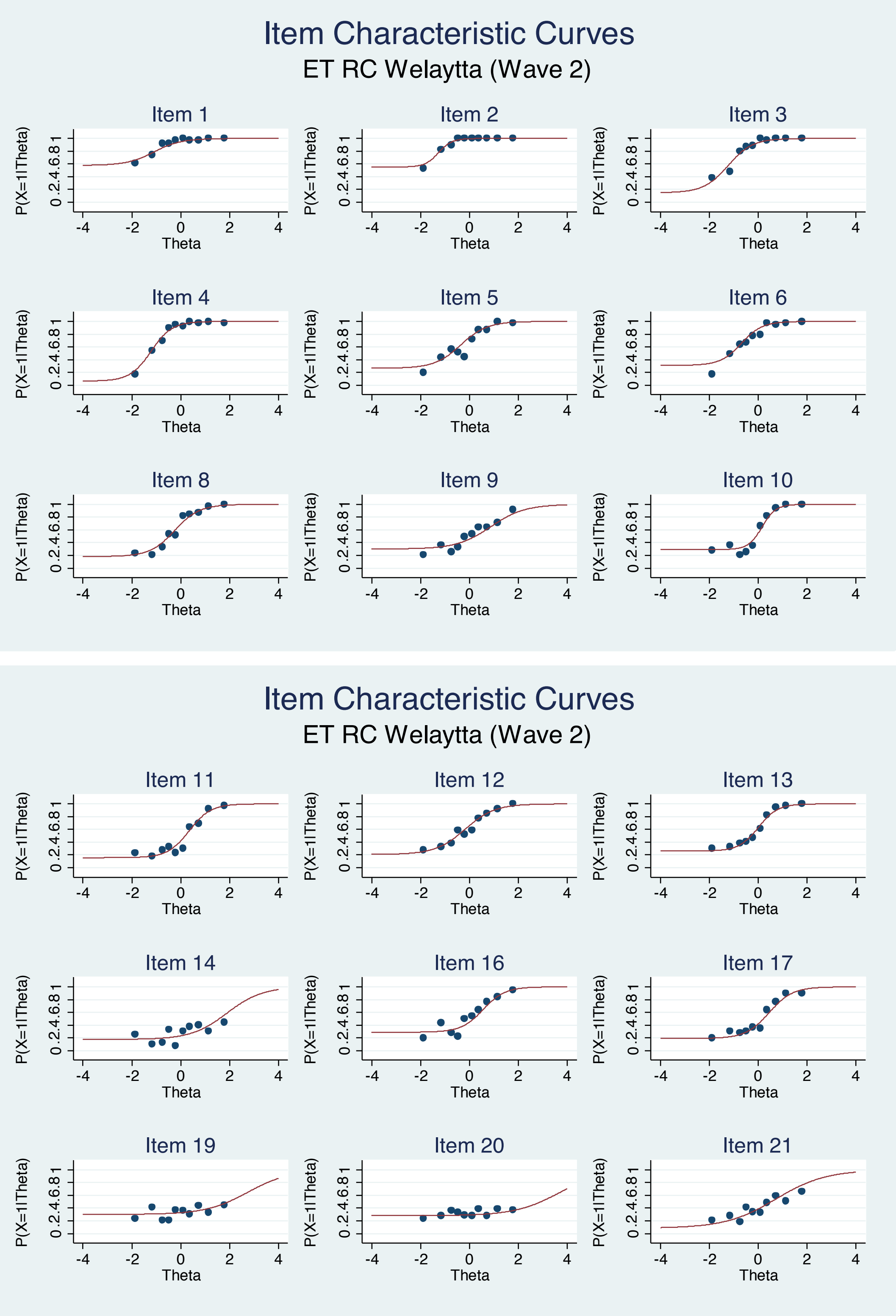
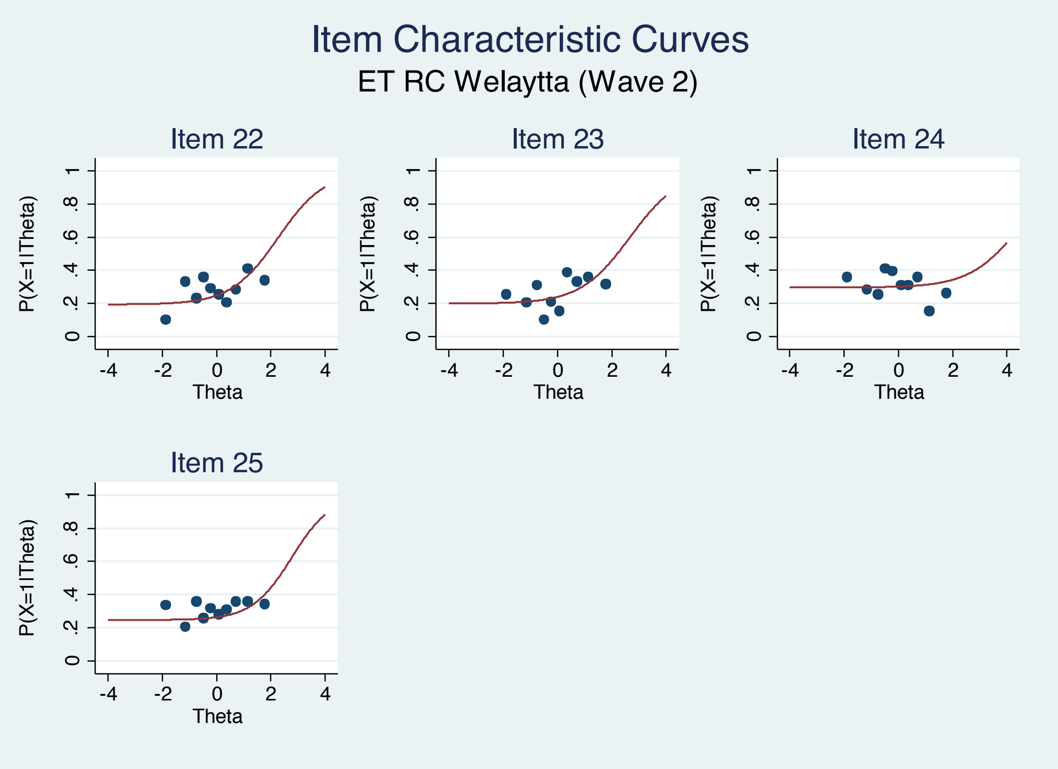

Item Characteristic Curves
ET RC Tigrinya (Wave 2)
Item 19
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
P(X=1|Theta)
.2 .4 .6 .8 1
-4 -2 0 2 4
Theta

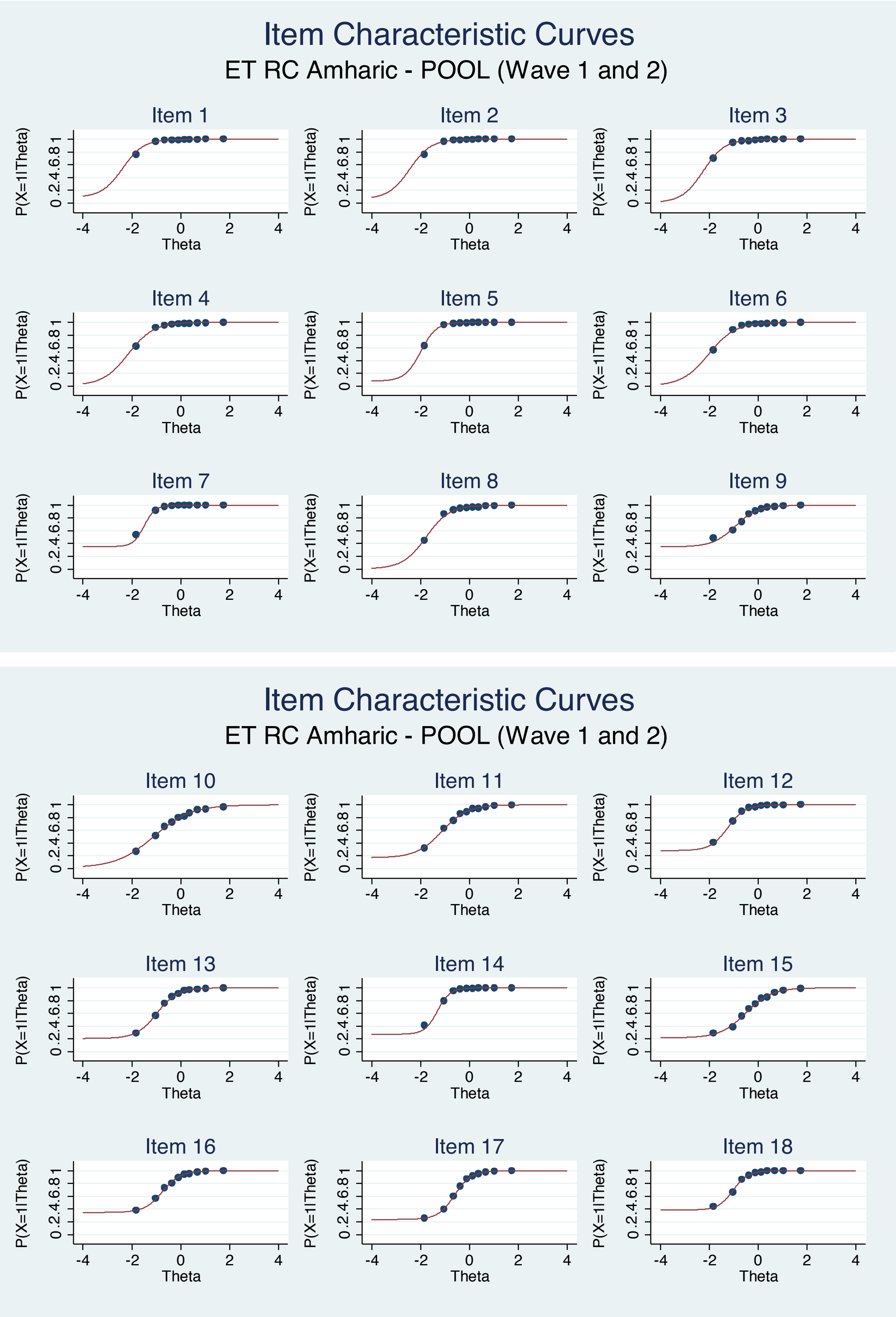
Figure B35. *ICC for reading comprehension test for Wolaytta*




Item Characteristic Curves
ET RC Welaytta (Wave 2)
Item 22
Item 23
Item 24
Item 25
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta

Pooled sample (Waves 1 and 2)

Figure B36. *ICC for reading comprehension test for Amharic*



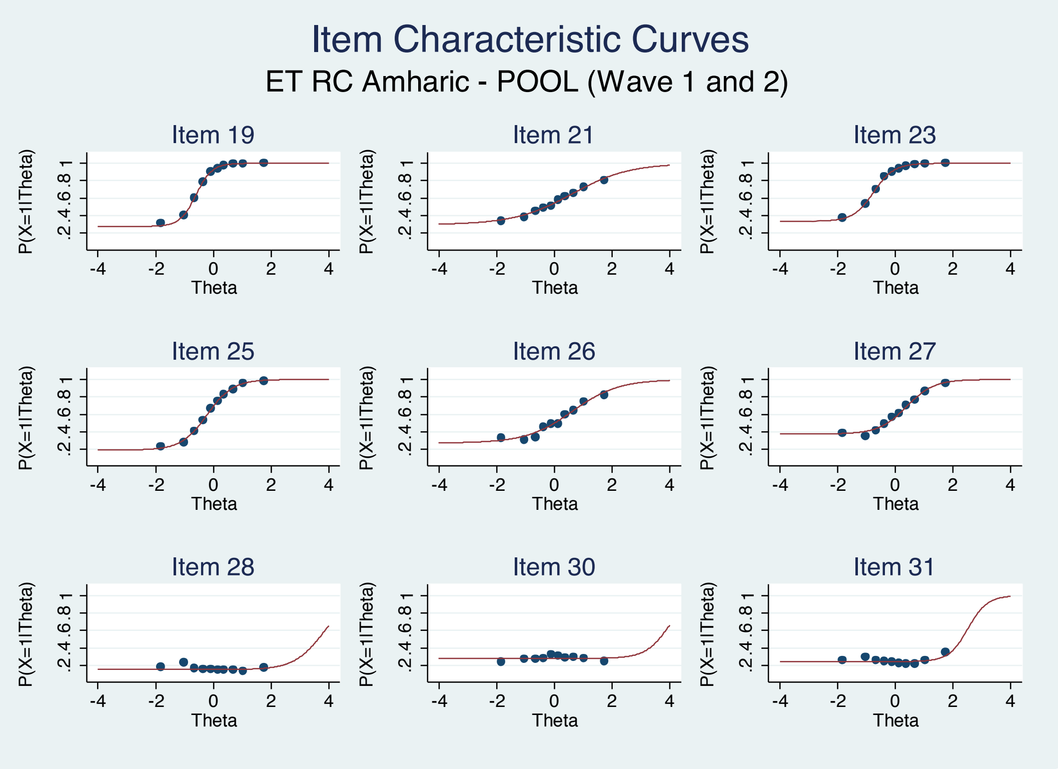
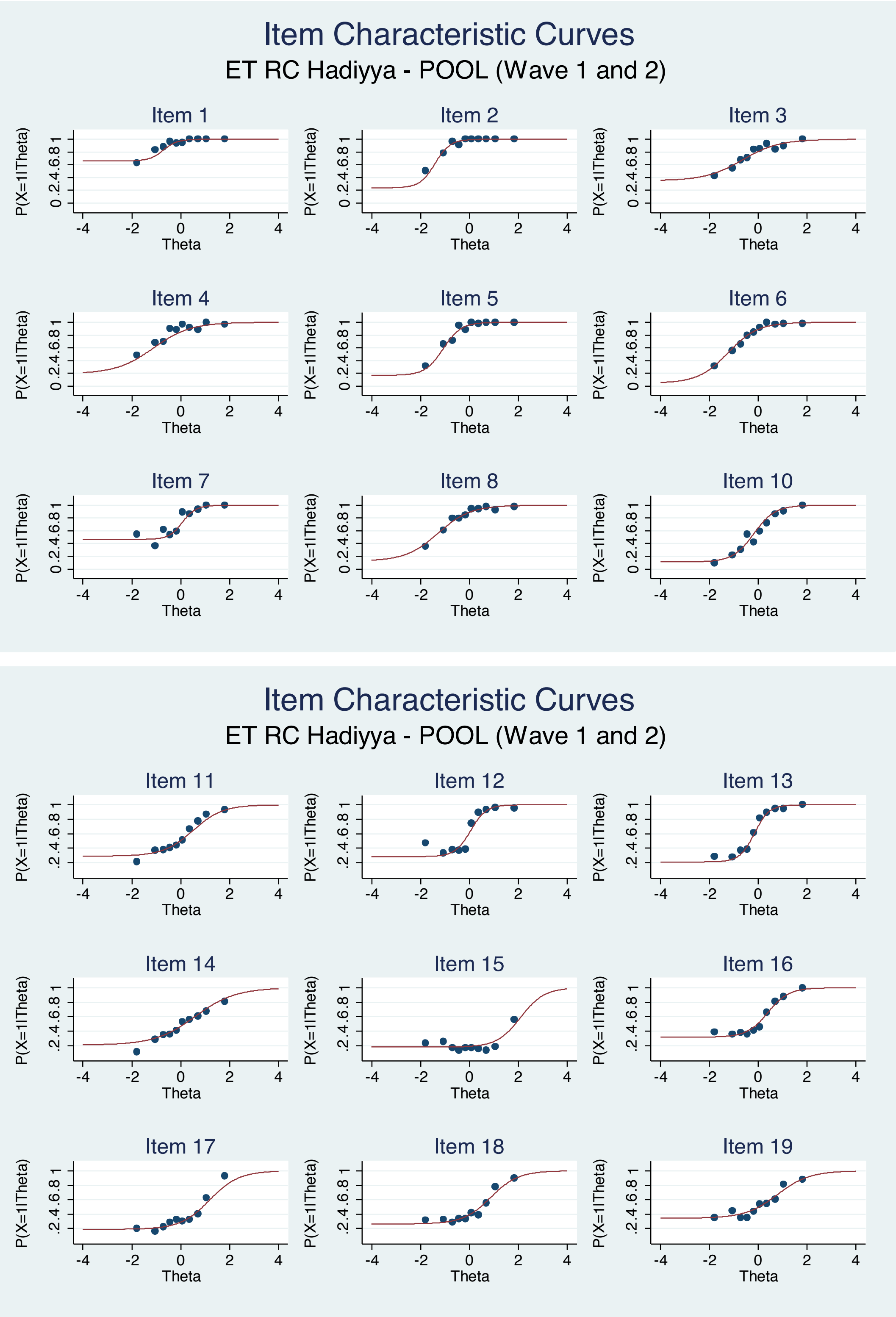

Item Characteristic Curves
ET RC Amharic - POOL (Wave 1 and 2)
Item 19
Item 21
Item 23
Item 25
Item 26
Item 27
Item 28
Item 30
Item 31
P(X=1|Theta)
.2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B37. *ICC for reading comprehension test for Hadiyya*



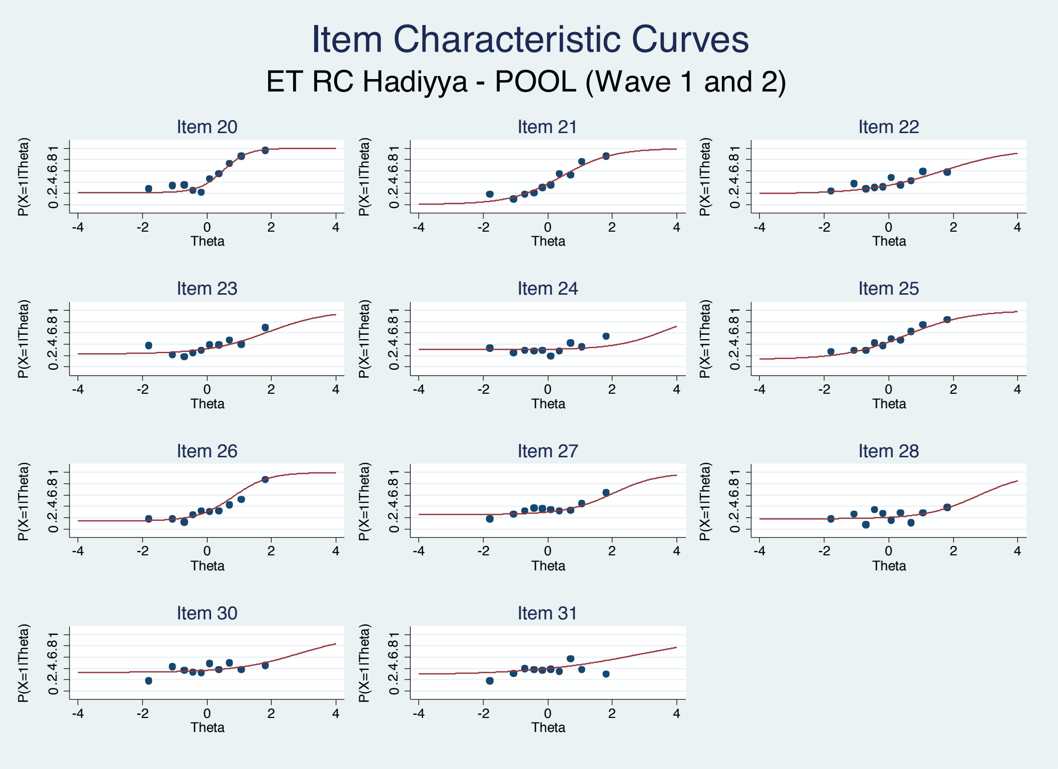
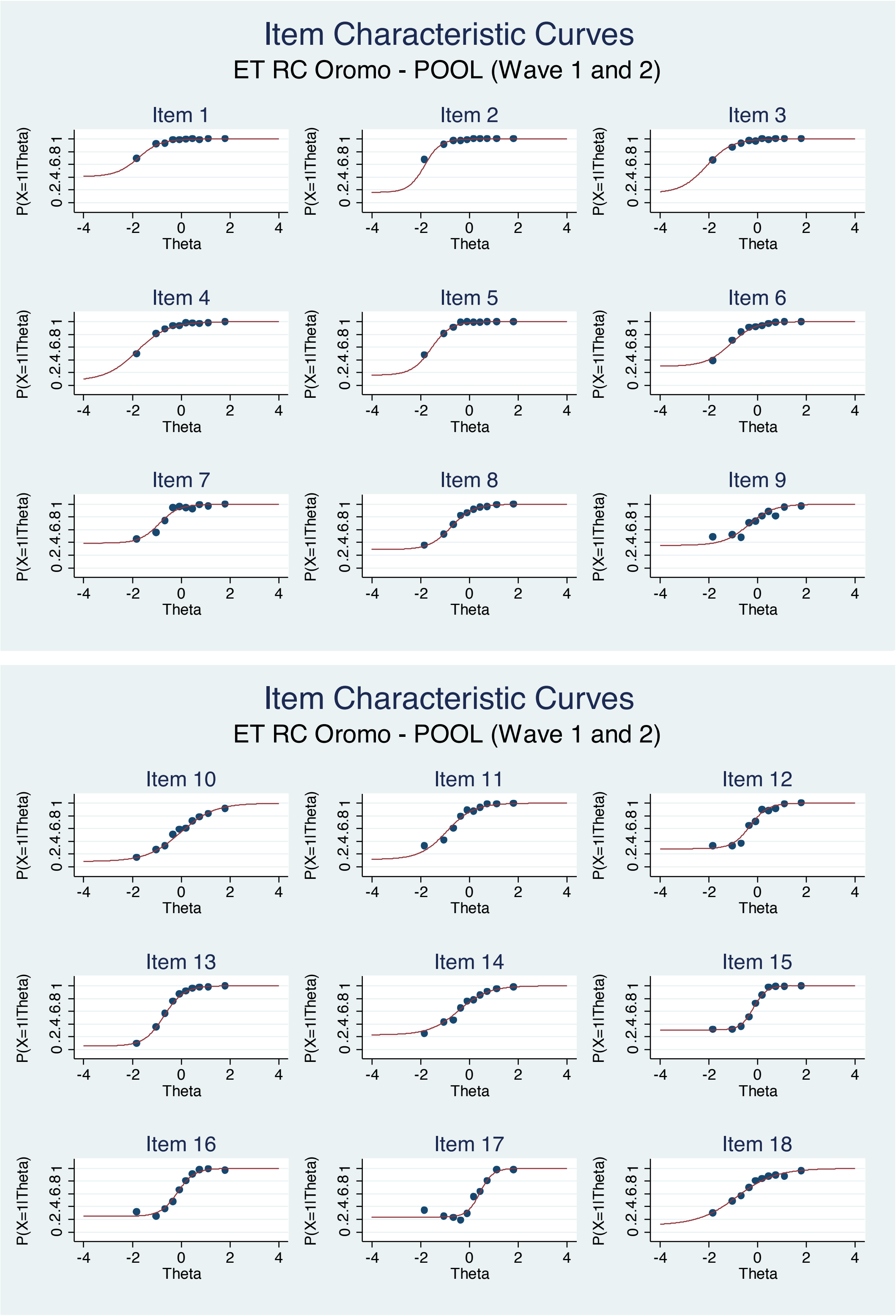

Item Characteristic Curves
ET RC Hadiyya - POOL (Wave 1 and 2)
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
Item 26
Item 27
Item 28
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B38. *ICC for reading comprehension test for Oromo*



