Educational Methods & Psychometrics—Vol. 4, Article No. 31
May 2026
https://doi.org/10.65301/emp.2026.265

CRITERION-REFERENCED INTERPRETATION THROUGH SCALE ANCHORING: ILLUSTRATED BY ANALYSIS OF DANISH RESULTS FROM PIRLS 2016

Svend Kreiner*

Section of Biostatistics, Department of Public Health, University of Copenhagen

Marianne Müller

School of Health Professions, Bern University of Applied Sciences, Bern, Switzerland

Tine Nielsen

Department of Applied Research in Education and social Sciences, UCL University College, Odense M, Denmark

Measurement by IRT and Rasch models is quantitative and interval-scaled, which is useful for statistical applications studying associations between educational test results and other variables. Quantitative measurement is also useful for ranking of students, but rarely provides information that is useful during formative classroom testing, since it is difficult to interpret quantitative test results in terms of what students can and cannot do. It has been suggested that information of that kind needs so-called criterion-referenced tests with items that are different from the kind of items of conventional educational tests. This paper disagrees. It argues that formative classroom testing needs interpretation of test results from conventional tests and describes how to provide and validate criterion-referenced *interpretation* by analysis of estimates of so-called scale-anchored probabilities of responses to items. Many studies, including the study of *Progress in International Reading Literacy* (PIRLS), interpret test results by scale anchoring, but interpretations are not criterion-referenced. This paper describes the assumptions and requirements of criterion-referenced interpretation and illustrates it on data from PIRLS.

Key words: Criterion-referenced interpretation, scale anchoring, IRT and Rasch models, Log-linear Rasch models, PIRLS.

1. Introduction

1.1 Criterion-referenced testing

Glaser (1963) distinguished between two fundamentally different types of educational test results: norm-referenced scores and criterion-referenced scores, requiring that a criterion-referenced measure should provide "explicit information as to what the individual can or cannot do", "which is independent of reference to the performance of others". Six years later, Popham and Husek (1969) noticed that criterion-referenced measurement had "been the subject of many discussions over the years", but also that "other than adding new terms to the technical lexicon, the two constructs have made little difference in measurement practice".

Correspondence should be made to Svend Kreiner, Section of Biostatistics, Department of Public Health, University of Copenhagen, Øster Farigmagsgade 5, B, 2099, Copenhagen K, Denmark. Email: svend.kreiner@mail.tele.dk

The 1963 paper by Glaser and the 1969 paper by Popham and Husak were reprinted together with a collection of symposium papers edited by Popham (1971) in the hope that it would encourage colleagues to "tackle the tricky technical problems facing us all" (Popham 1971, preface). Despite these efforts and after developing tests for more than a decade[1], Popham (2014) refers to "half a century wasted" for several reasons, including "slapping the label *criterion-referenced* on tests rather than on test-based interpretations".

To Popham and Husek (1969), the fundamental difference between the two types of tests lies in the definition of items. Today, it is not relevant to distinguish between norm-referenced and criterion- referenced testing. Measurement by educational tests is quantitative measurement supported by IRT or Rasch models, and Wright (1980, p. 196), discussing probabilistic models for attainment tests, insisted that interpretation of test scores from Rasch models should do the same as Glaser required of criterion-referenced tests by providing information on students' "strengths and weaknesses". In other words, interpretation should be criterion-referenced.

1.2 Interpretability

De Vet et al. (2011) define interpretability as "the degree to which one can assign qualitative meaning – that is, clinical or commonly understood connotations – to an instrument's quantitative score" (p. 228). The definition applies to measurement supported by both classical and contemporary test theory. Hambleton and Zenisky (2013) describe two ways to interpret test scores supported by conventional IRT models, both of which are important for our paper: interpretation by item mapping and interpretation by scale anchoring.

Interpretation by item mapping extends the interpretation by Wright maps and ICC curves illustrated by Wright and Masters (1982). Wright maps provide information on the locations of dichotomous items and Thurstonian thresholds of polytomous items, and interpretation focuses on whether events defined by responses to items are smaller or larger than 50 %. This information is useful for interpretation, but it is not enough to distinguish between what students can and cannot do. To address this issue, Hambleton and Zenisky (2013) replace the 50% probability criterion with larger probabilities, and assess response probabilities at a small set of so-called benchmarks on the measurement scale.

Interpretation by scale anchoring generates estimates of probabilities and expected frequencies of the responses to items at the selected benchmarks. Forsyth (1991) and Beaton and Allen (1992) describe two ways to calculate scale-anchored probabilities and frequencies. The direct method uses data counting the responses to items for persons close to the benchmark values, while the smoothing method uses the estimates of the item parameters of an IRT model to calculate probabilities given the person parameter associated with the benchmark. In large-sample studies, the two estimates will be close to equivalent, if the IRT model actually fits the data.

The benchmarks of Hambleton and Zenisky and of Beaton and Allen do not depend on subject matter theory and do not refer to the criteria against which to interpret test scores. They are either norm-referenced benchmarks or equidistant benchmarks of convenience.