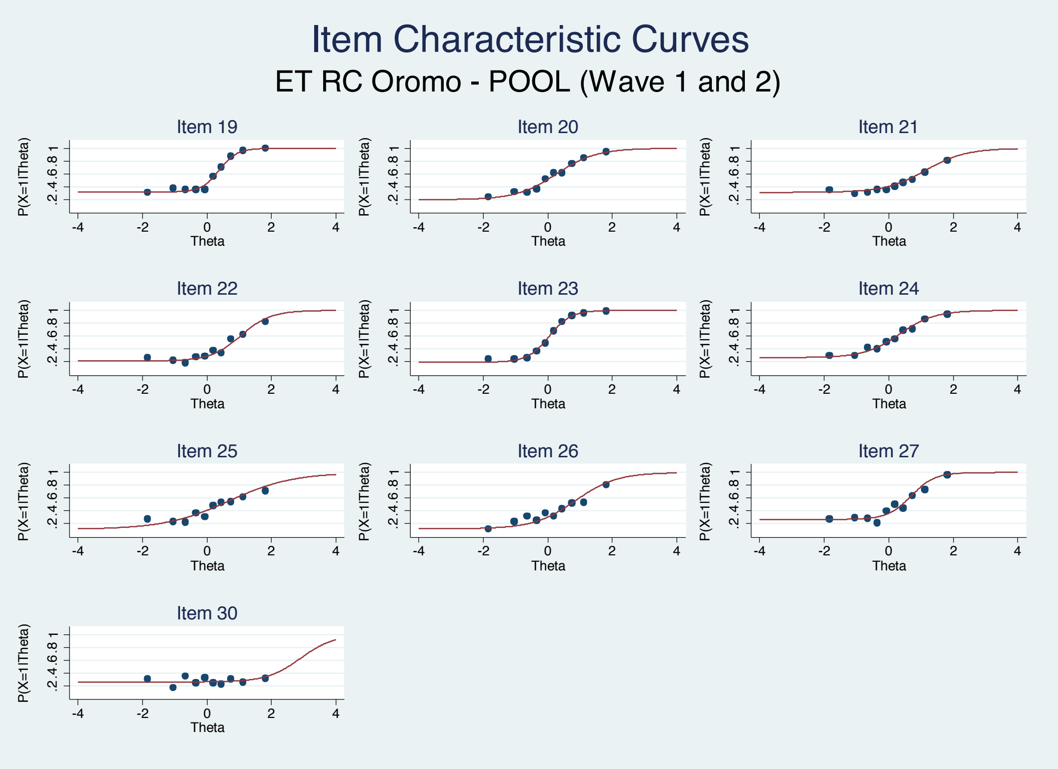
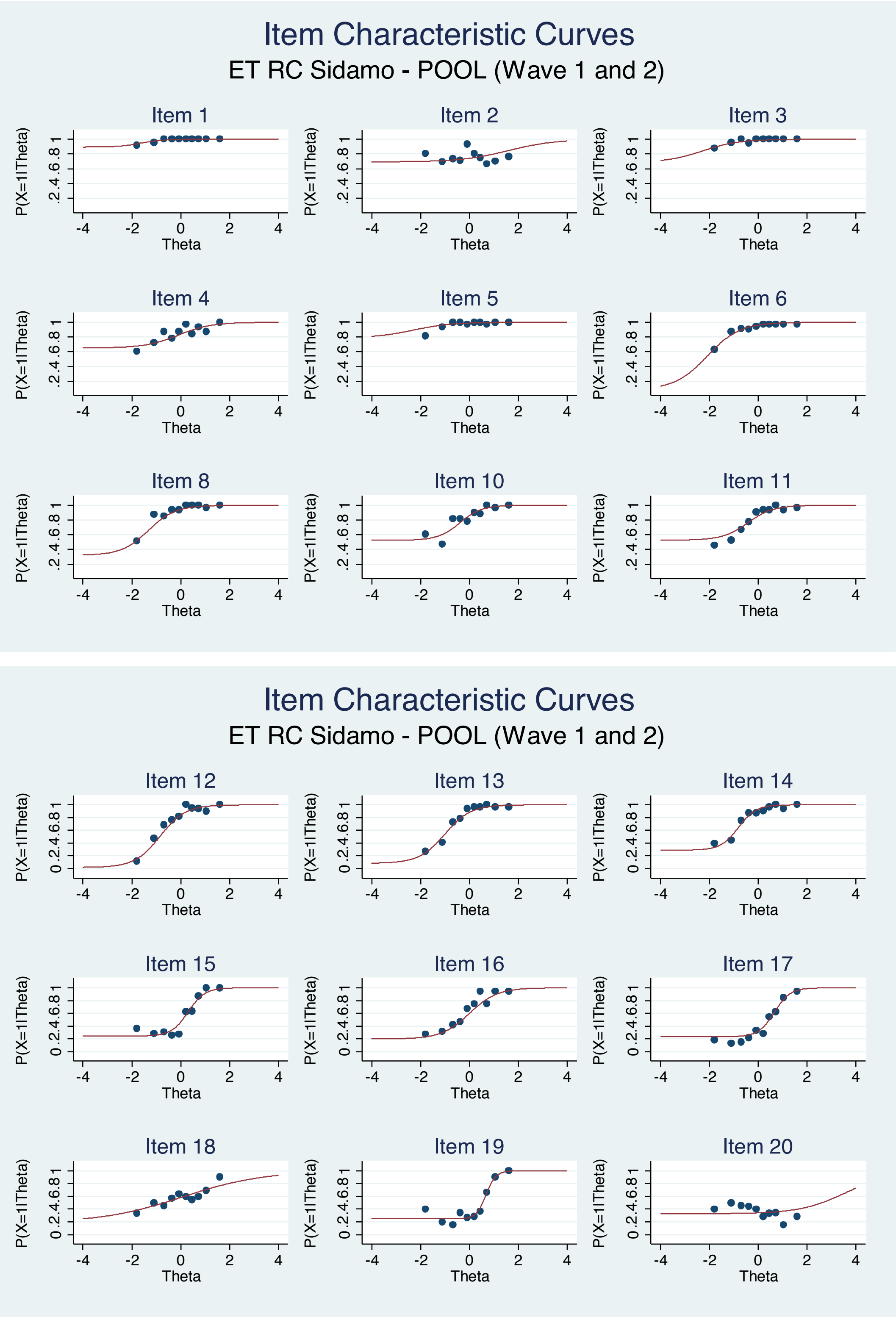

Item Characteristic Curves
ET RC Oromo - POOL (Wave 1 and 2)
Item 19
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
Item 26
Item 27
Item 30
P(X=1|Theta)
.2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B39. *ICC for reading comprehension test for Sidamo*



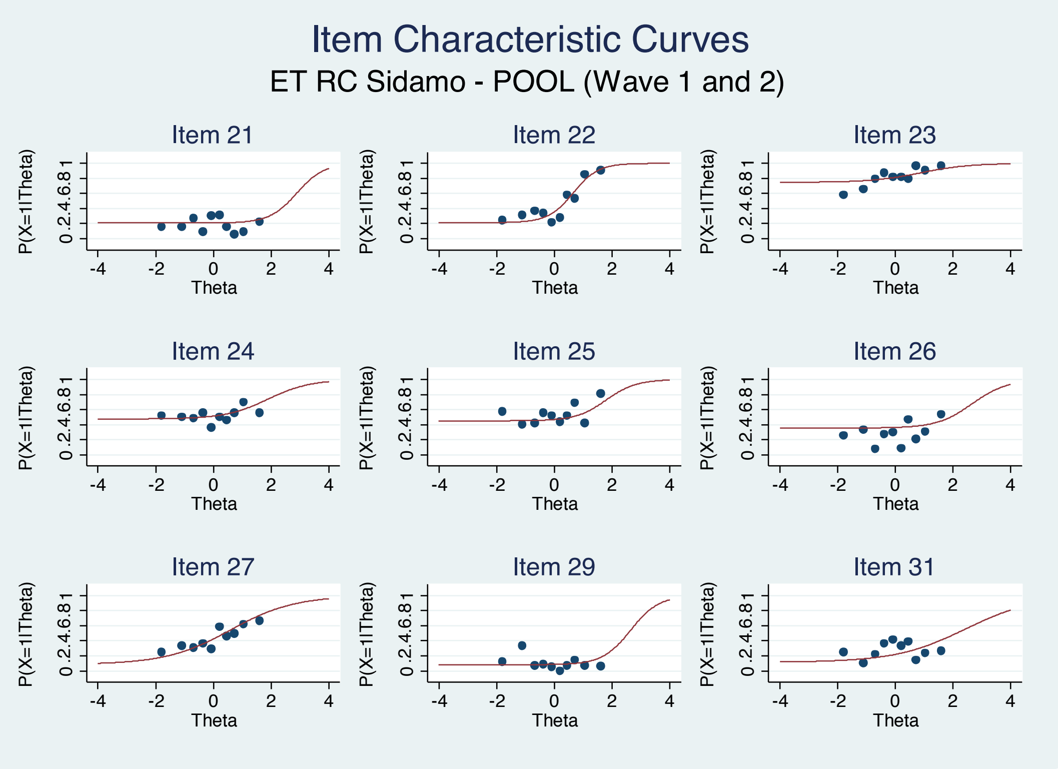
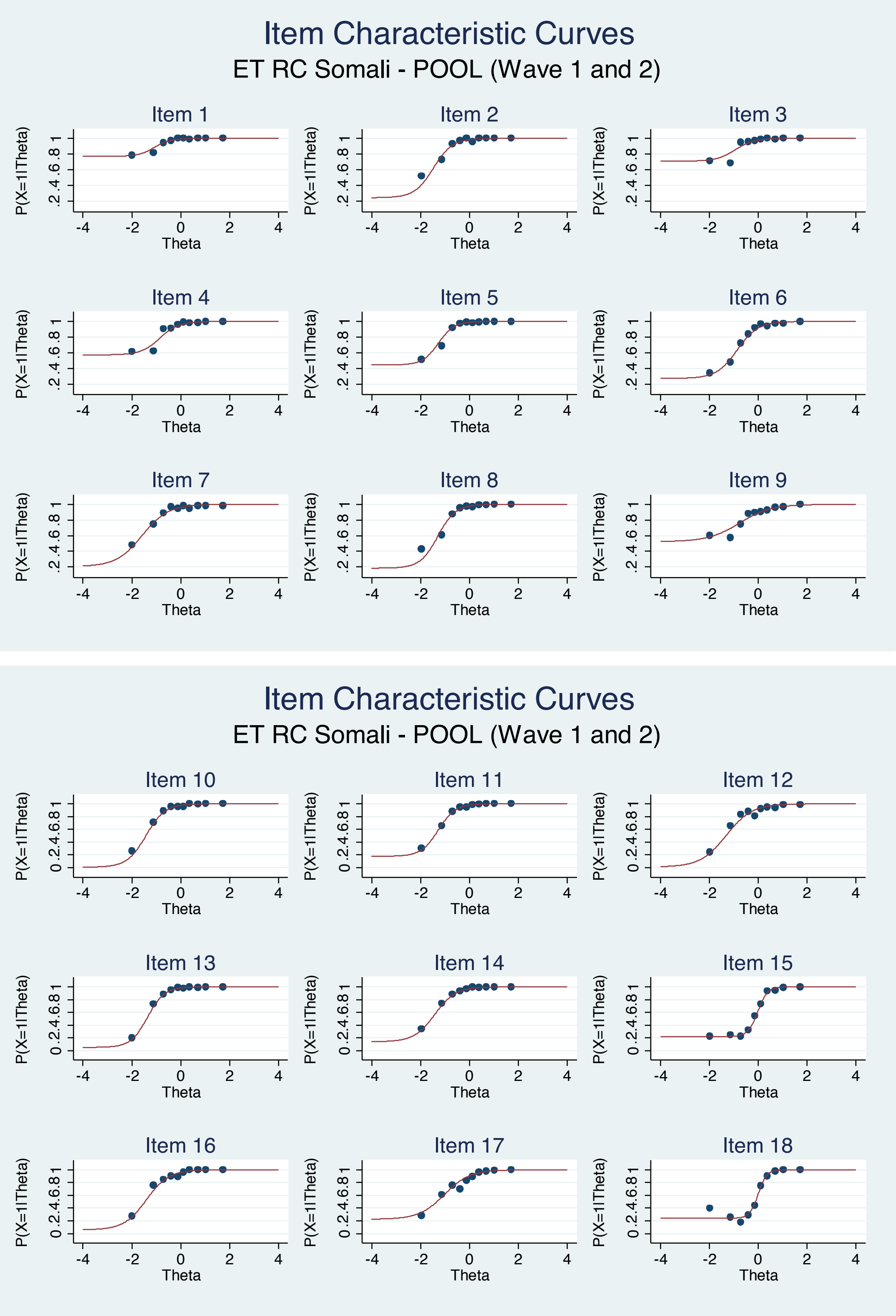

Item Characteristic Curves
ET RC Sidamo - POOL (Wave 1 and 2)
Item 21
Item 22
Item 23
Item 24
Item 25
Item 26
Item 27
Item 29
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B40. *ICC for reading comprehension test for Somali*



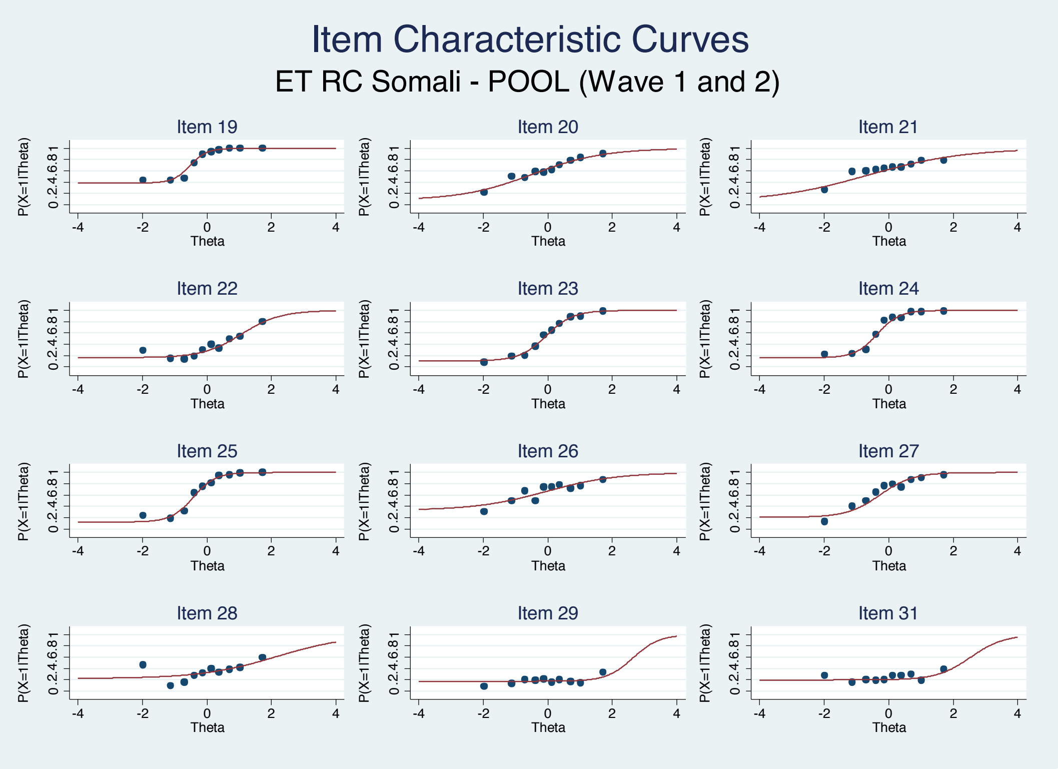
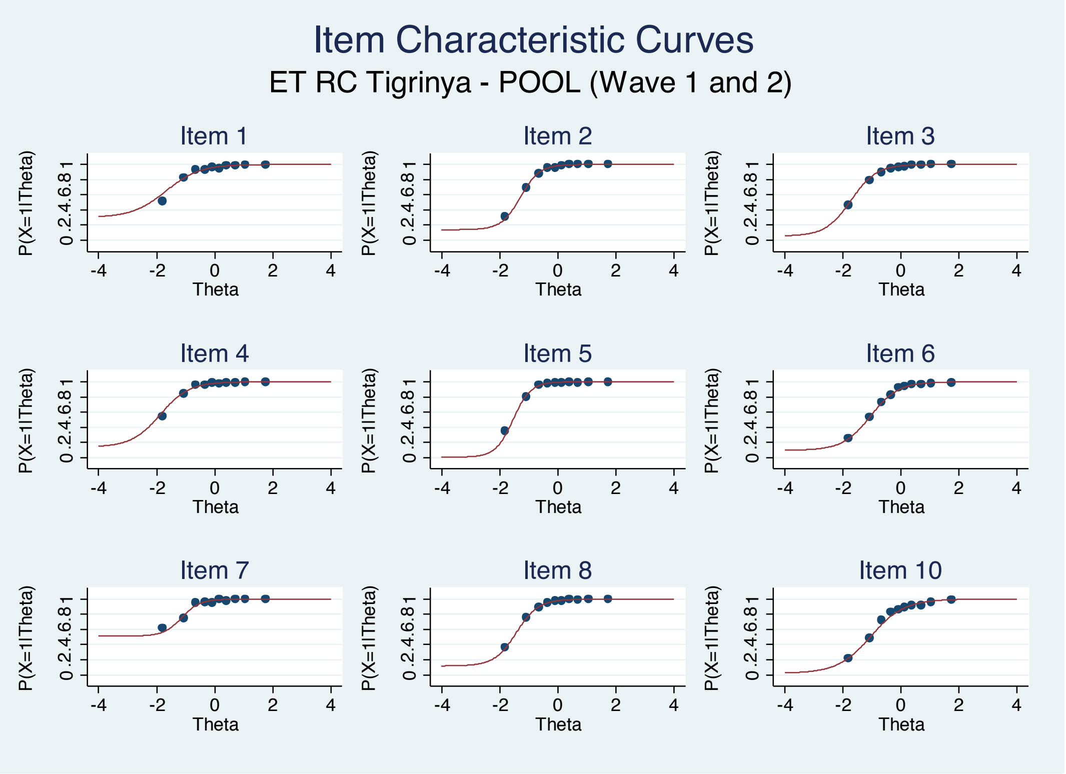
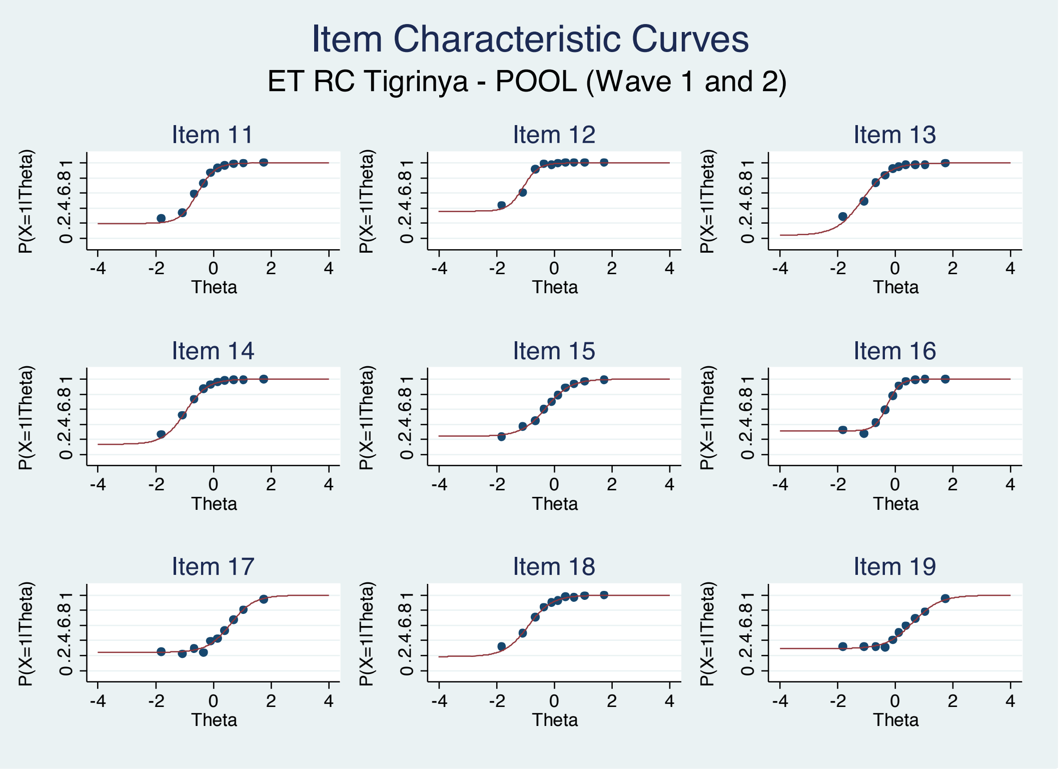

Item Characteristic Curves
ET RC Somali - POOL (Wave 1 and 2)
Item 19
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
Item 26
Item 27
Item 28
Item 29
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B41. *ICC for reading comprehension test for Tigrinya*





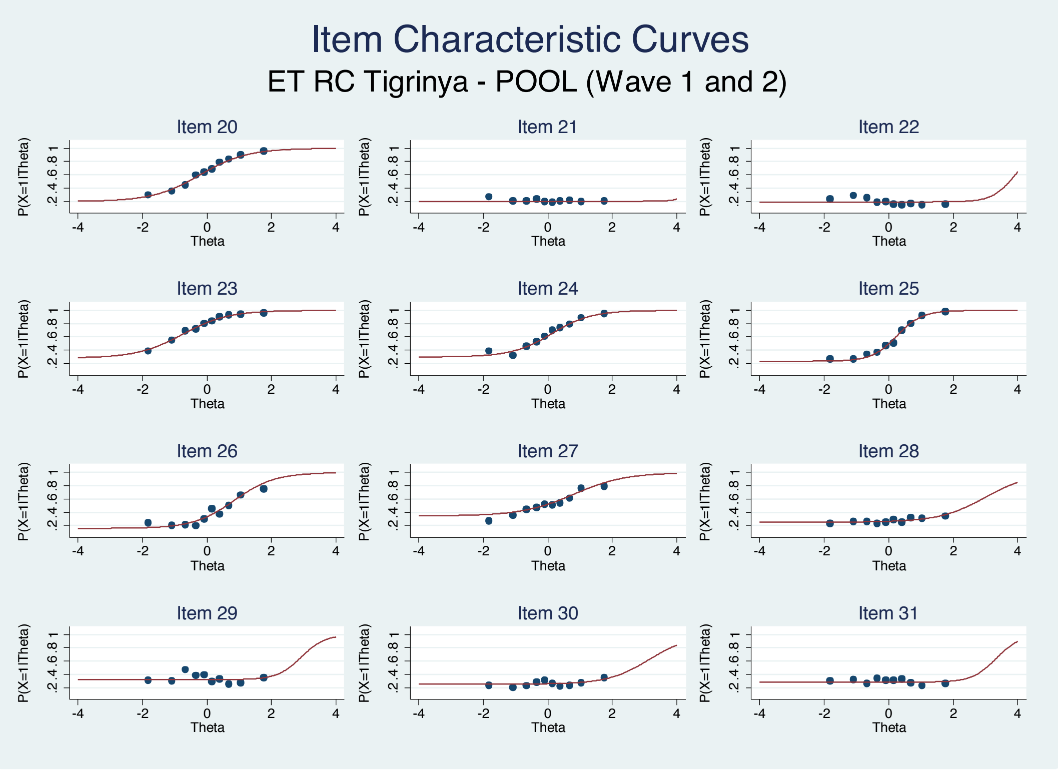
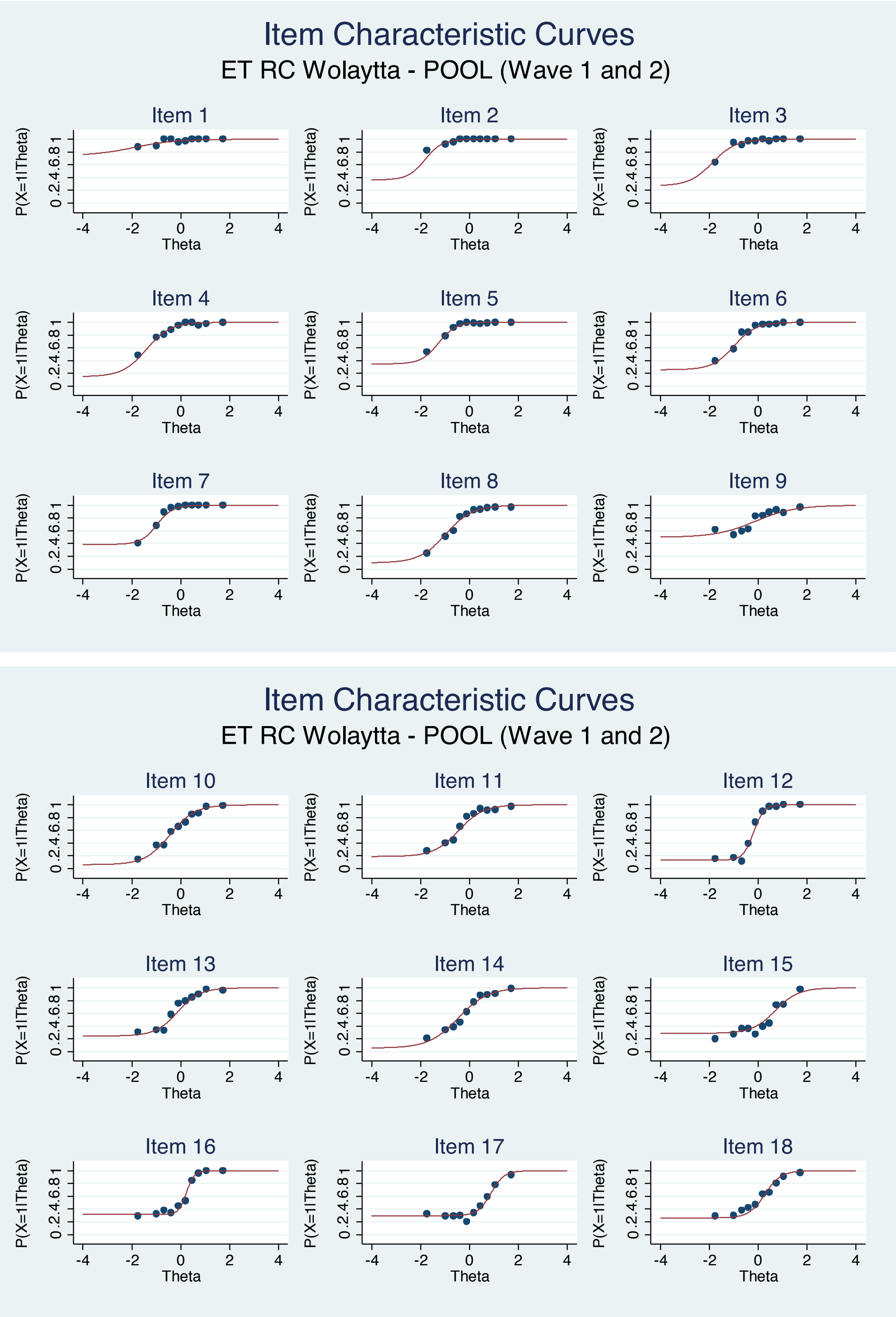
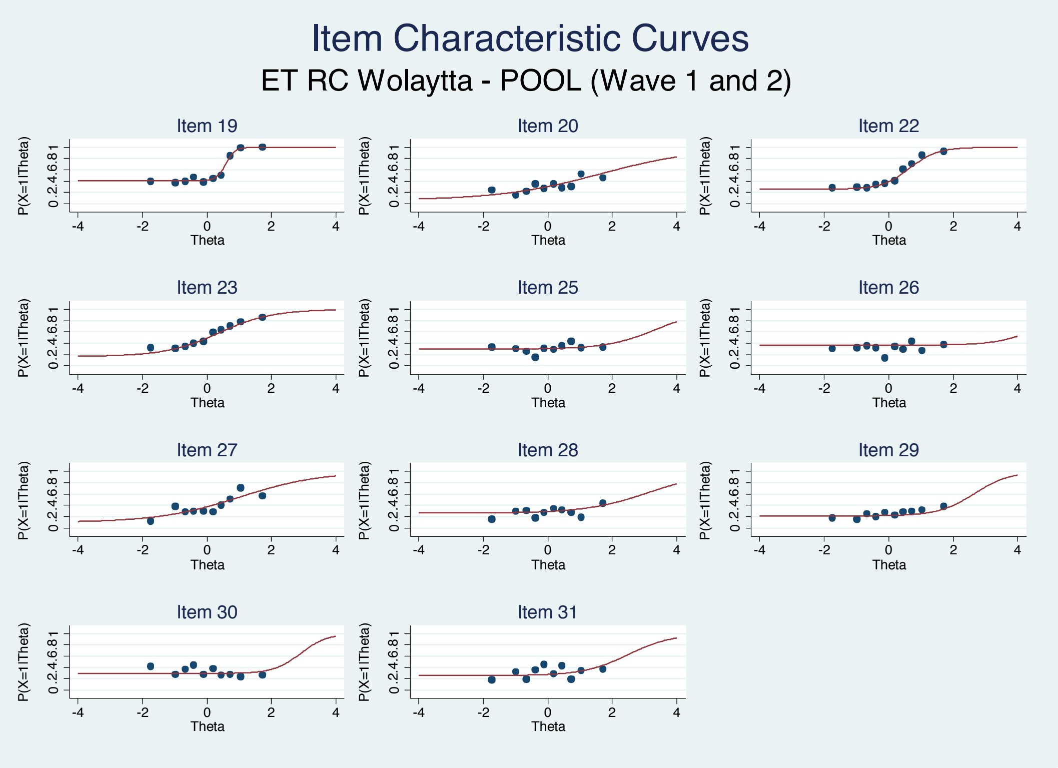

Item Characteristic Curves
ET RC Tigrinya - POOL (Wave 1 and 2)
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
.2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Figure B42. *ICC for reading comprehension test for Wolaytta*




Item Characteristic Curves
ET RC Wolaytta - POOL (Wave 1 and 2)
Item 19
Item 20
Item 22
Item 23
Item 25
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta

Appendix C: Item analysis performed by language and area

Maths

Table C1. *Item analysis performed for Amharic-speaking children*

Note: o = Item dropped + = Item corrected

Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24	x				
25					
26					
27					
28					
29					
30					
31					

Table C2. *Item analysis performed for Hadiyya-speaking children*

Note: o = Item dropped + = Item corrected

Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1					
2		x	x	x	x
3					
4					
5		x	x	x	x
6					
7					
8					
9					
10					
11					
12					
13				+	
14	x				
15	o				
16				+	+
17					
18	x				
19	x				
20	x				
21					
22	x				
23	x				
24	o				
25	o				
26					
27					
28					
29	x				
30					
31	x				

Table C3. *Item analysis performed for Oromo-speaking children*

Note: o = Item dropped + = Item corrected

Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14	o				
15					
16					
17					
18					
19					
20	o				
21	x				
22	o				
23					
24	o				
25	x				
26					
27					
28					
29					
30	x				
31					

Table C4. *Item analysis performed for Sidamo-speaking children*

Note: o = Item dropped + = Item corrected

Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1	x				
2			+		
3	x				
4					
5					
6			+		
7					
8			+		+
9					
10		x	x	x	x
11			+		
12		x	x	x	x
13					+
14	x				
15		x	x	x	x
16	x				
17					
18	x				
19	x				
20	o				
21	x				
22	o				
23		+			
24	x				
25	x				
26					
27					
28					
29	x				
30	x				
31	x				

Table C5. *Item analysis performed for Somali-speaking children*

Note: o = Item dropped + = Item corrected

Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14	x				
15	o				
16					
17					
18					
19					
20	x				
21	x				
22	x				
23					
24	o				
25	x				
26					
27					
28	x				
29					
30					
31	x				

Table C6. *Item analysis performed for Tigrinya-speaking children*

Note: o = Item dropped + = Item corrected

Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1					
2					
3	o				
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16	o				
17					
18	x				
19	x				
20	x				
21	x				
22	x				
23					
24	o				
25	o				
26					
27					
28					
29					
30					
31	x				

Table C7. *Item analysis performed for Wolaytta-speaking children*

Note: o = Item dropped + = Item corrected

Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14	x				
15					
16					+
17					
18	x				
19	x				
20	o				
21	o				
22	o				
23					
24					
25	o				
26					
27					
28					
29					
30					
31	x				

Reading comprehension

Table C8. *Item analysis performed for Amharic-speaking children*

Note: o = Item dropped + = Item corrected

Item	Itemwithpoorfit	DIF			
		W1M	W1F	W2M	W2F
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20	o				
21					
22	o				
23					
24	o				
25					
26					
27					
28	x				
29	o				
30	x				
31	x				

Table C9. *Item analysis performed for Hadiyya-speaking children*

Note: o = Item dropped + = Item corrected

Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1					
2					
3					
4					
5					
6					
7					
8					
9	o				
10					
11					
12					
13					
14					
15	x				
16					
17					
18					
19					
20					
21					
22					
23					
24	x				
25					
26					
27					
28	x				
29	o				
30	x				
31	x				

Table C10. *Item analysis performed for Oromo-speaking children*

Note: o = Item dropped + = Item corrected

Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28	o				
29	o				
30	x				
31	o				

Table C11. *Item analysis performed for Sidamo-speaking children*

Note: o = Item dropped + = Item corrected

Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1	x				
2	x				
3	x				
4	x				
5	x				
6				+	
7	o				
8					
9	o				
10				+	
11				+	
12					
13				+	
14					
15					
16					
17					
18		+			
19					
20	x				
21	x				
22					
23	x				
24	x				
25	x				
26	x				
27					
28	o				
29	x				
30	o				
31	x				

Table C12. *Item analysis performed for Somali-speaking children*

Note: o = Item dropped + = Item corrected

Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21					
22					
23					
24					
25					
26					
27					
28	x				
29	x				
30	o				
31	x				

Table C13. *Item analysis performed for Tigrinya-speaking children*

Note: o = Item dropped + = Item corrected

Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1					
2					
3					
4					
5					
6					
7					
8					
9	o				
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21	x				
22	x				
23					
24					
25					
26					
27					
28	x				
29	x				
30	x				
31	x				

Table C14. *Item analysis performed for Wolaytta-speaking children*

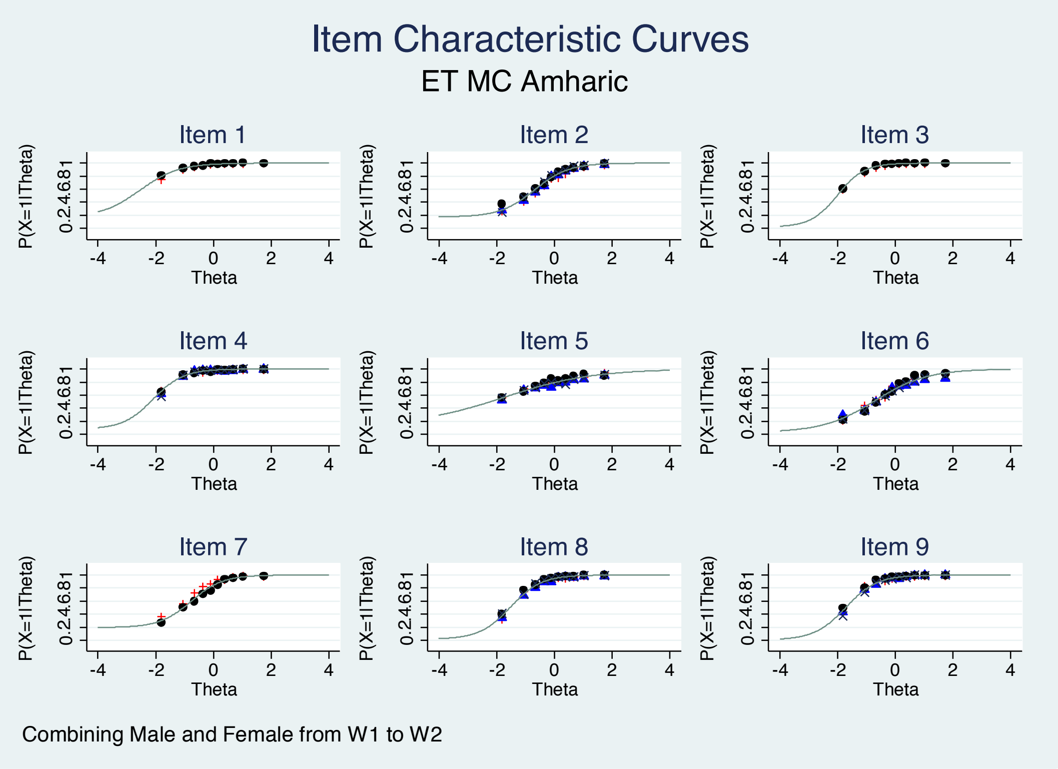
Note: o = Item dropped + = Item corrected

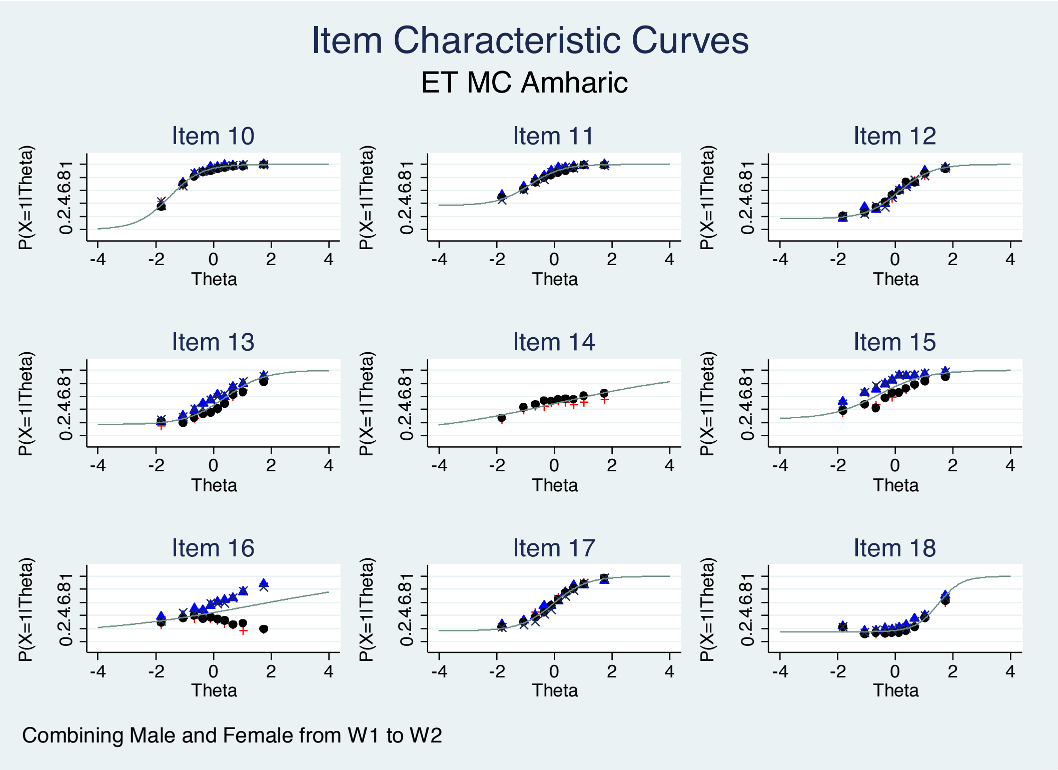
Item	Item with poor fit	DIF			
		W1M	W1F	W2M	W2F
1	x				
2					
3					
4					
5					
6					
7					
8					
9					
10					
11					
12					
13					
14					
15					
16					
17					
18					
19					
20					
21	o				
22					
23					
24	o				
25	x				
26	x				
27					
28	x				
29	x				
30	x				
31	x				

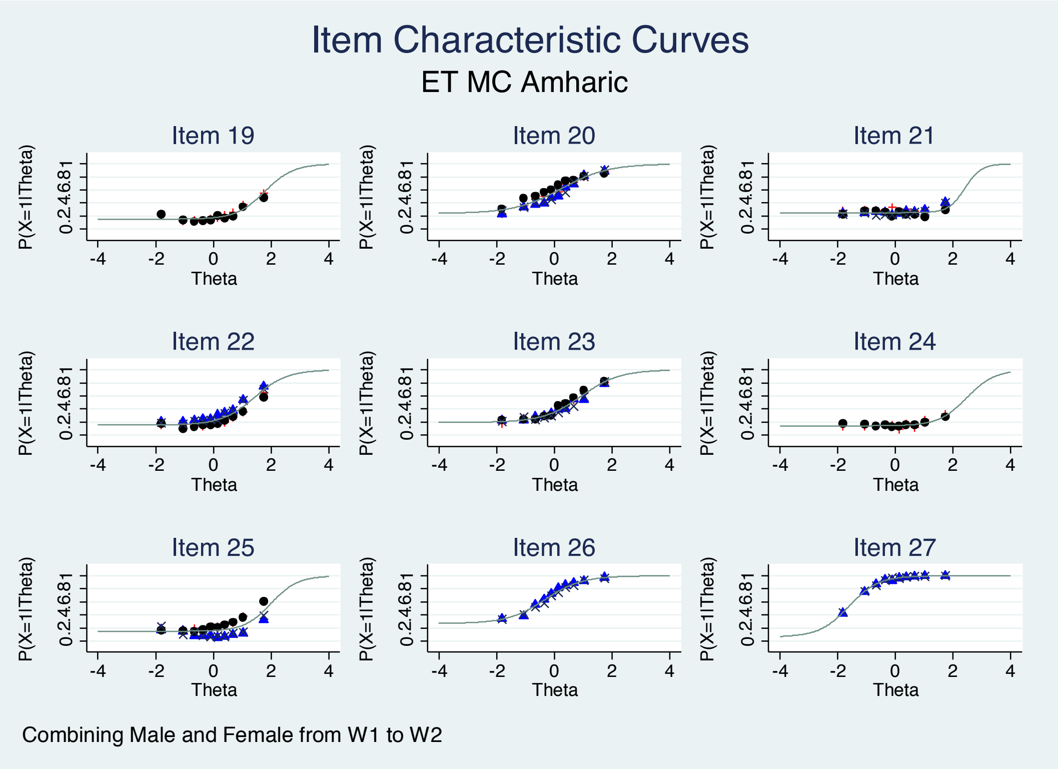
Appendix D: DIF analysis performed by language and area (pooled sample)

Maths

Figure D1. *Item DIF analysis by groups for Amharic*




Item Characteristic Curves
ET MC Amharic
Item 10
Item 11
Item 12
Item 13
Item 14
Item 15
Item 16
Item 17
Item 18
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta
Combining Male and Female from W1 to W2


Item Characteristic Curves
ET MC Amharic
Item 19
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
Item 26
Item 27
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta
Combining Male and Female from W1 to W2

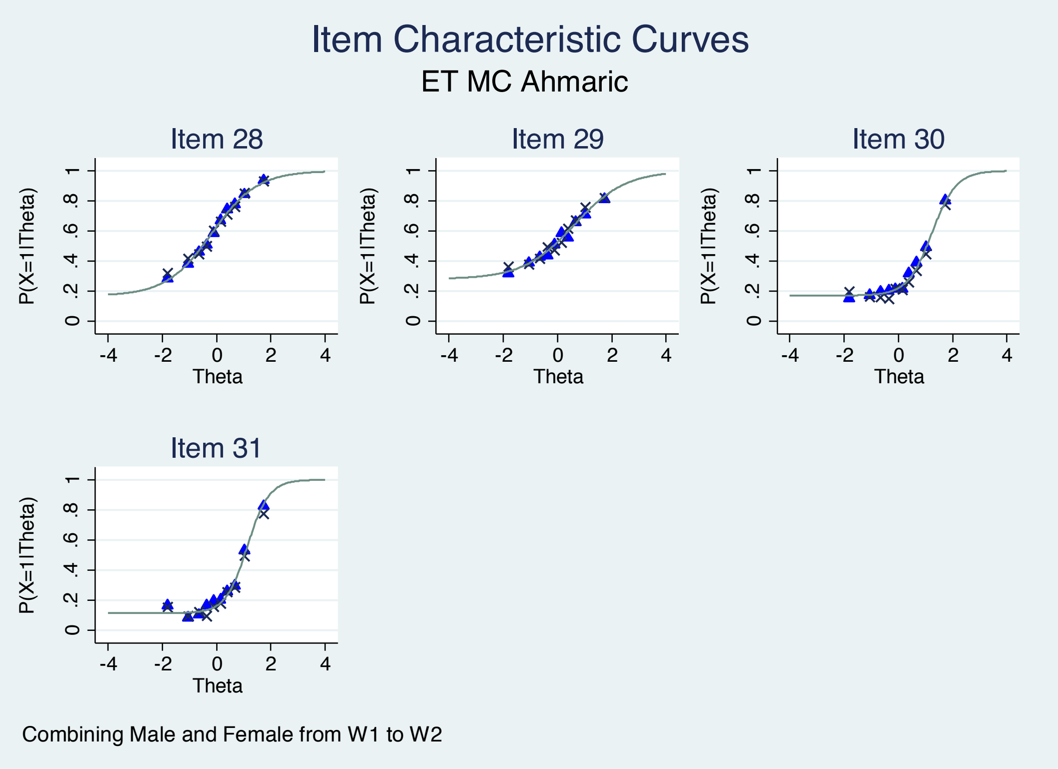
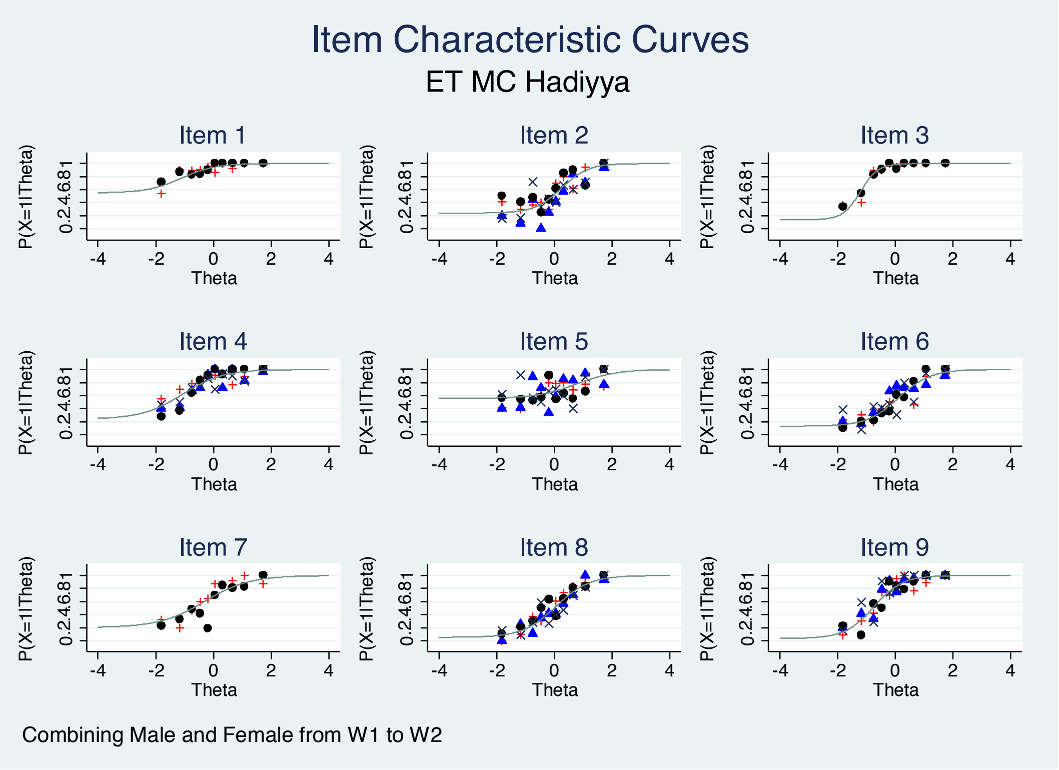
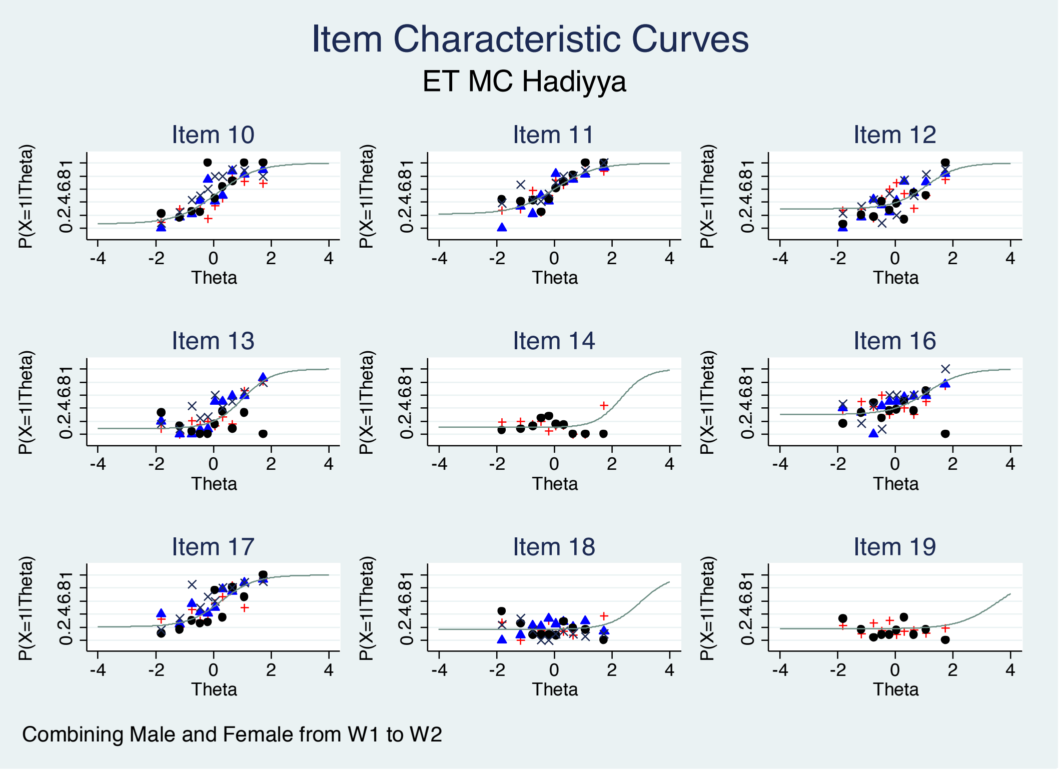

Item Characteristic Curves
ET MC Ahmaric
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0
.2
.4
.6
.8
1
-4
-2
0
2
4
Theta
Combining Male and Female from W1 to W2