---

[1] According to Popham (2009, p. 102-103)

1.3 Applications of educational tests.

Popham (2009) distinguishes between formative and accountability testing. Accountability testing is required when somebody "in a position of authority doubts whether a particular educational program is doing a good job" (Popham, 2009, p. 96). The focus in accountability studies is always on quantitative measurement and differences that statistical analyses can disclose. Test results collected during accountability studies are not meant to be useful to teachers in classroom applications, but interpretation of test results may be essential when those responsible for accountability studies need to share their results in meaningful terms with politicians, civil servants, and the public.

PIRLS (Progress in International Reading Literacy Study) is an international accountability study of functional literacy among fourth-grade students. PIRLS[2] recognizes that policy and curriculum improvements require more information than what statistical analyses of quantitative test results are able to make available. For this reason, PIRLS partitions the range of test scores into a number of ordinal proficiency categories that they interpret through anchor scaling as defined by Beaton and Allen (1992). PIRLS does not claim that the proficiency categories are criterion-referenced, and Bruggink et al. (2022) propose applications of PIRLS in classrooms that do not involve PIRLS's proficiency categories. We therefore take it that they do not consider the interpretation inherent in the classification to be useful to teachers.

Popham states that the purpose of formative testing is to "elicit evidence so that teachers can make accurate inferences about their students' unseen knowledge and skills" and "adjust their ongoing instructional procedures", and Wright (1980) reminds us that "the only justification for testing under these circumstances is the intention to use test results to help test takers" by providing information on students' "strengths and weaknesses" (p. 196).

In other words, interpretation of results from formative tests may be helpful in the same way as criterion-referenced testing, but it is the interpretation that has to be criterion-referenced. To address this issue, we will do what Haertel (1985) claims that many others have done. We will illustrate how to "press" quantitative test results from PIRLS 2016 "into service to meet the demands of criterion-referenced testing" to make them useful for formative classroom testing.

1.5 Criterion-referenced interpretation

The criterion-referenced interpretation that this paper describes is interpretation through scale anchoring, but it uses scale anchoring for two different purposes and departs from the way that Beaton and Allen (1992) did it. It derives criterion-referenced *benchmarks* by analysis of a large number of scale-anchored distributions, and it interprets the proficiencies of students within stages separated by benchmarks at the midpoints of the stages rather than at the benchmarks.

The assumptions of the criterion-referenced benchmarks are the same as the assumptions of the so-called construct maps defined by Wilson (2005, 2023). They assume that there is a qualitative order of levels or stages inherent in the construct and that it is possible to specify *waypoints* that separate stages with quantitatively *and* qualitatively different response distributions. Wilson uses such waypoints as benchmarks for the interpretation of test scores by Wright maps. Section 5 of this paper will show how to define and use waypoints as *criterion-*

2 In the following we will use the acronym PIRLS to refer to the study in itself as well as to the International Study Center conducting the study.

referenced benchmarks by a systematic analysis of a large range of scale-anchored distributions.

The set of criterion-referenced benchmarks is supposed to refer explicitly to the criteria to which interpretation refers, and it defines a model of a construct map where the waypoints defined by the benchmarks separate qualitative stages where students meet some but not all the criteria that interpretation refers to. The extreme waypoint should refer to students who meet all the criteria.

During our search for papers on criterion-referenced testing, we came across two papers that have been important for our work on criterion-referenced interpretation. Haertel (1985) outlines a conceptual and methodological framework for criterion-referenced assessment, and Clifford (2016) defined a framework for the interpretation of response probabilities in terms that refer to the *development* of abilities.

1.6 The outline of the paper

Development of a criterion-referenced assessment of educational tests is a stepwise procedure described in Figure 1.

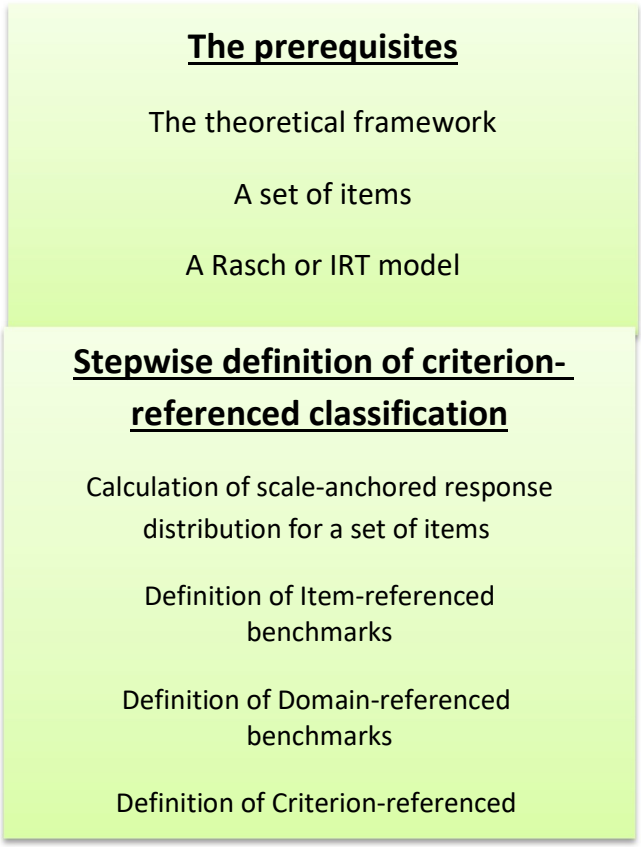


Figure 1.
The methodological components of the procedure for the definition of criterion-referenced proficiency categories.