Figure D2. *Item DIF analysis by groups for Hadiyya*





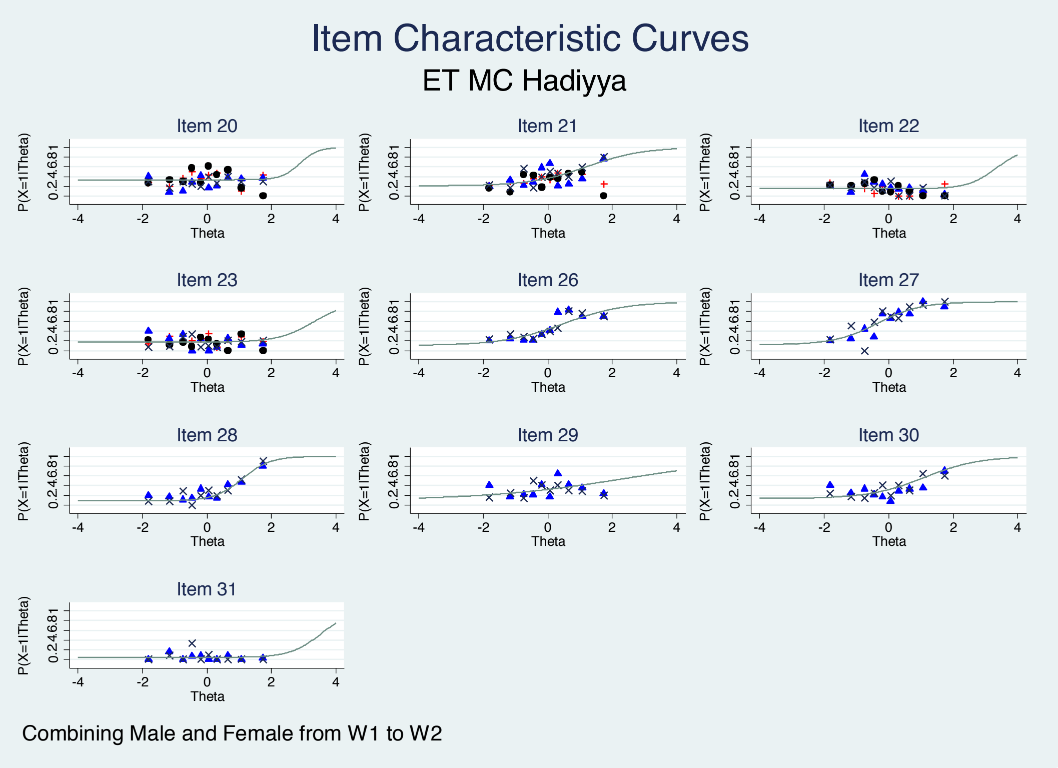
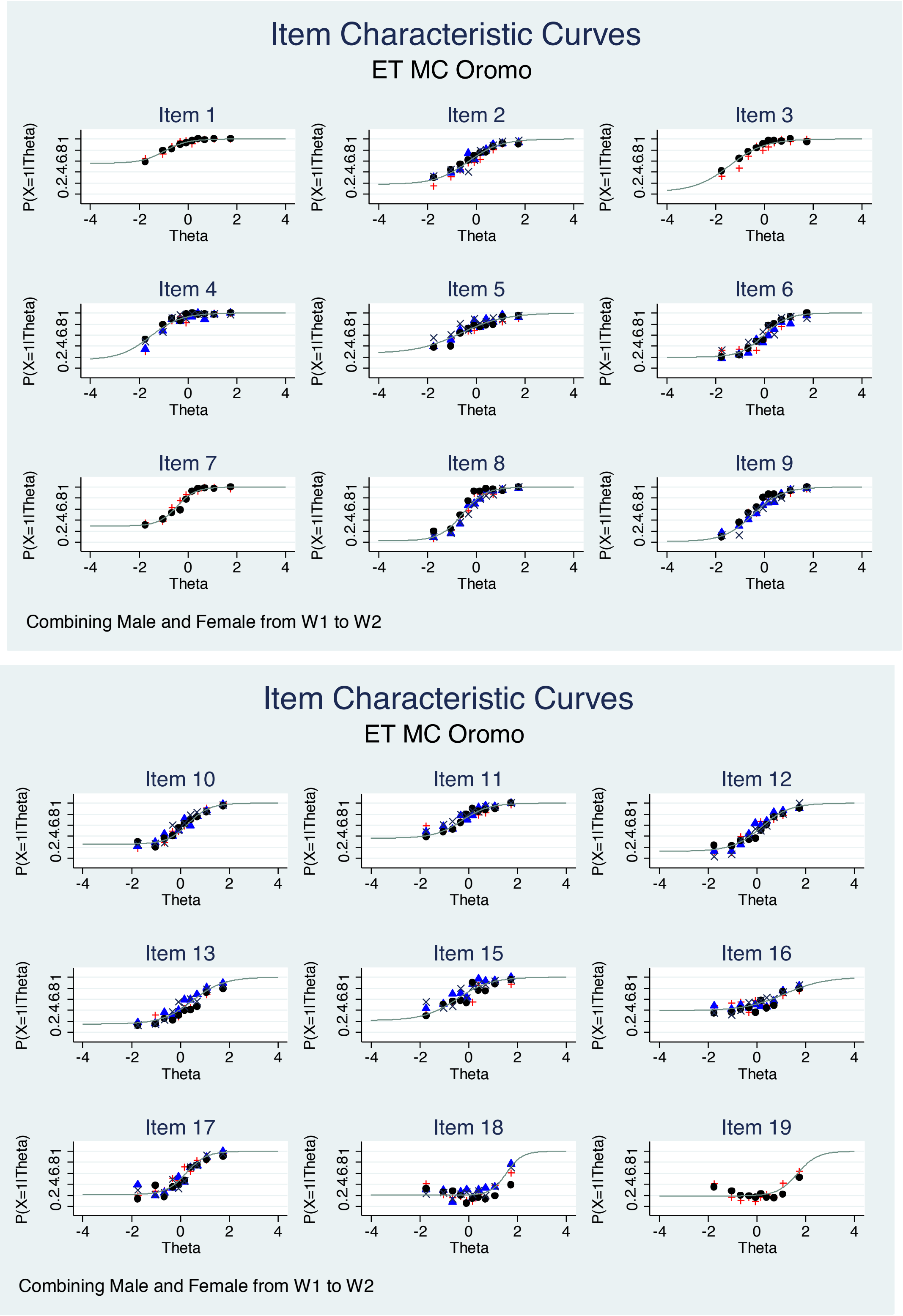

Item Characteristic Curves
ET MC Hadiyya
Item 20
Item 21
Item 22
Item 23
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4
-2
0
2
4
Theta
Combining Male and Female from W1 to W2

Figure D3. *Item DIF analysis by groups for Oromo*



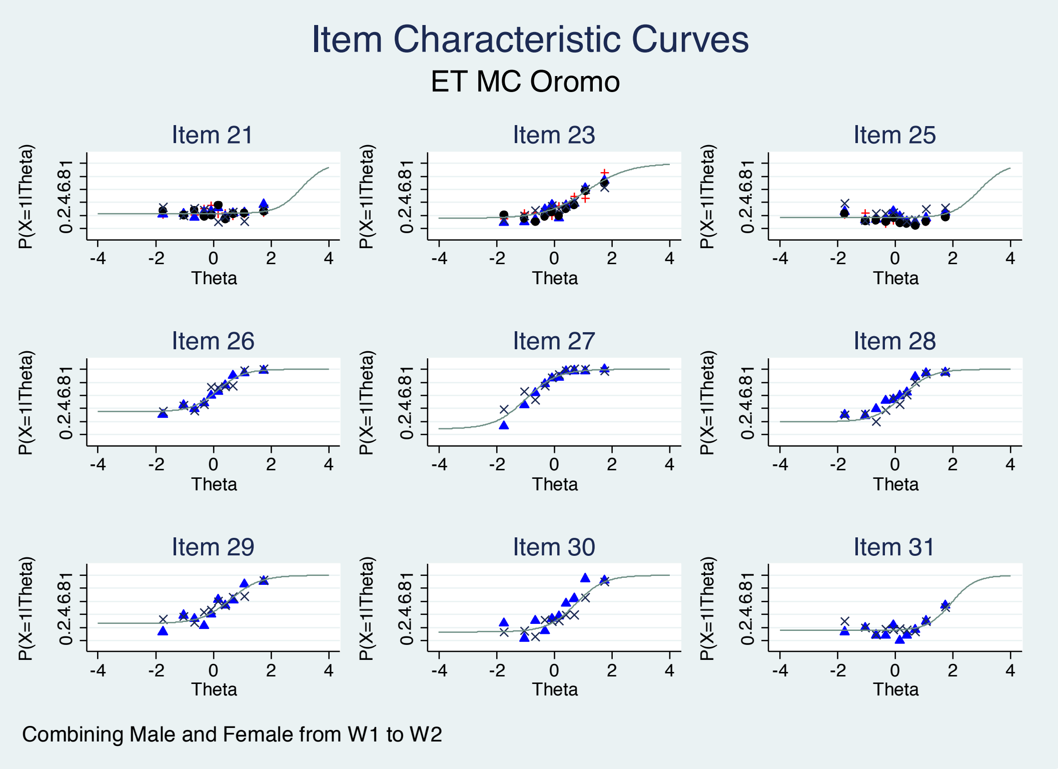
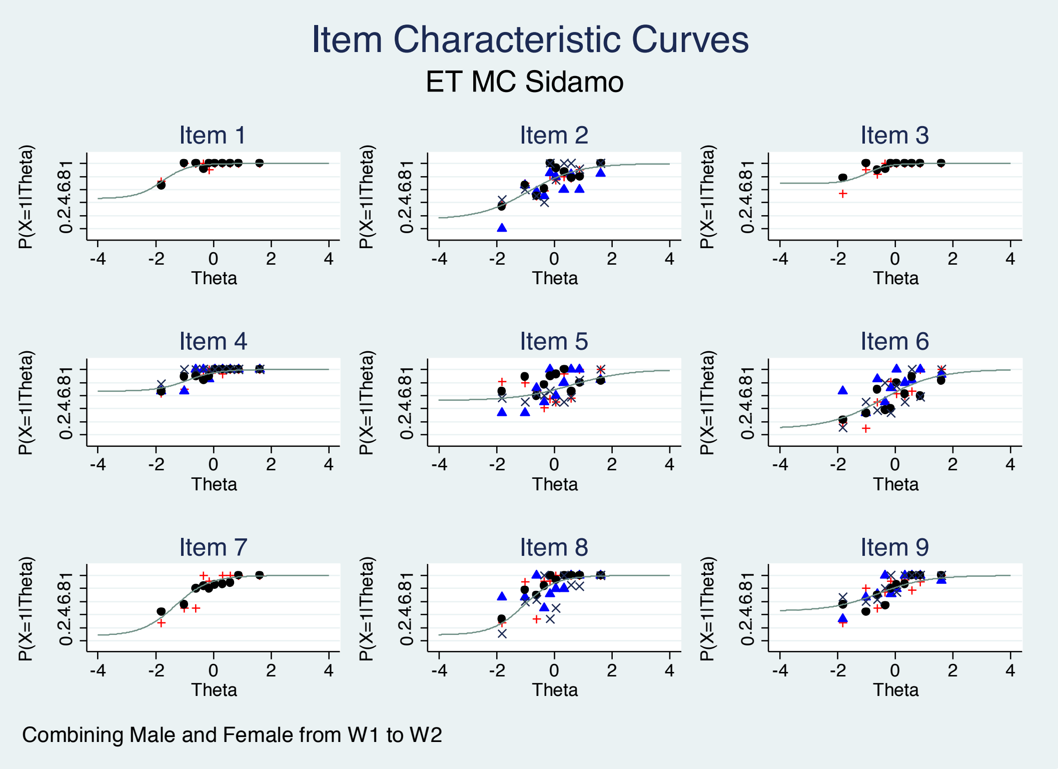
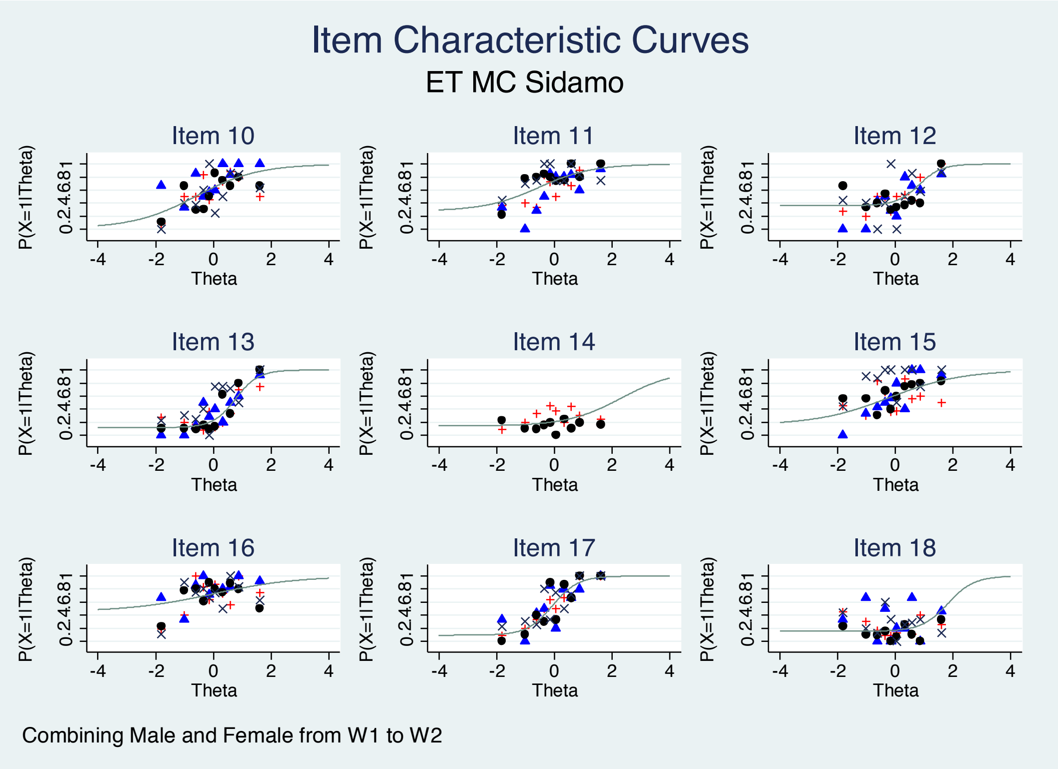

Item Characteristic Curves
ET MC Oromo
Item 21
Item 23
Item 25
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta
Combining Male and Female from W1 to W2

Figure D4. *Item DIF analysis by groups for Sidamo*



Combining Male and Female from W1 to W2



Combining Male and Female from W1 to W2

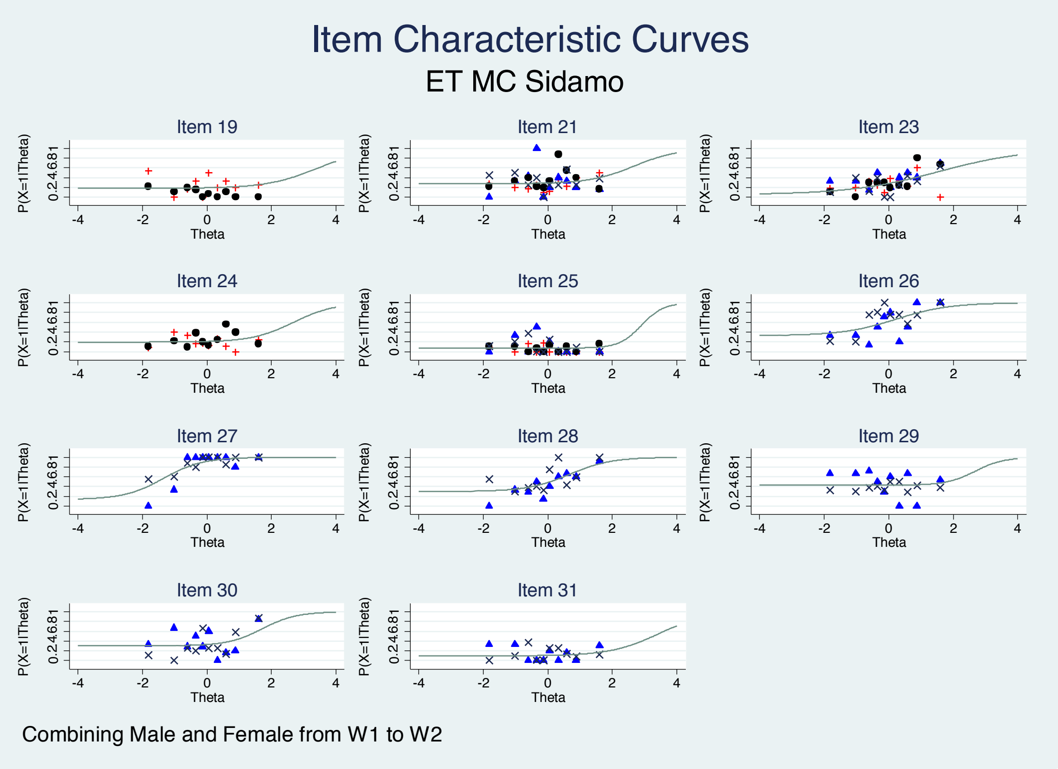
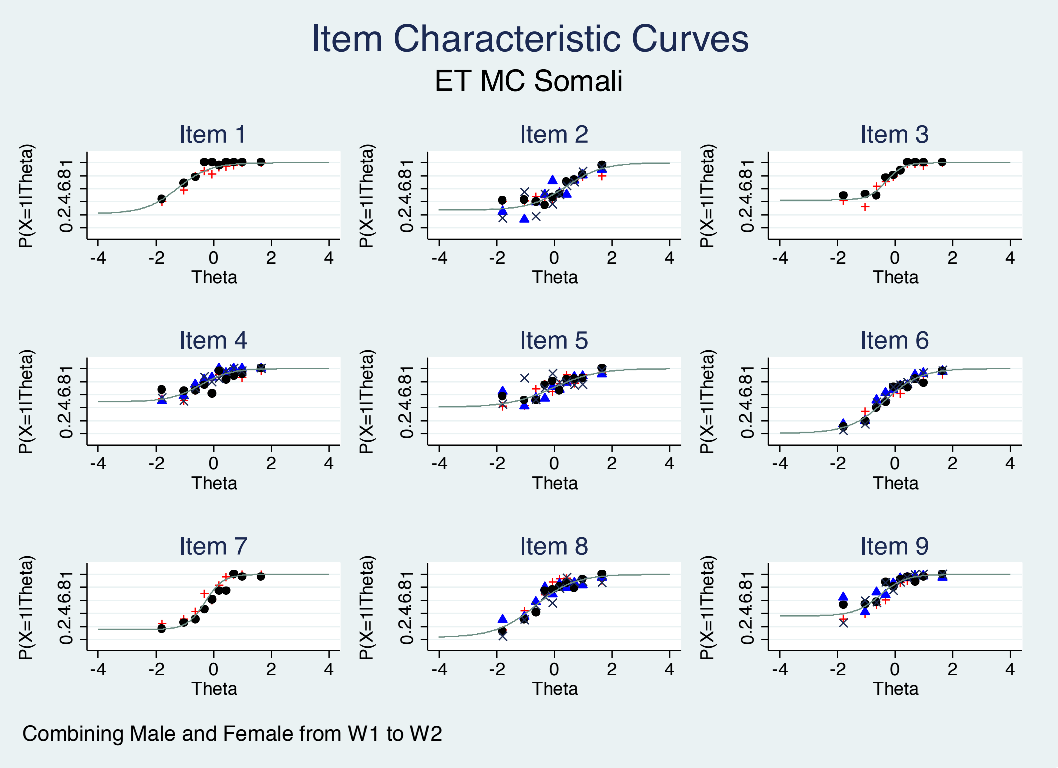
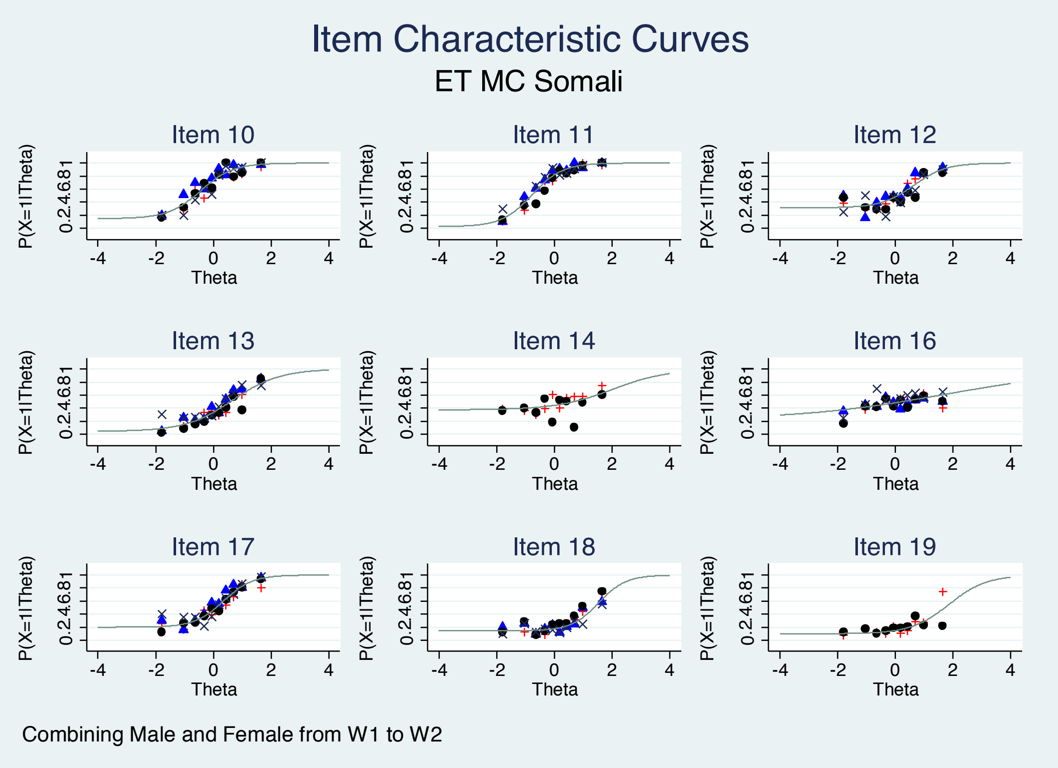

Item Characteristic Curves
ET MC Sidamo
Item 19
Item 21
Item 23
Item 24
Item 25
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4
-2
0
2
4
Theta
Combining Male and Female from W1 to W2

Figure D5. *Item DIF analysis by groups for Somali*





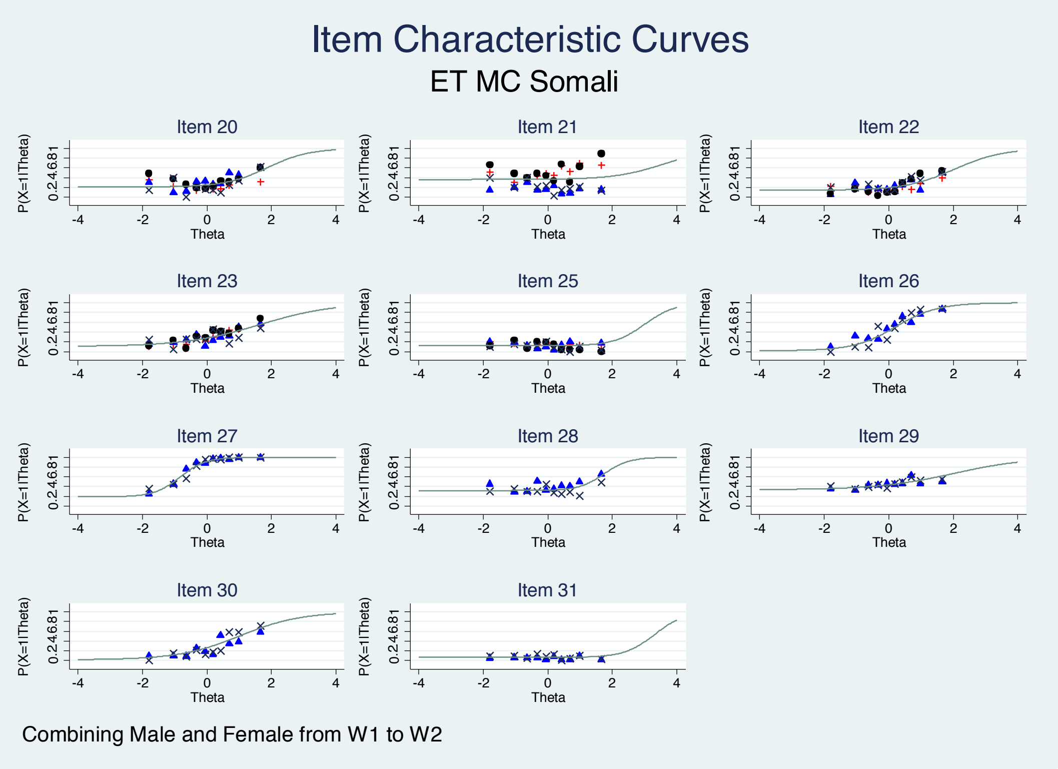
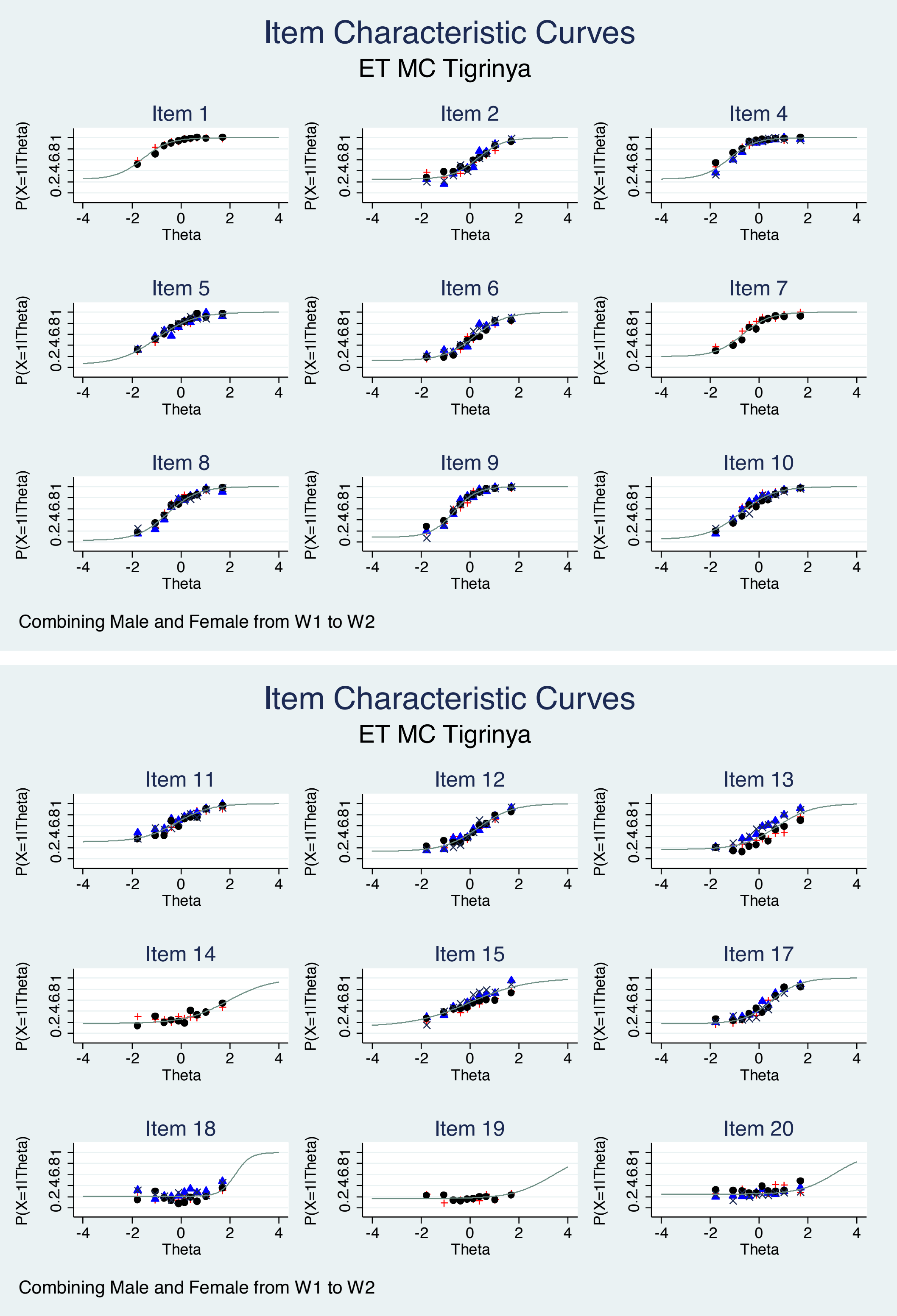

Item Characteristic Curves
ET MC Somali
Item 20
Item 21
Item 22
Item 23
Item 25
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4
-2
0
2
4
Theta
Combining Male and Female from W1 to W2

Figure D6. *Item DIF analysis by groups for Tigrinya*



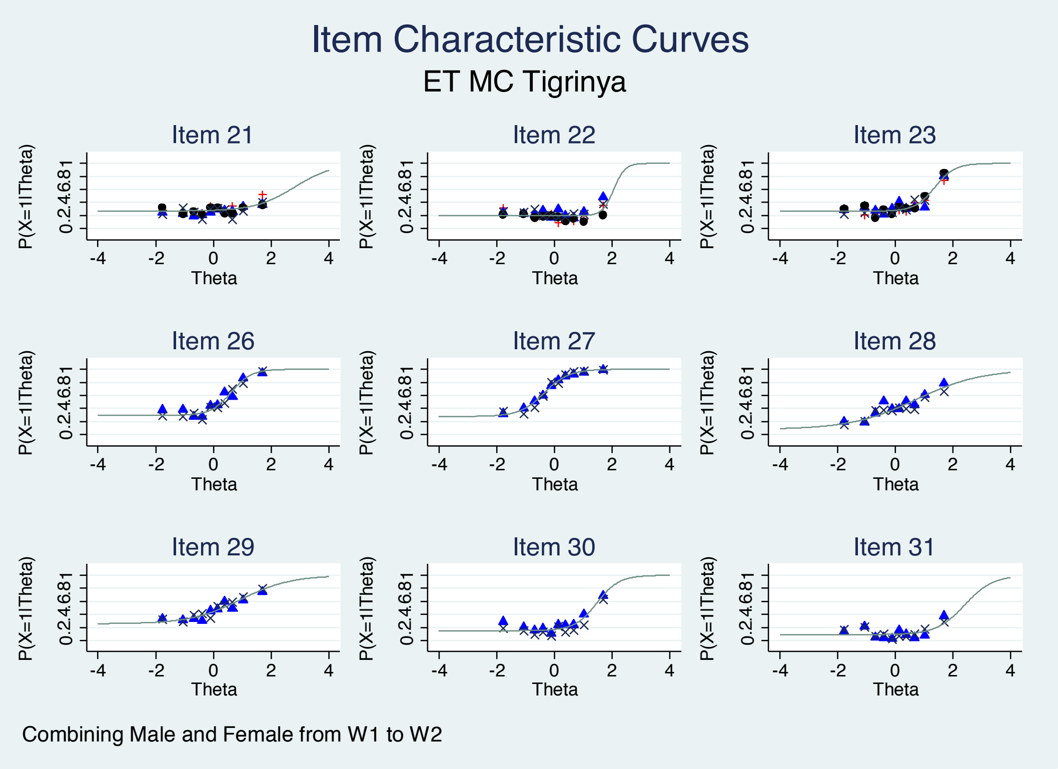
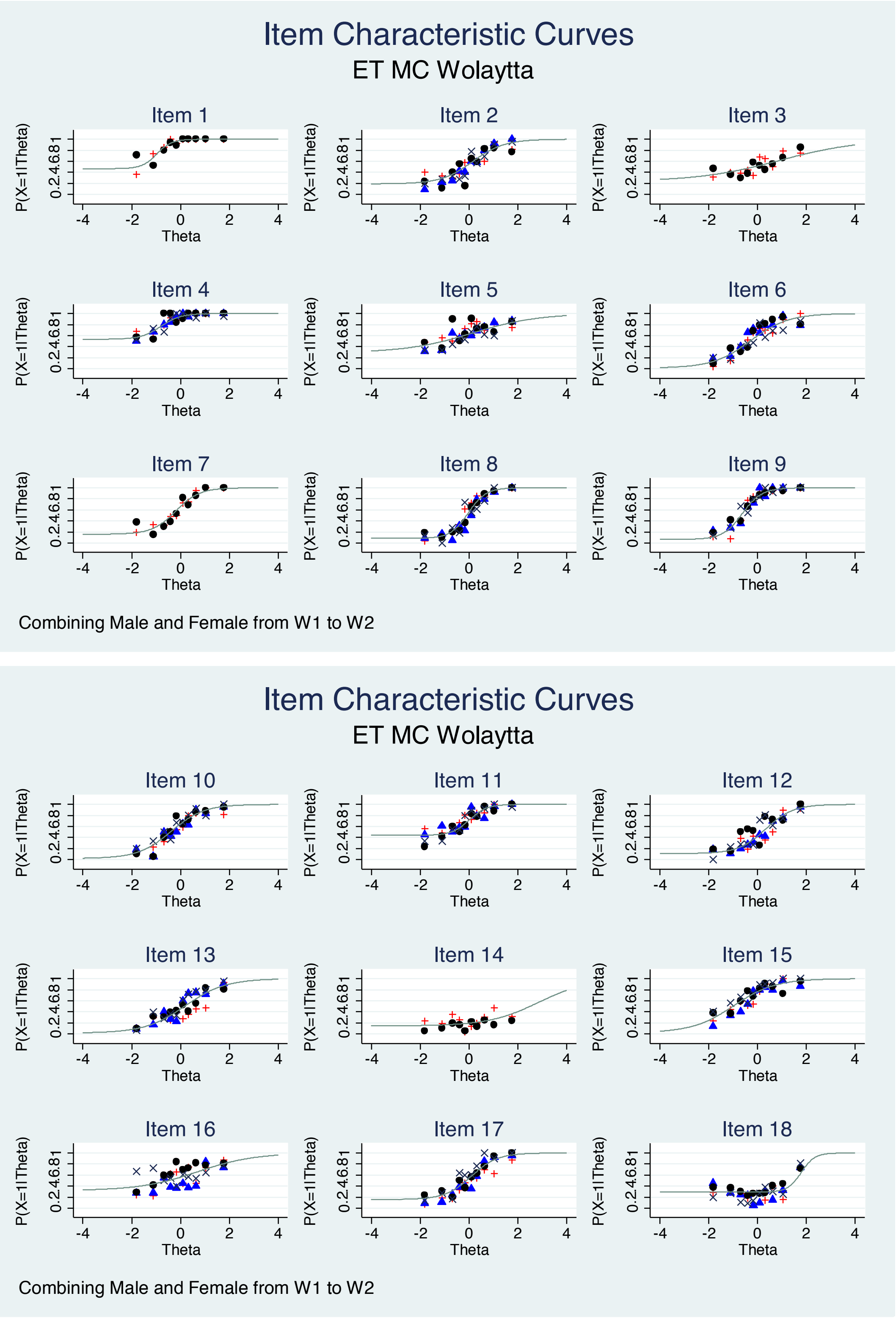
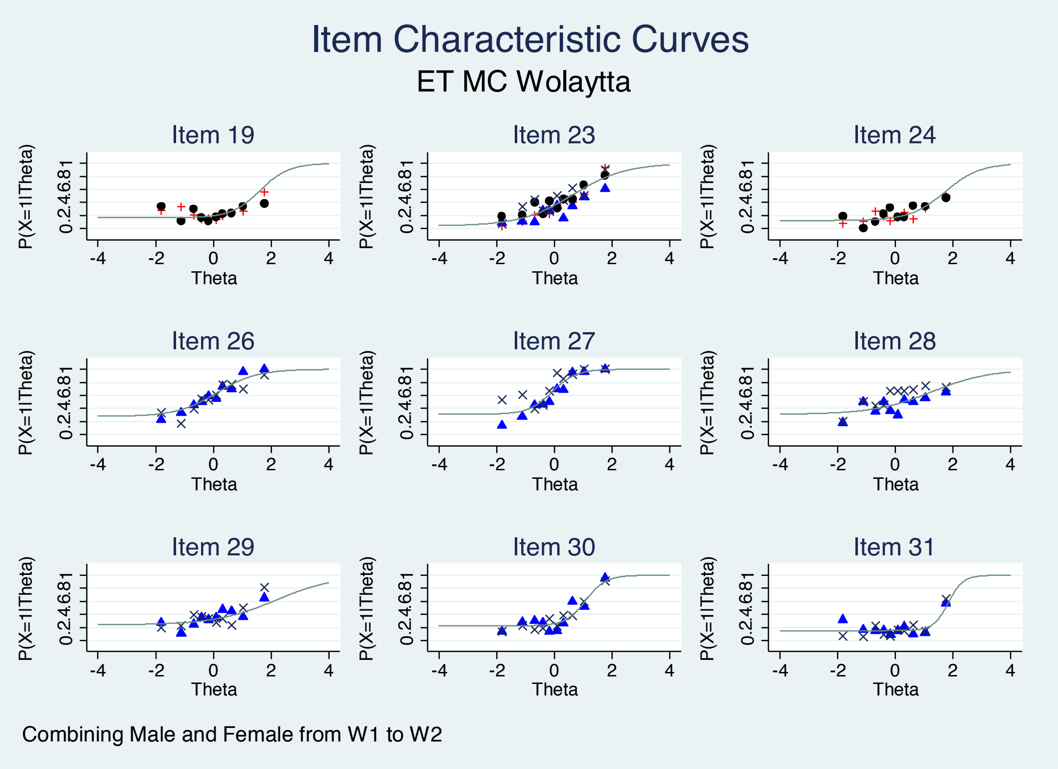

Item Characteristic Curves
ET MC Tigrinya
Item 21
Item 22
Item 23
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta
Combining Male and Female from W1 to W2

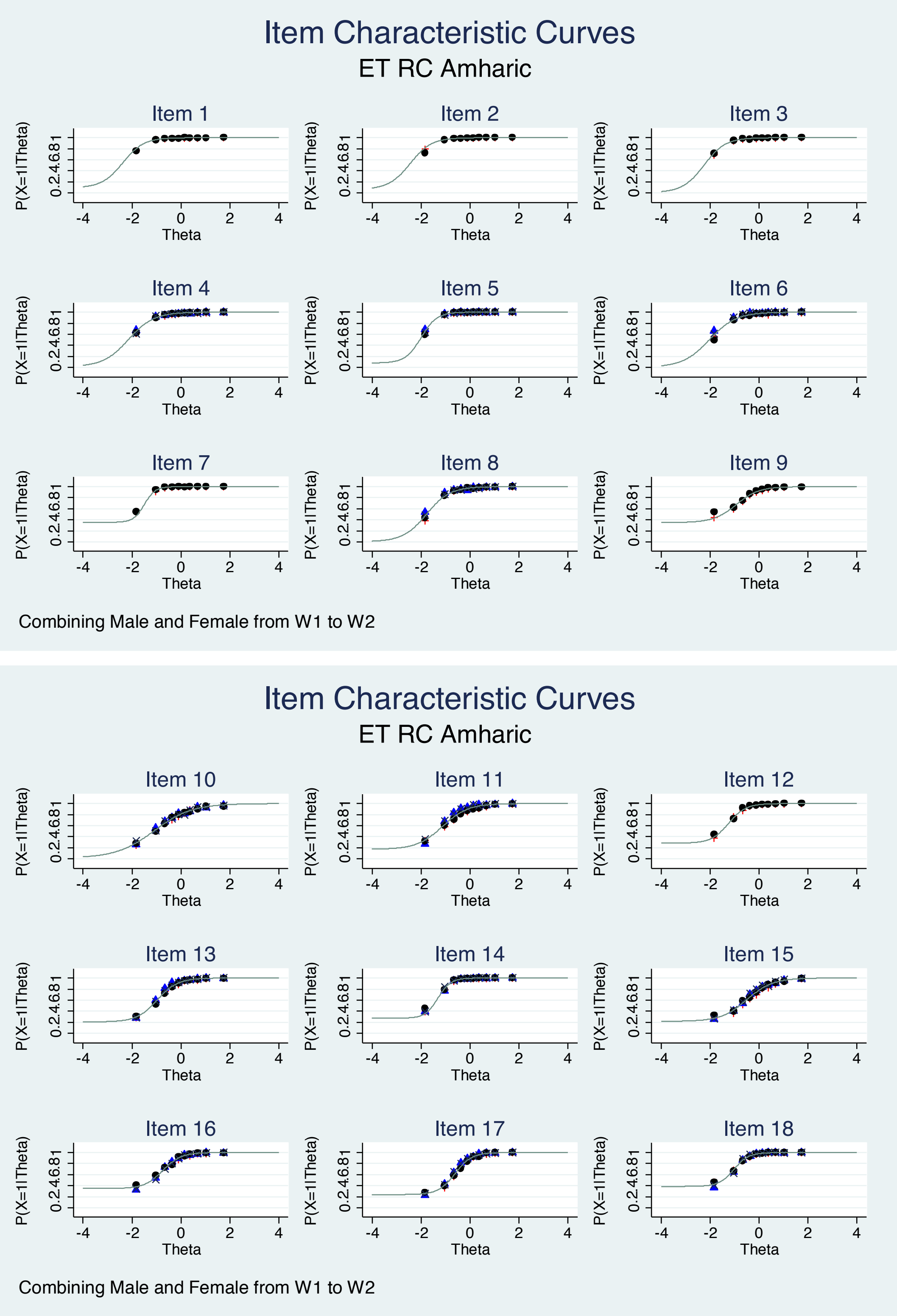
Figure D7. *Item DIF analysis by groups for Wolaytta*




Item Characteristic Curves
ET MC Wolaytta
Item 19
Item 23
Item 24
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta
Combining Male and Female from W1 to W2

Reading comprehension

Figure D8. *Item DIF analysis by groups for Amharic*



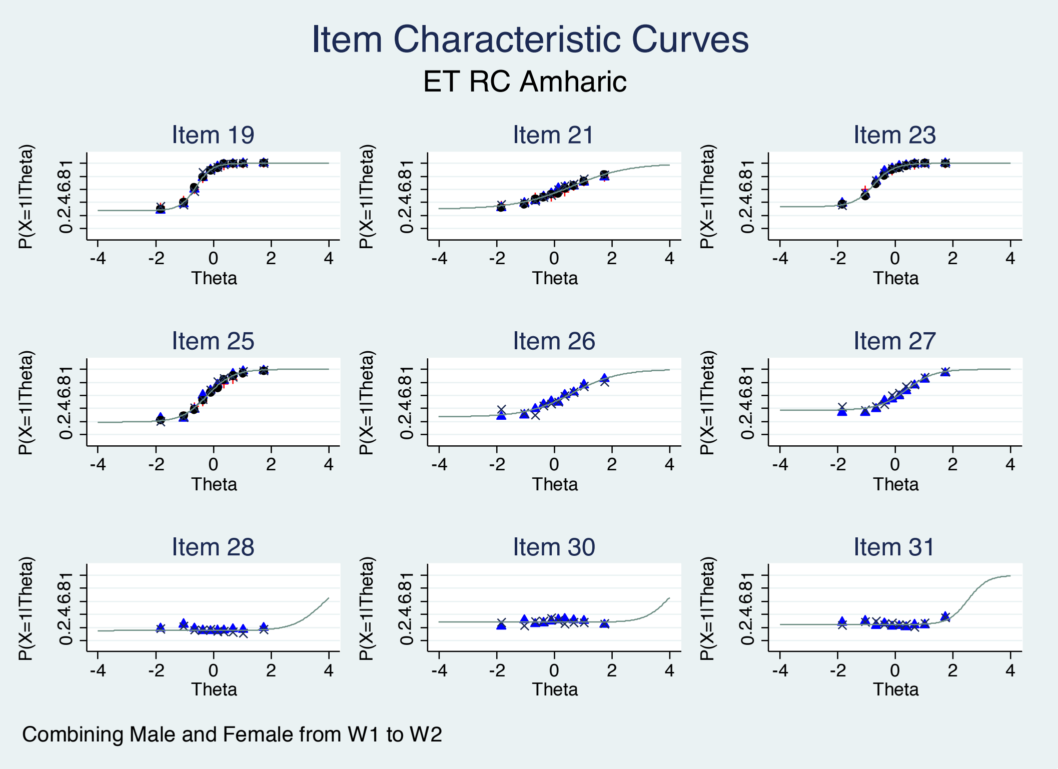
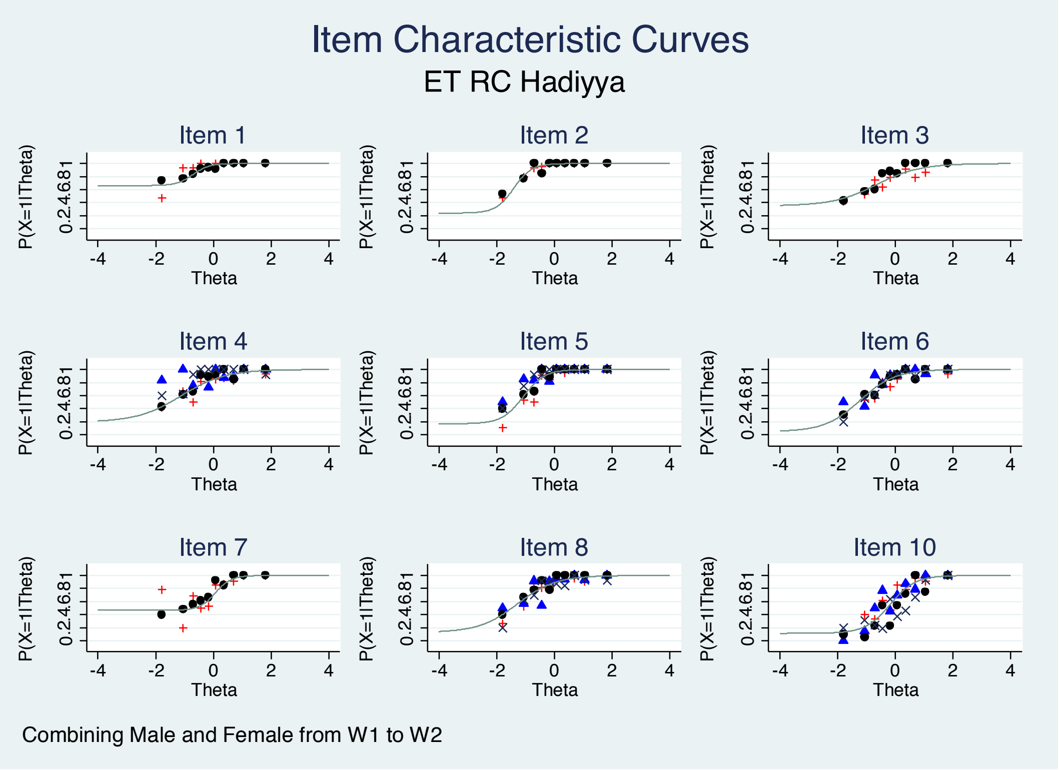
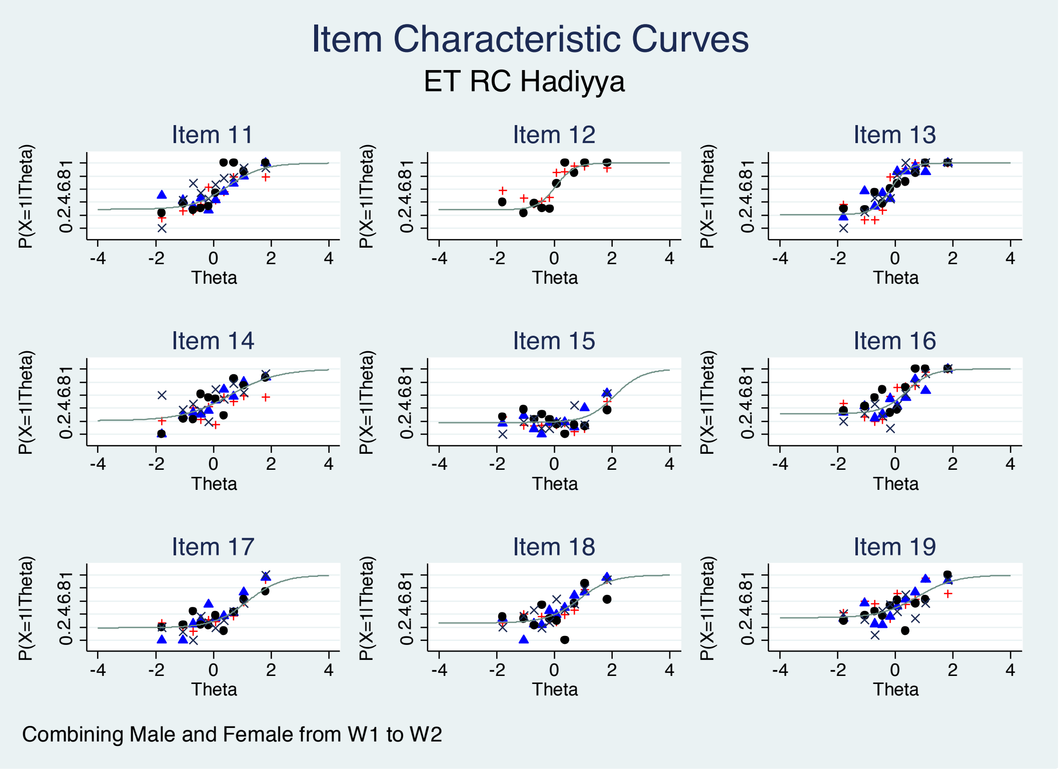

Item Characteristic Curves
ET RC Amharic
Item 19
Item 21
Item 23
Item 25
Item 26
Item 27
Item 28
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4
-2
0
2
4
Theta
Combining Male and Female from W1 to W2

Figure D9. *Item DIF analysis by groups for Hadiyya*





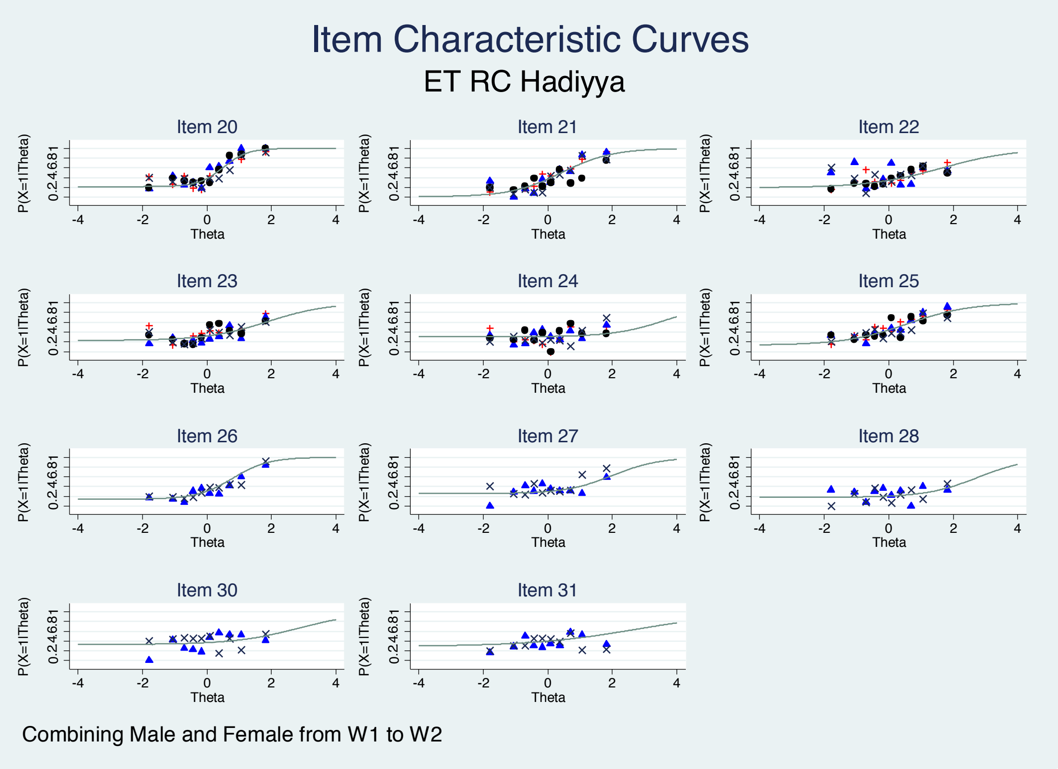
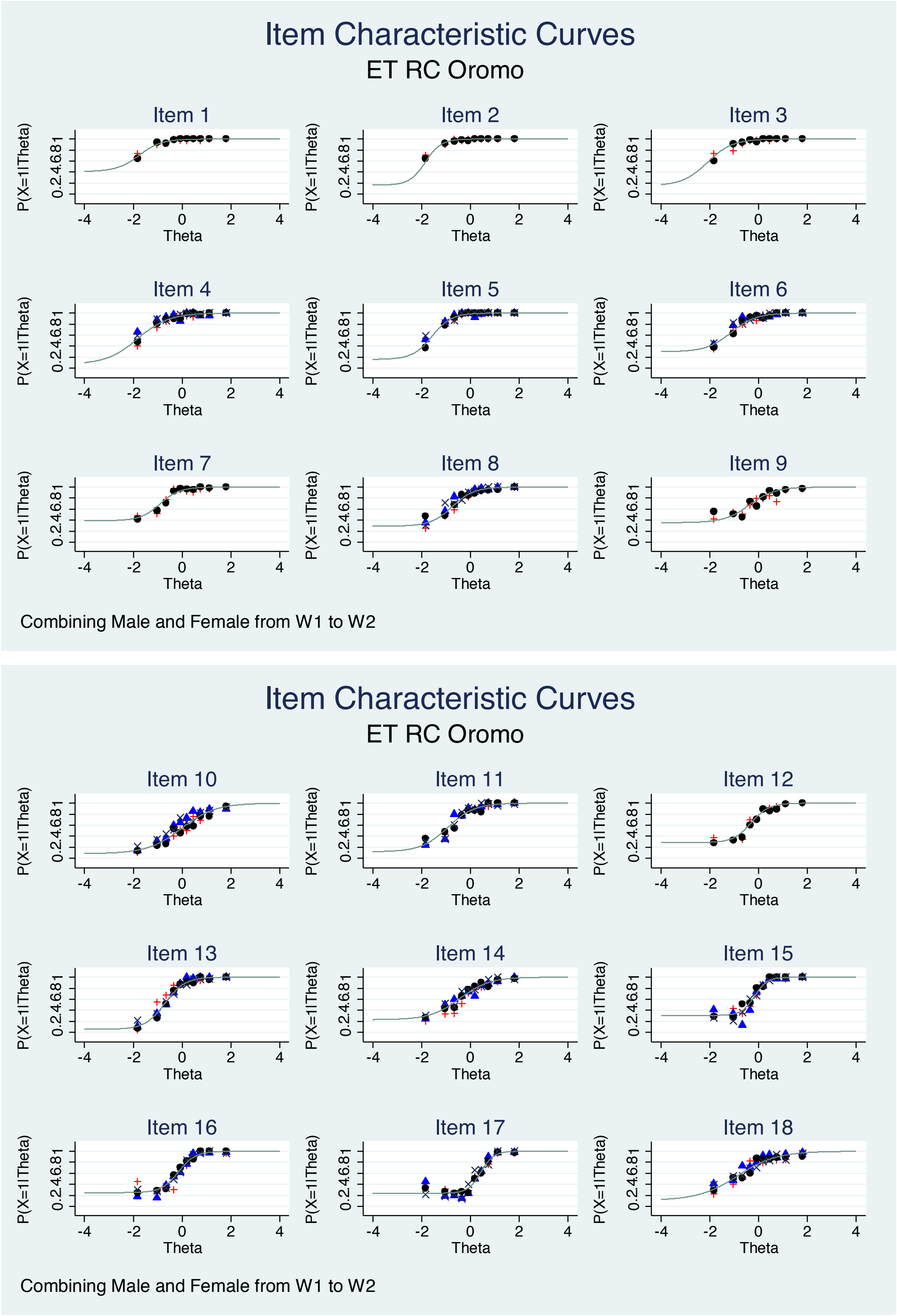

Item Characteristic Curves
ET RC Hadiyya
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
Item 26
Item 27
Item 28
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4
-2
0
2
4
Theta
Combining Male and Female from W1 to W2

Figure D10. *Item DIF analysis by groups for Oromo*



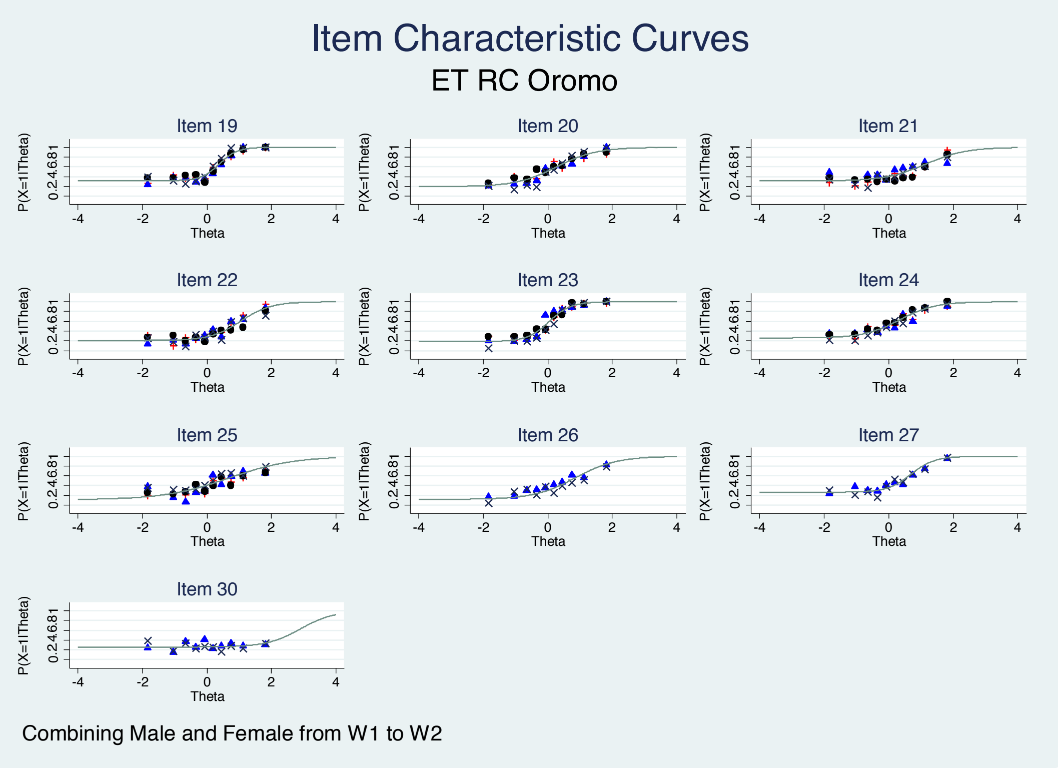
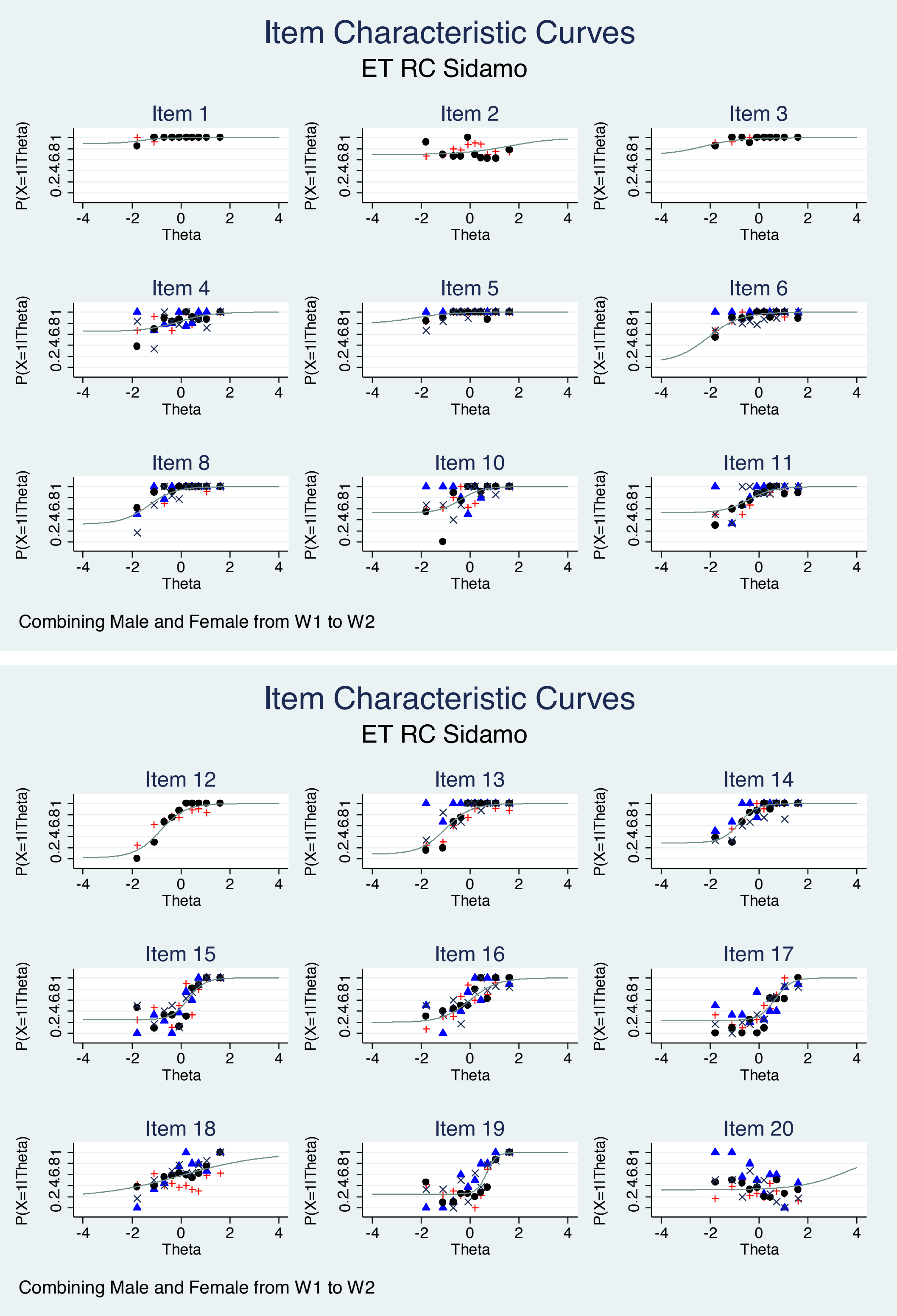

Item Characteristic Curves
ET RC Oromo
Item 19
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
Item 26
Item 27
Item 30
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4 -2 0 2 4
Theta
Combining Male and Female from W1 to W2

Figure D11. *Item DIF analysis by groups for Sidamo*



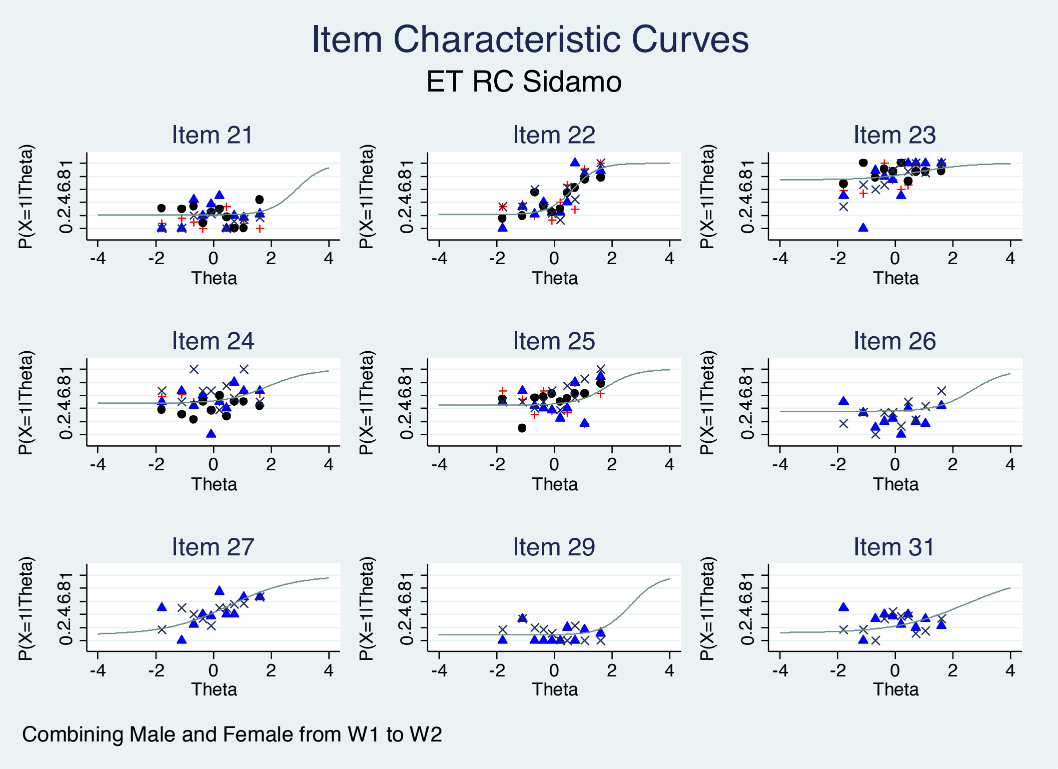
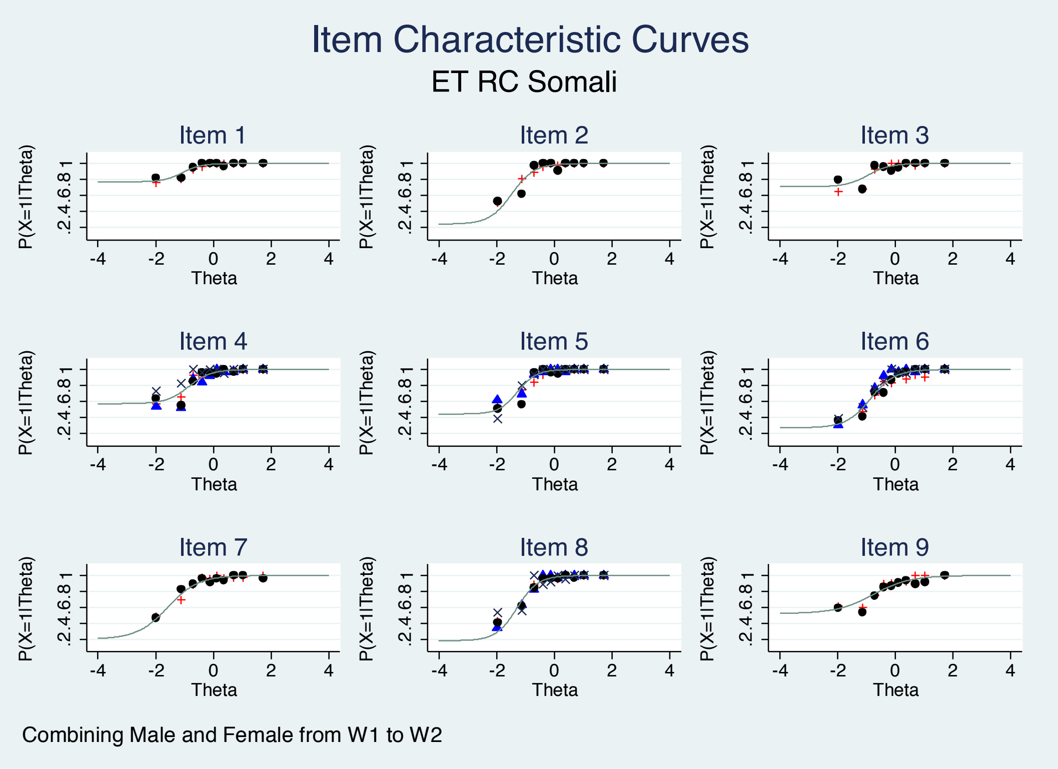
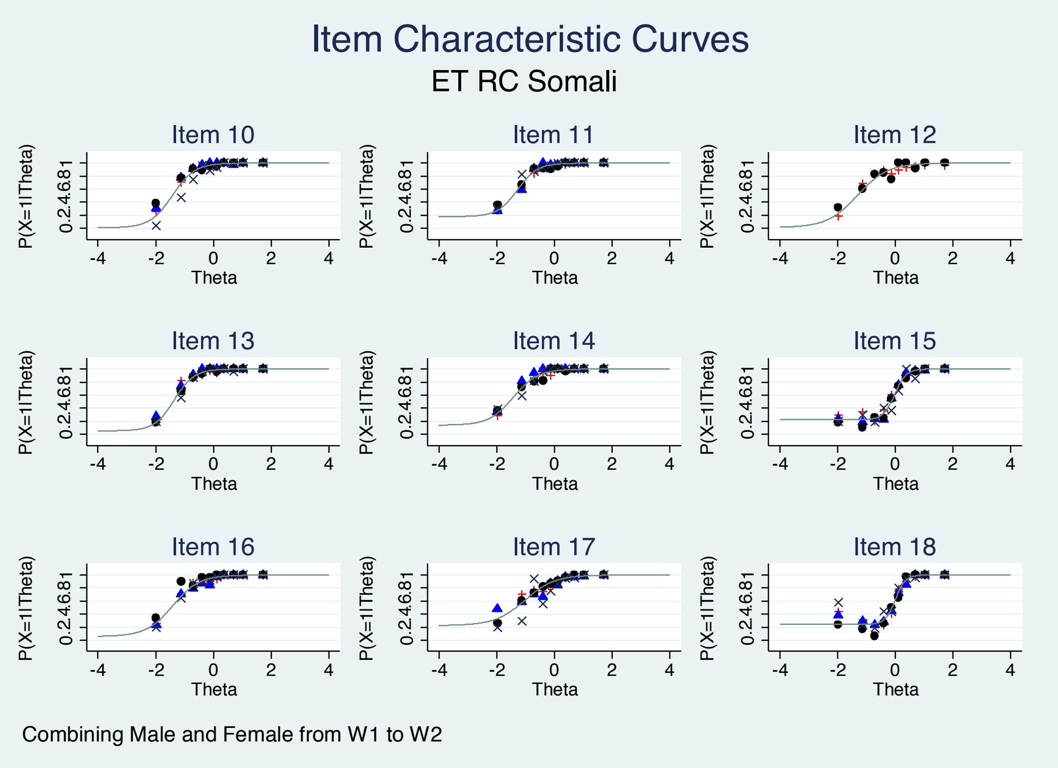

Item Characteristic Curves
ET RC Sidamo
Item 21
Item 22
Item 23
Item 24
Item 25
Item 26
Item 27
Item 29
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4
-2
0
2
4
Theta
Combining Male and Female from W1 to W2

Figure D12. *Item DIF analysis by groups for Somali*





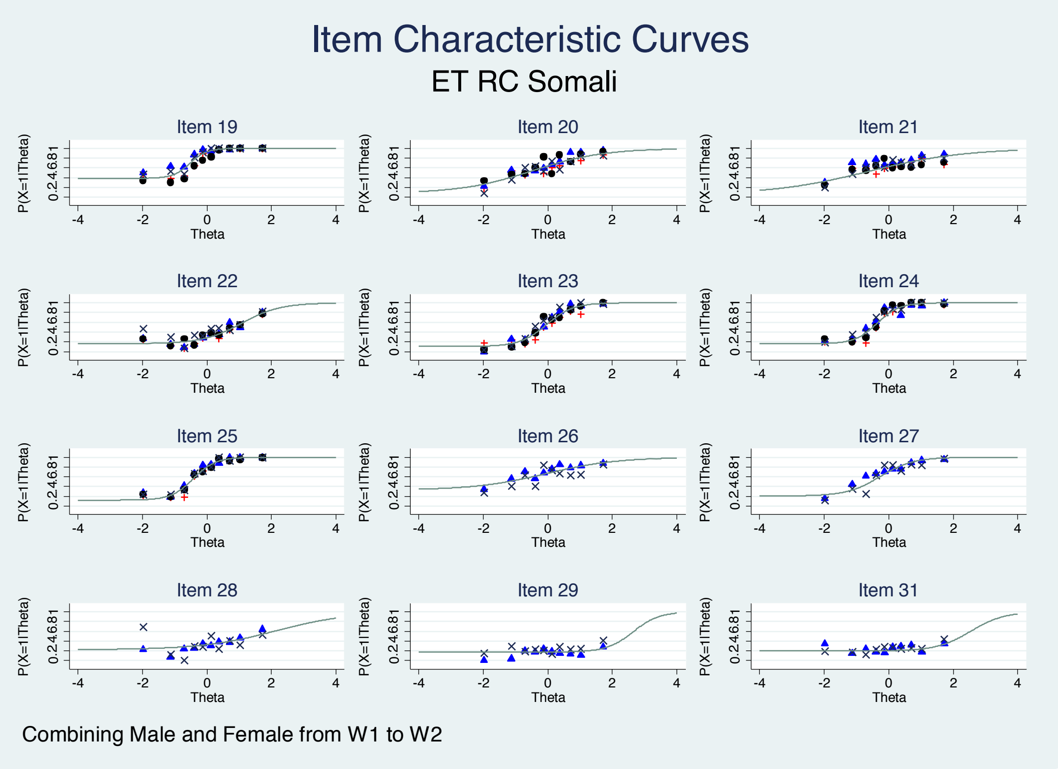
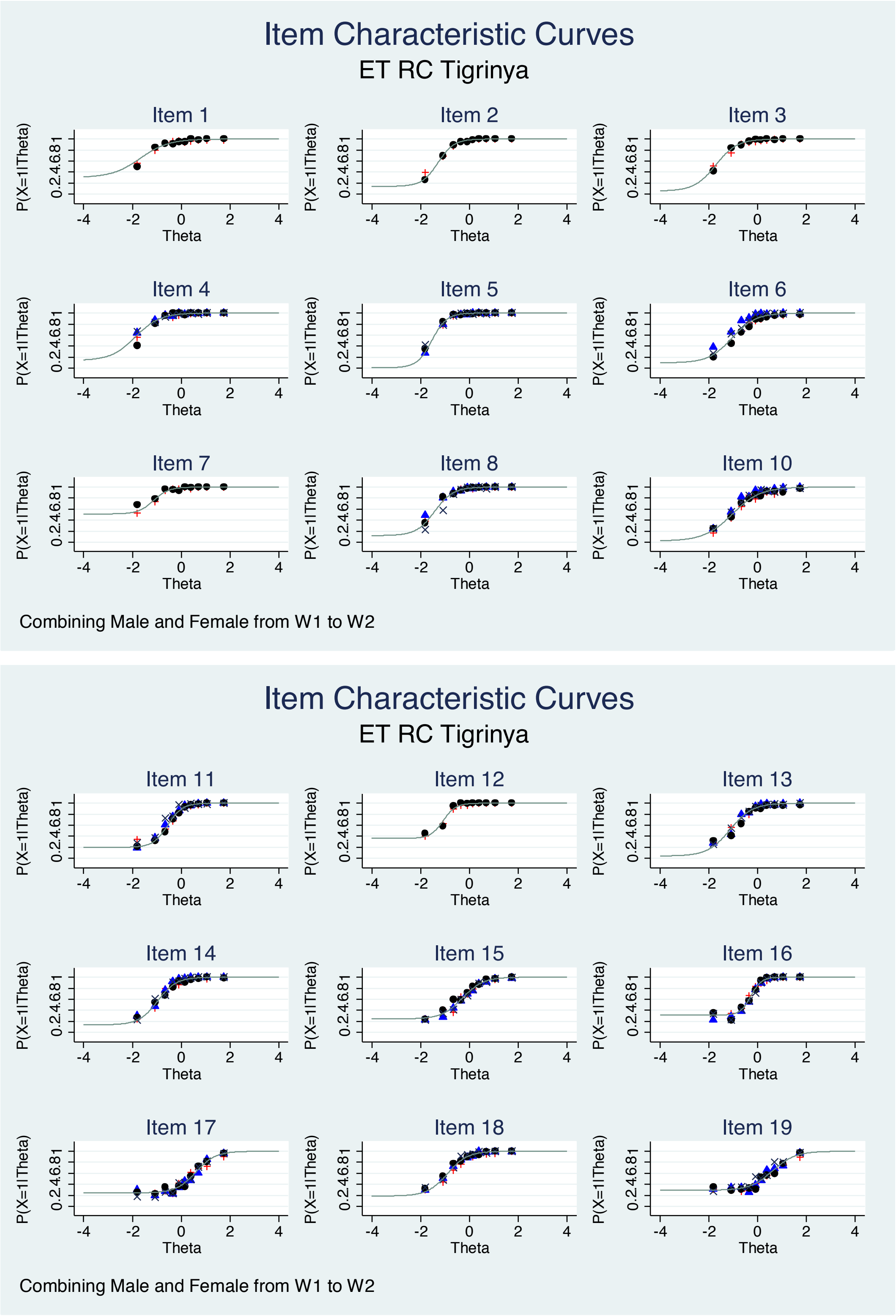

Item Characteristic Curves
ET RC Somali
Item 19
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
Item 26
Item 27
Item 28
Item 29
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4
-2
0
2
4
Theta
Combining Male and Female from W1 to W2

Figure D13. *Item DIF analysis by groups for Tigrinya*



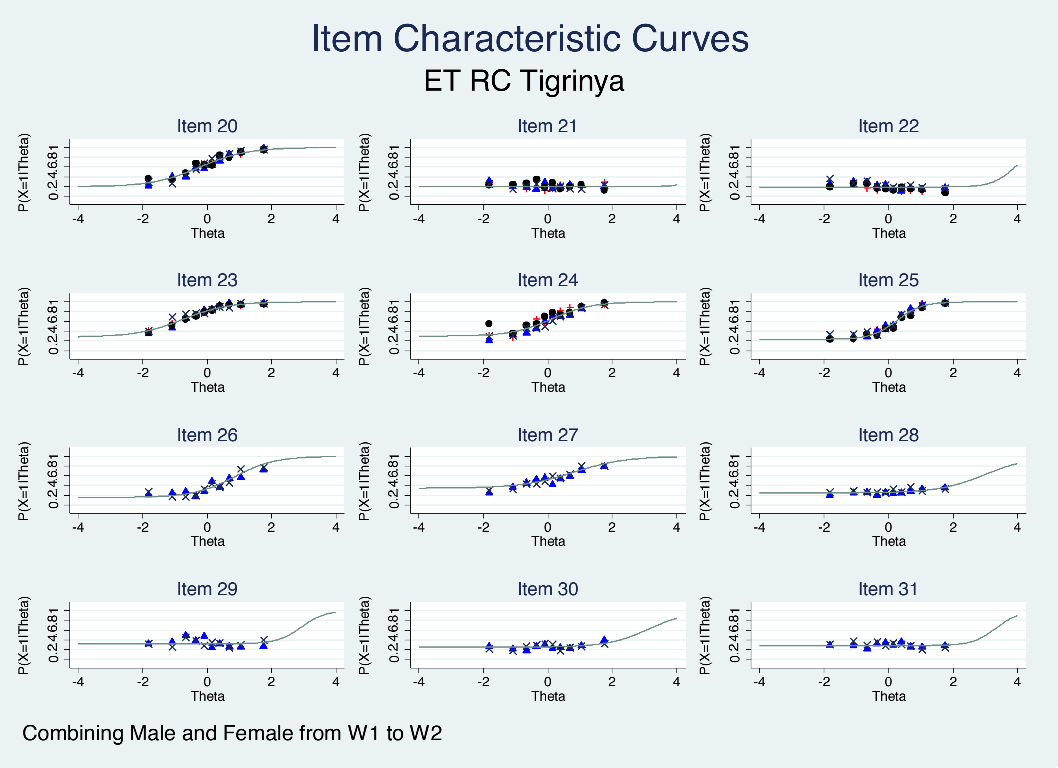
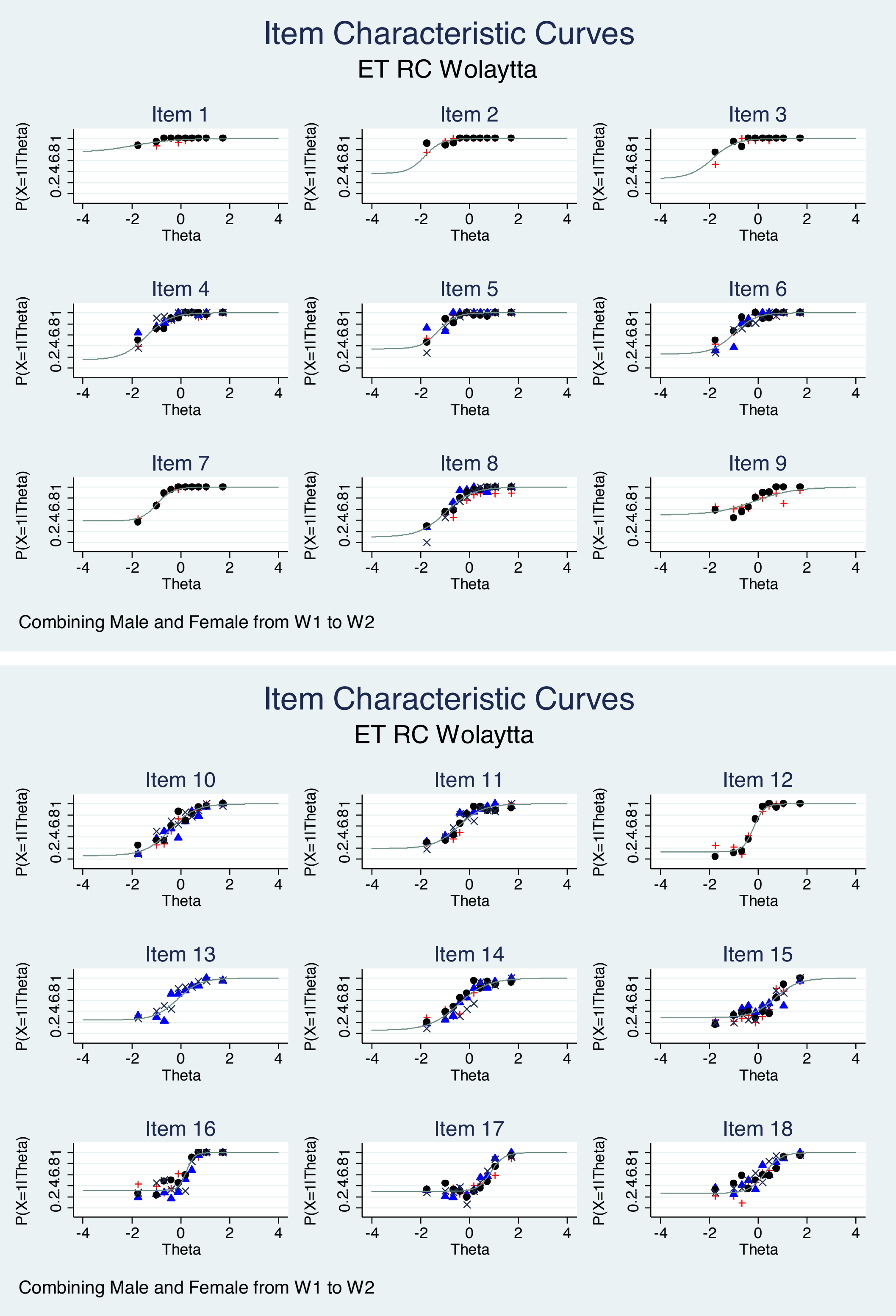

Item Characteristic Curves
ET RC Tigrinya
Item 20
Item 21
Item 22
Item 23
Item 24
Item 25
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0 .2 .4 .6 .8 1
-4
-2
0
2
4
Theta
Combining Male and Female from W1 to W2

Figure D14. *Item DIF analysis by groups for Wolaytta*



Item Characteristic Curves

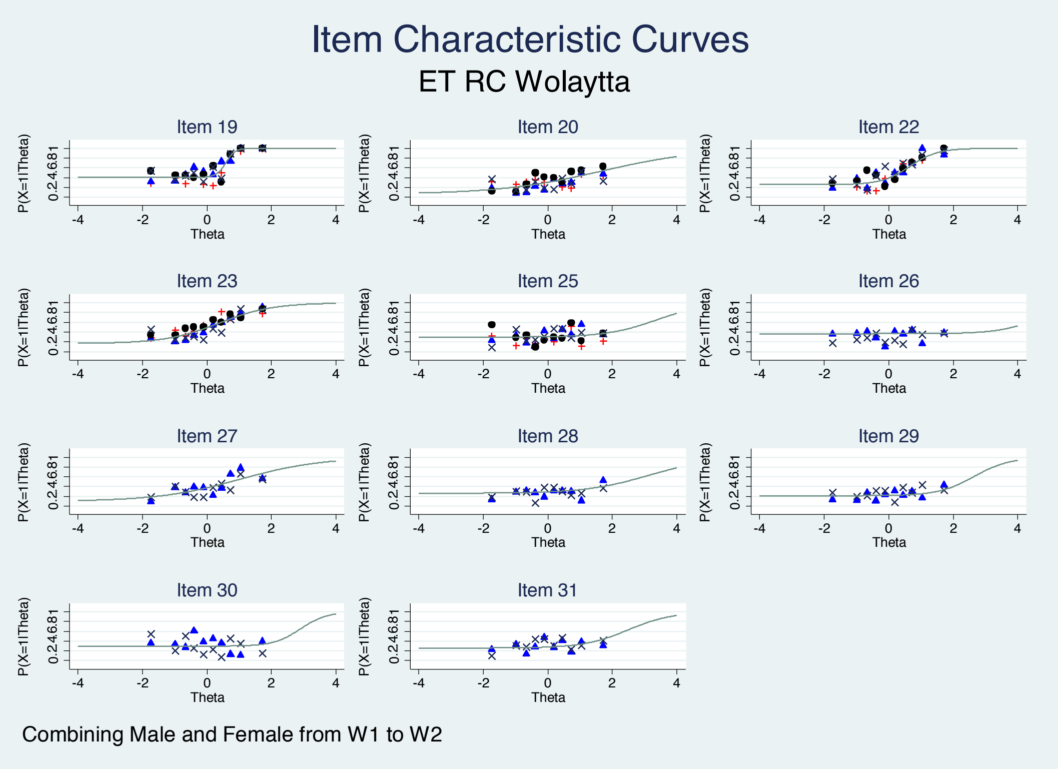
ET RC Wolaytta

Combining Male and Female from W1 to W2

Item Characteristic Curves

ET RC Wolaytta

Combining Male and Female from W1 to W2


Item Characteristic Curves
ET RC Wolaytta
Item 19
Item 20
Item 22
Item 23
Item 25
Item 26
Item 27
Item 28
Item 29
Item 30
Item 31
P(X=1|Theta)
0.2.4.6.81
-4
-2
0
2
4
Theta
Combining Male and Female from W1 to W2

Appendix E: Original and corrected IRT scores distribution

Maths

Figure E1. *Original and corrected IRT score distribution for Amharic*

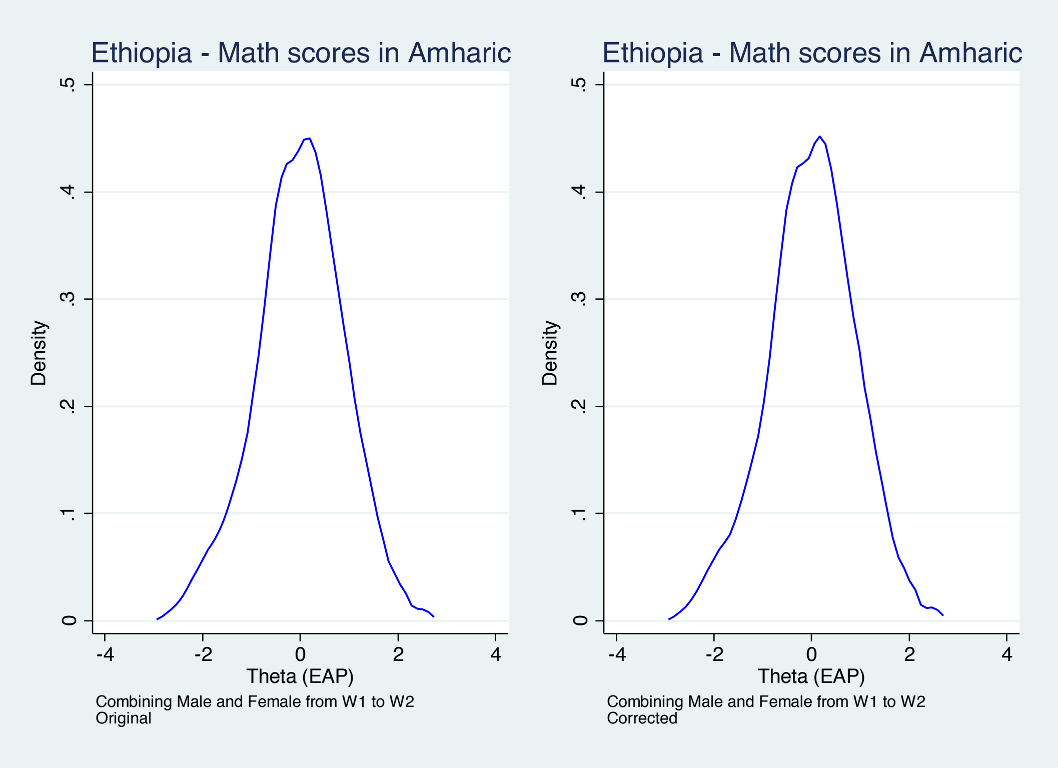


Figure E2. *Original and corrected IRT score distribution for Hadiyya*

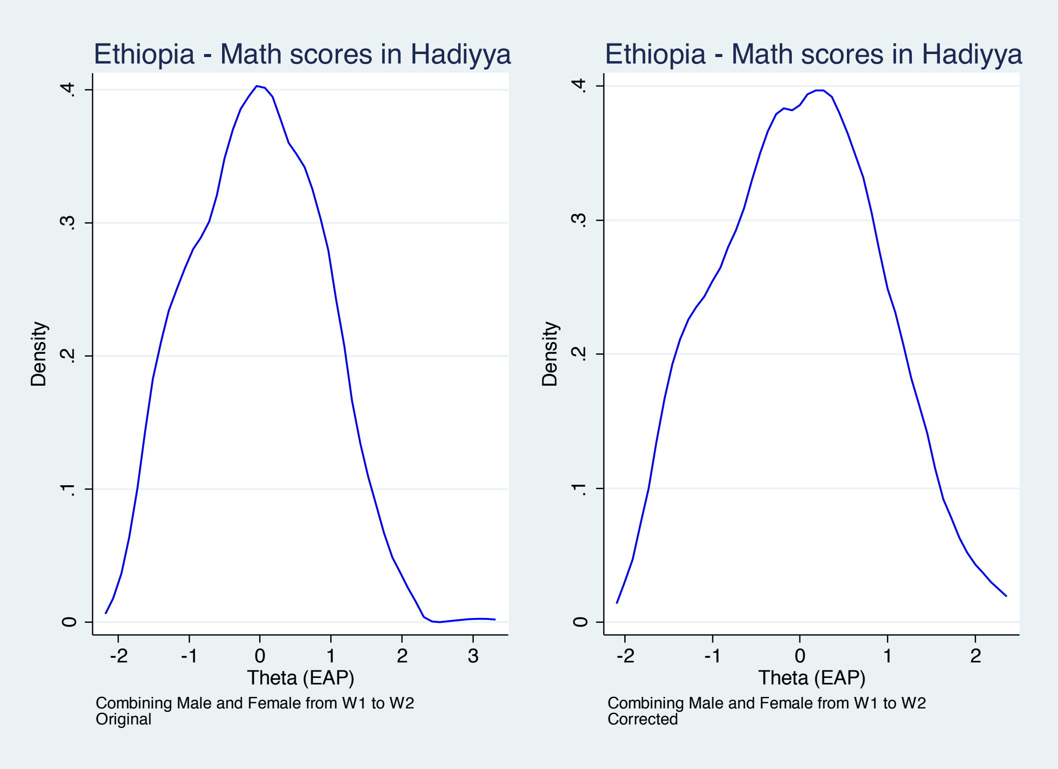


Figure E3. *Original and corrected IRT score distribution for Oromo*

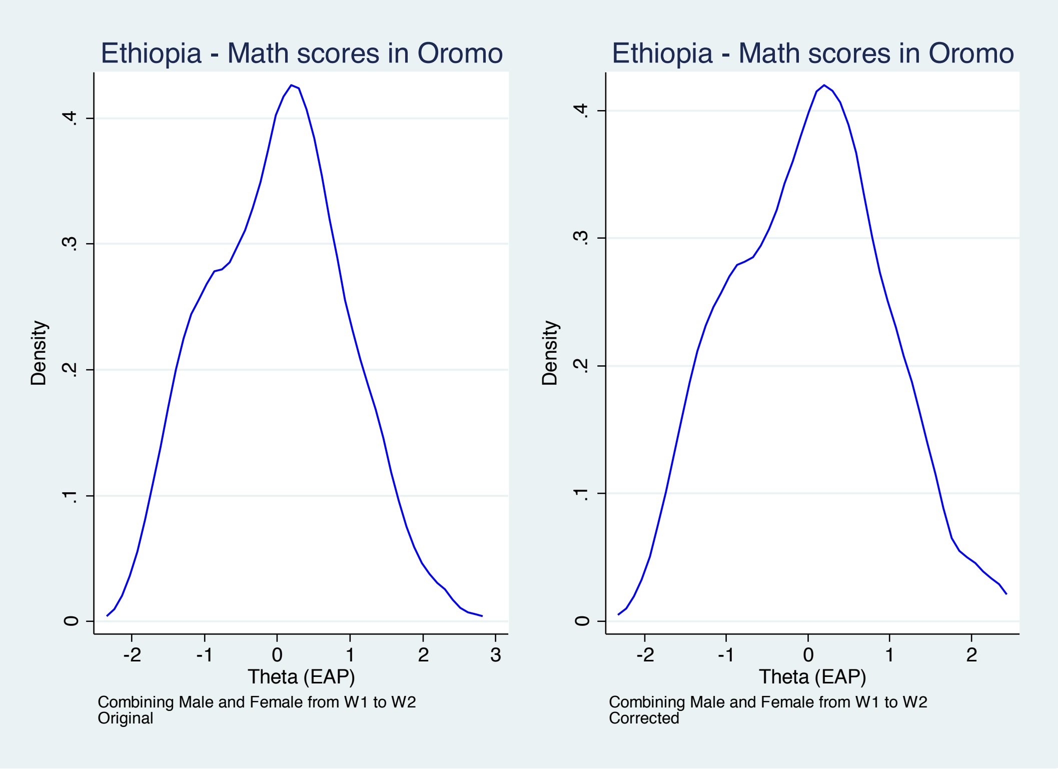


Figure E4. *Original and corrected IRT score distribution for Sidamo*

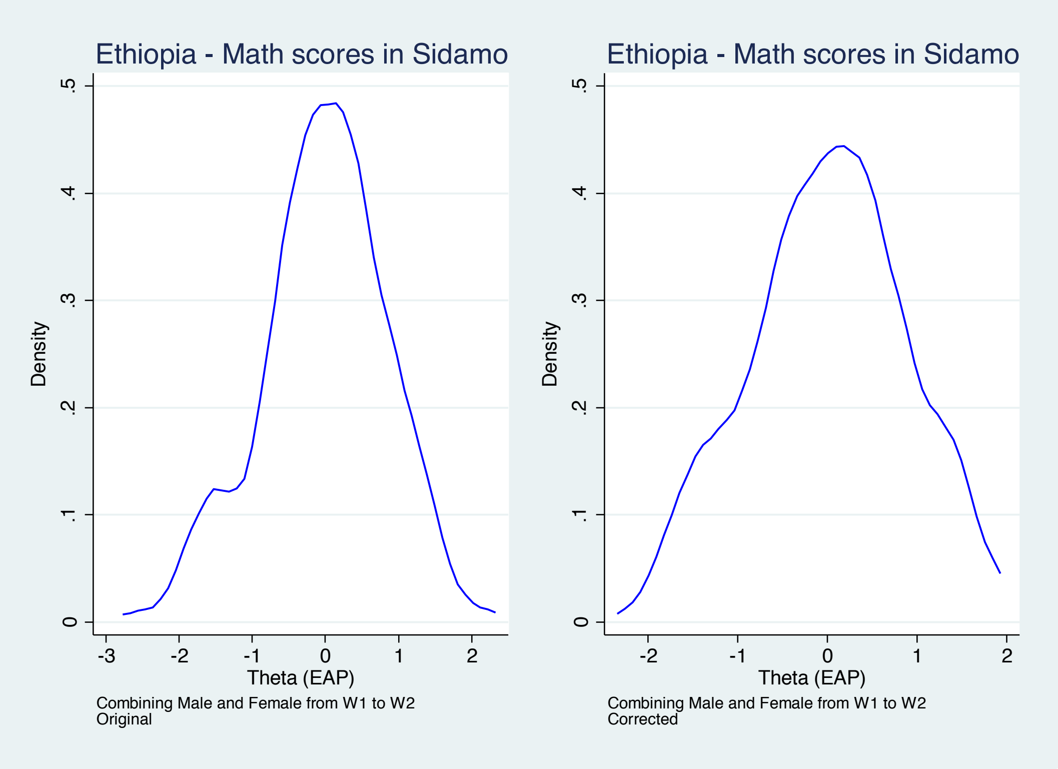


Figure E5. *Original and corrected IRT score distribution for Somali*

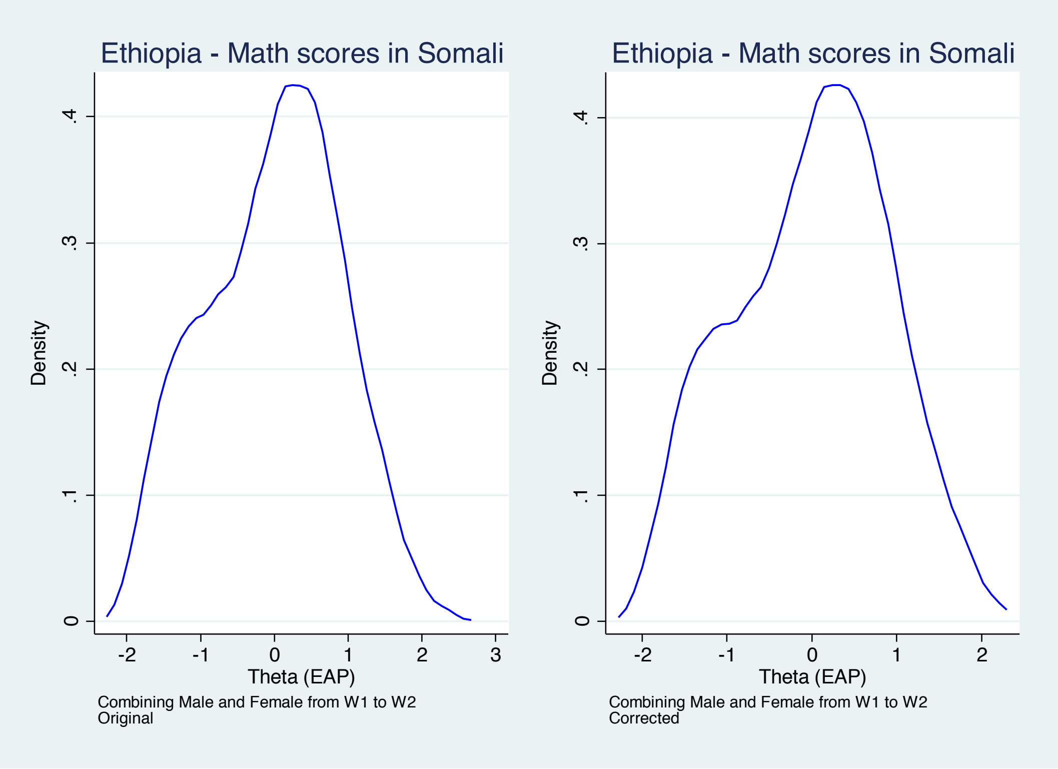


Figure E6. *Original and corrected IRT score distribution for Tigrinya*

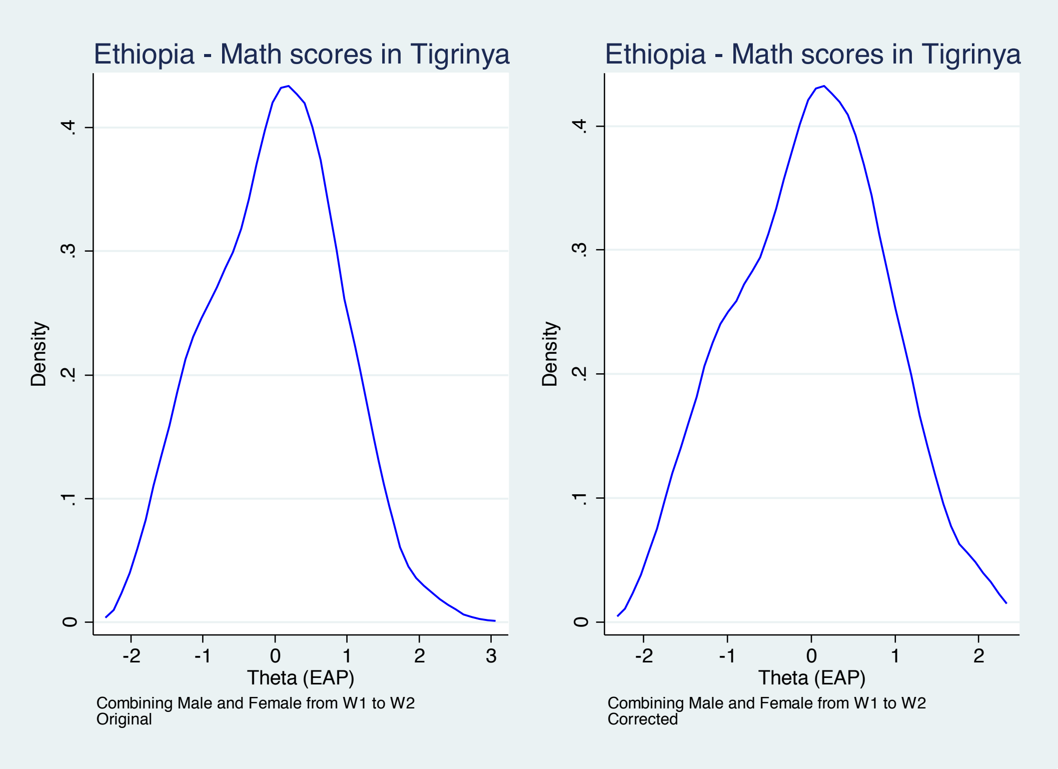
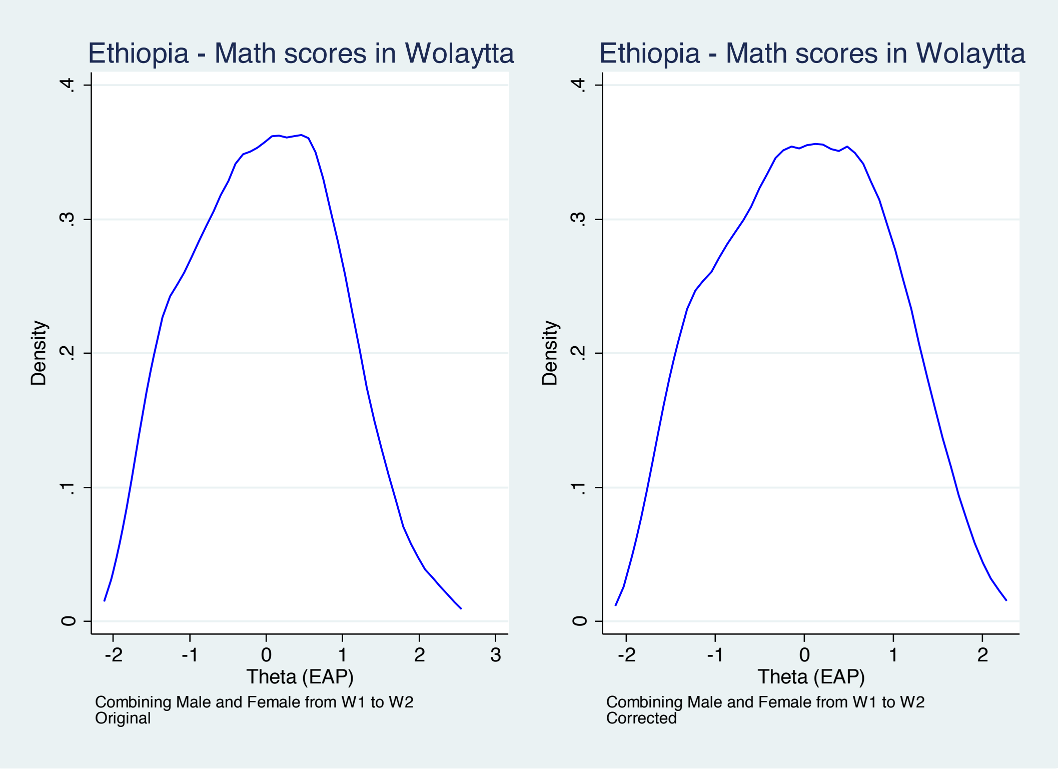


Figure E7. *Original and corrected IRT score distribution for Wolaytta*



Reading comprehension

Figure E8. *Original and corrected IRT score distribution for Amharic*

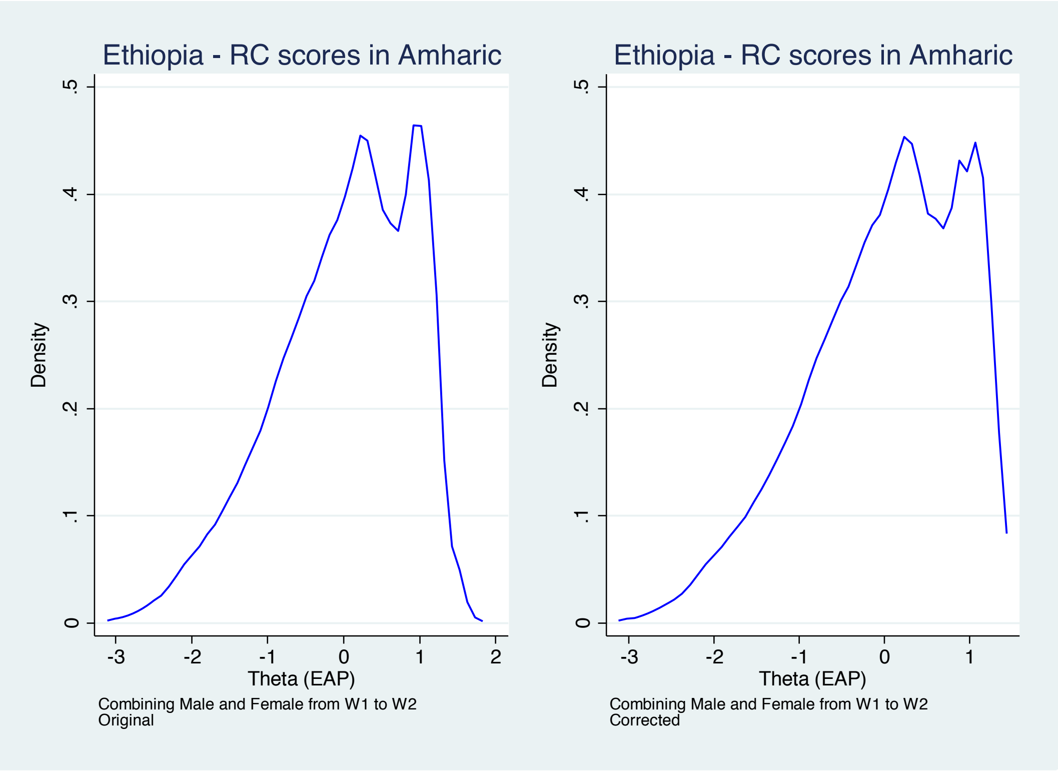


Figure E9. *Original and corrected IRT score distribution for Hadiyya*

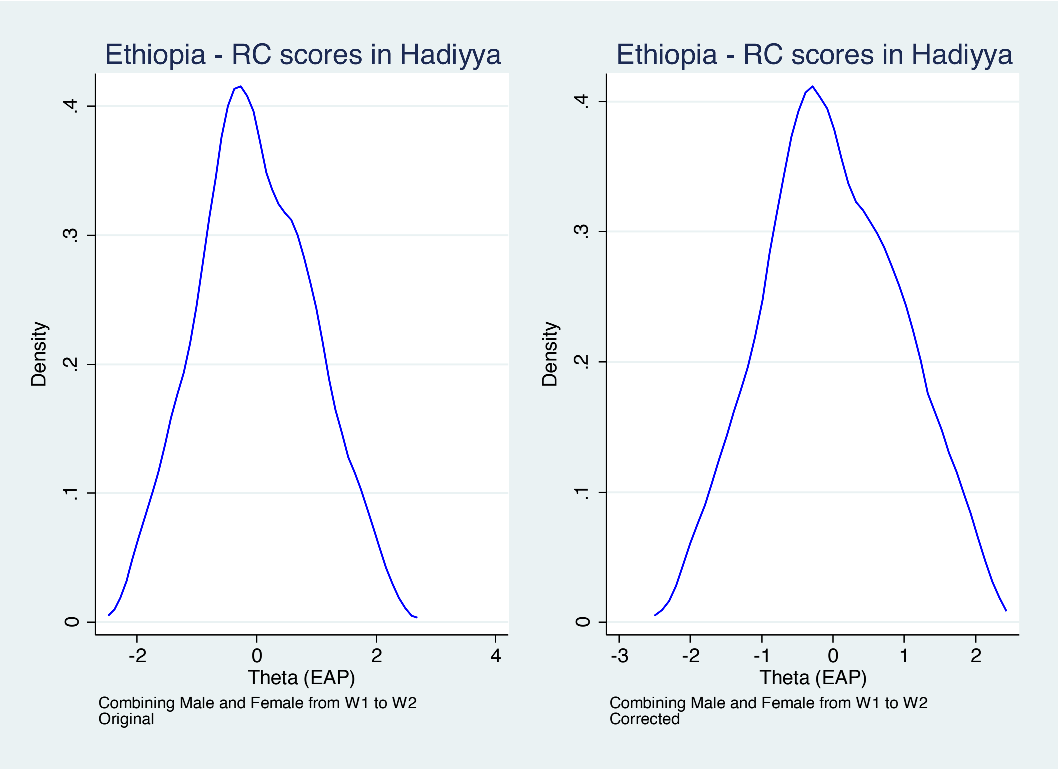


Figure E10. *Original and corrected IRT score distribution for Oromo*

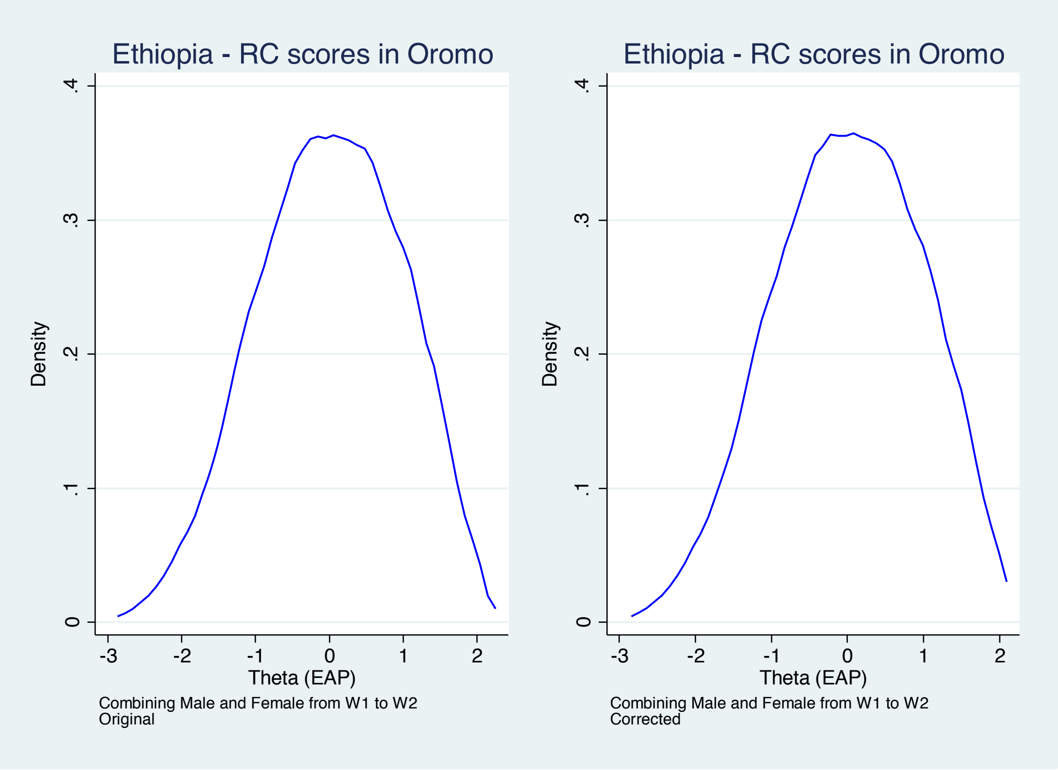


Figure E11. *Original and corrected IRT score distribution for Sidamo*

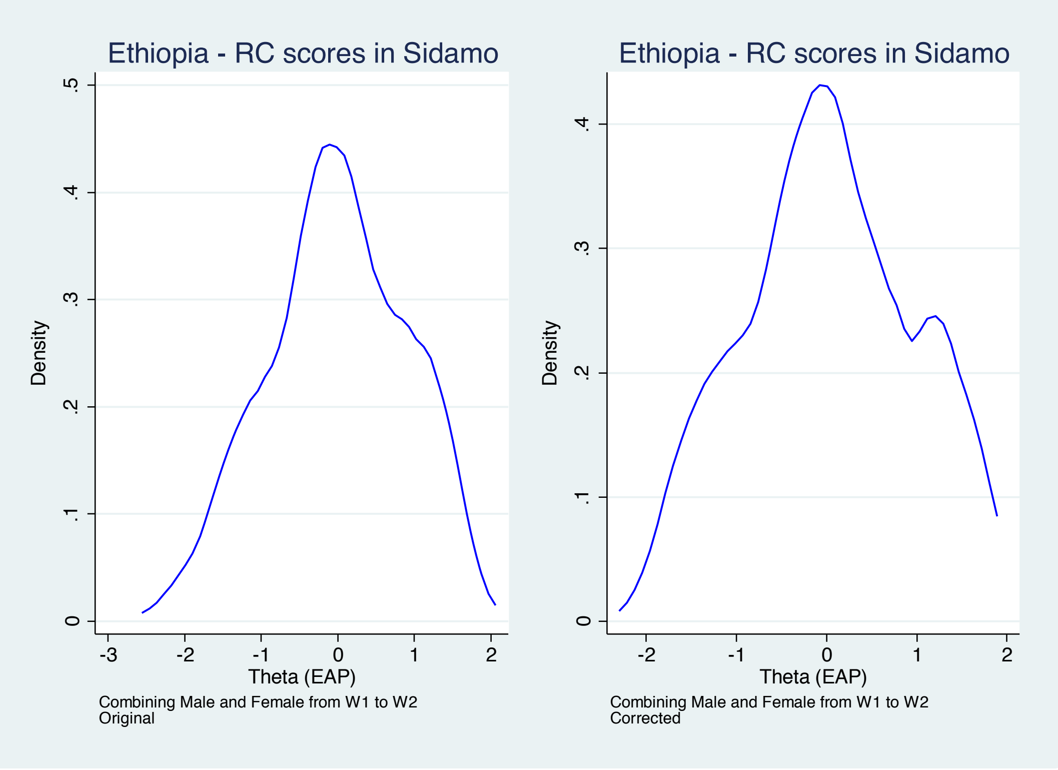


Figure E12. *Original and corrected IRT score distribution for Somali*

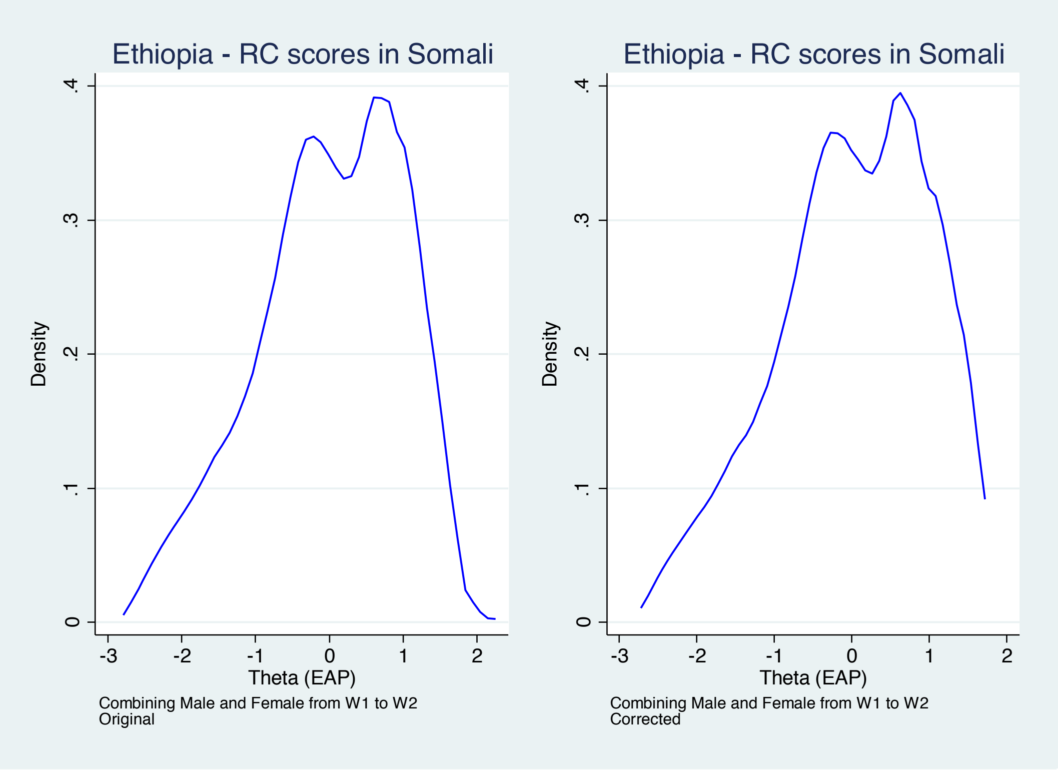


Figure E13. *Original and corrected IRT score distribution for Tigrinya*

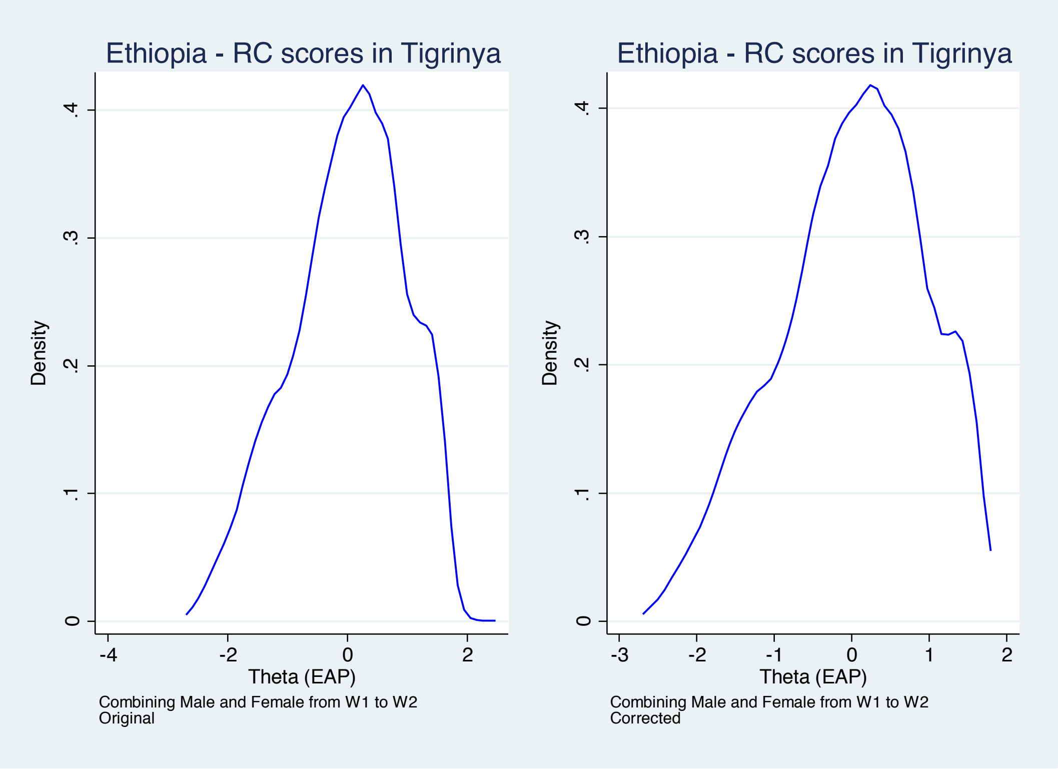
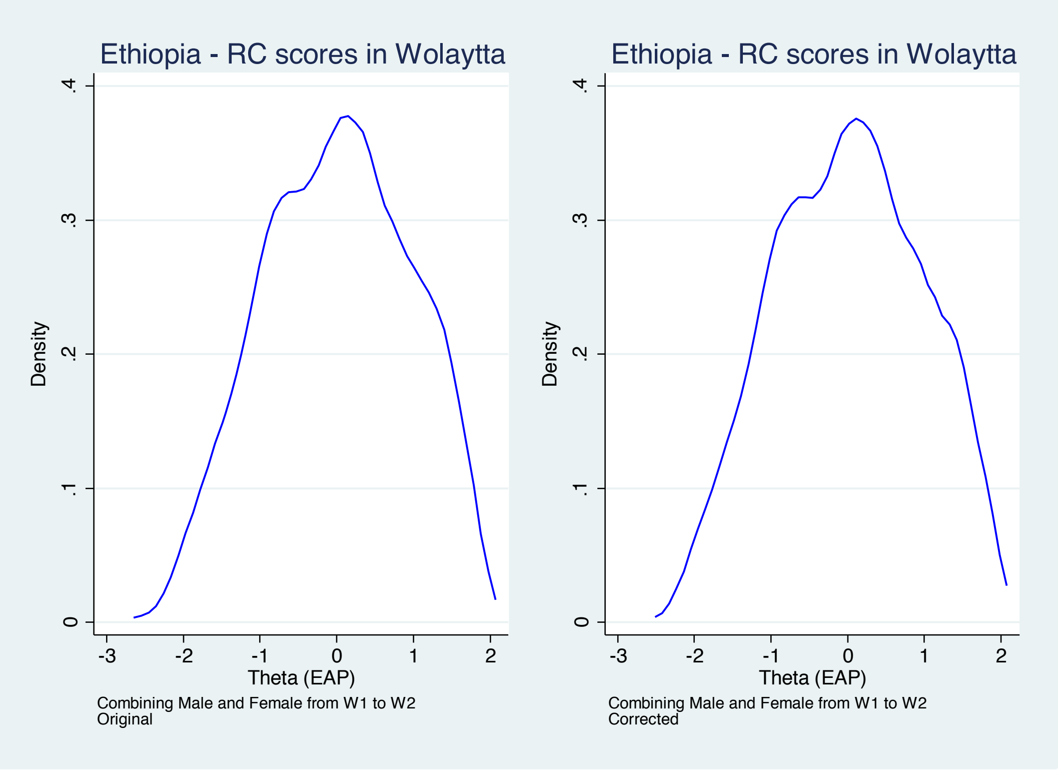


Figure E14. *Original and corrected IRT score distribution for Wolaytta*



The Reliability and Validity of Achievement Tests in the Second Young Lives School Survey in Ethiopia

This technical note gives details of the reliability and validity of the assessments used in the second school survey carried out by Young Lives in Ethiopia for the purpose of the construction of test scores on a common scale within each language for maths and reading comprehension. This document give details of the three-parameter model used to build the achievement scores in both content areas. We tested graphically for item fit and item bias (by gender and wave). Our results indicate that most of the items used have a good item fit as well as they did not show the presence of bias by wave or gender. Finally, we did an external validity analysis correlating the IRT scores (maths and reading comprehension) with individual and family characteristics, and the results showed that correlations were statistically significant with the expected signs.

Young Lives
An International Study of Childhood Poverty

About Young Lives

Young Lives is an international study of childhood poverty, involving 12,000 children in 4 countries over 15 years. It is led by a team in the Department of International Development at the University of Oxford in association with research and policy partners in the 4 study countries: Ethiopia, India, Peru and Vietnam.

Through researching different aspects of children's lives, we seek to improve policies and programmes for children.

Young Lives Partners

Young Lives is coordinated by a small team based at the University of Oxford, led by Professor Jo Boyden.

- *Ethiopian Development Research Institute, Ethiopia*
- *Pankhurst Development Research and Consulting plc, Ethiopia*
- *Save the Children (Ethiopia programme)*
- *Centre for Economic and Social Studies, Hyderabad, India*
- *Sri Padmavathi Mahila Visvavidyalayam (Women's University), Andhra Pradesh, India*
- *Grupo de Análisis para el Desarollo (GRADE), Peru*
- *Instituto de Investigación Nutricional (IIN), Peru*
- *Centre for Analysis and Forecasting, Vietnamese Academy of Social Sciences, Vietnam*
- *General Statistics Office, Vietnam*
- *Oxford Department of International Development, University of Oxford, UK*

Contact:
Young Lives
Oxford Department of
International Development,
University of Oxford,
Mansfield Road,
Oxford OX1 3TB, UK
Tel: +44 (0)1865 281751
Email: younglives@younglives.org.uk
Website: www.younglives.org.uk

Young Lives
An International Study of Childhood Poverty