Section 2 provides a brief description of PIRLS. The criteria on which our criterion-referenced classification of test results from PIRLS will be founded are implied by PIRLS's theoretical framework.

Section 3 describes a measurement model that fits Danish data from PIRLS.

Section 4 discusses measurement issues. Measurement by log-linear Rasch models is quantitative, similar to measurement by conventional Rasch models. This section includes an example that illustrates the need for additional qualitative information on proficiency in addition to conventional quantitative measures of ability information on estimates of person parameters provided by IRT and Rasch models and defines the scale-anchored response distributions that we use to define item-referenced benchmarks.

Section 5 describes the three-step procedure that defines criterion-referenced categories. The first step estimates the scale-anchored distribution of responses to items and defines so-called item-referenced benchmarks. The second step defines domain-referenced benchmarks, and the third step defines the criterion-referenced benchmarks that we use to define criterion-referenced proficiency categories. Finally, this section proposes a criterion-referenced classification scheme that describes seven stages of development of reading literacy.

Section 6 explores the validity and reliability of criterion-referenced interpretation and examines the criterion validity by comparison of the stages of development defined in Section 5 and the PIRLS's benchmarks and proficiency categories.

Finally, Section 7 recapitulates what we have attempted to do and discusses a number of methodological issues.

2. PIRLS

2.1 The theoretical framework

PIRLS distinguishes between two different reading purposes: reading for literary experience and reading to acquire and use information (Mullis & Martin, 2015). To take this into account, PIRLS selected and adapted texts with fictional and non-fictional contents together with questions or items associated with the texts.

In addition to this, PIRLS distinguishes between the following four comprehension processes and developed items where responses to questions only need one specific comprehension process.

Focusing on and retrieving explicitly stated information (INF)
Making straightforward Inferences (IFR)
Interpreting and Integrating Ideas and Information (I&I)
Evaluating and critiquing content and textual elements (E&C)

Haertel (1985) introduced the notion of item domains where "test items of a clearly defined type comprise an *item domain*" and claimed that "there is a general agreement that the criterion-referenced test score must be referable to a well-defined domain" (p. 24). Since responding to items in PIRLS items is supposed to depend on one of the four comprehension processes, it follows that PIRLS´s reading test covers four different item domains and that criterion-referenced assessment of a student´s reading proficiency must refer to the degree to which the student is able or unable to master all four comprehension processes.

2.2 Booklet 16

In 2016, the PIRLS assessment inventory consisted of 12 texts assembled in booklets. Our example uses data from Danish students' responses to the items in Booklet 16, because PIRLS

has published the texts of Booklet 16 together with the items[3]. Our analyses will illustrate how Booklet 16 would function as a stand-alone formative test.

Booklet 16 consists of two texts: "Mary's Red Hen" providing literary experience, and "The Green Turtle" providing factual information. PIRLS used four types of items for each text:

Dichotomous multiple-choice (MC) items
Dichotomous items with open responses
Trinary items with open responses
Polytomous items with open responses scored from zero to three

Figure 2 shows the distribution of Booklet 16 items across the theoretical frame of reference. Since content validity requires that the items cover all parts of the theoretical framework, Figure 2 shows that the reading test of Booklet 16 is content valid.

Reading purpose	Comprehension process			
	Retrieving information	Straightforward inferencing	Evaluating and critiquing	Interpreting and integrating
Literary experience	3 MC items 1 trinary	3 MC items 1 dichotomous	1 MC item 2 dichotomous	1 MC item 3 dichotomous 1 polytomous
Factual information	1 MC item 2 dichotomous 1 trinary	4 MC items 2 dichotomous 1 trinary	2 MC items 1 dichotomous	1 trinary 1 polytomous

Figure 2.
Distribution of Booklet 16 items across the PIRLS framework.

2.3 Scaling methodology

PIRLS applies non-Rasch IRT models for the analysis of the international data and Rasch models for secondary analyses of data at national levels (Martin et al. 2017) but does not explain why they use different models for international and national analyses.

In 2001, PIRLS fixed the origin and unit of measurement of their person parameter scale (θ_{PIRLS}) in such a way that the proficiency distribution across all PIRLS countries had a mean equal to 500 and a standard deviation (SD) equal to 100. To make results comparable across different PIRLS surveys, the definition of θ_{PIRLS} has been the same since 2001. Since many countries have joined PIRLS since 2001, it follows that the interpretation of θ_{PIRLS} values in terms of means and standard deviations of a relevant population no longer applies[4].

2.4 Interpretation through scale anchoring in PIRLS

PIRLS recognized that it takes much more than a value on a quantitative scale to be a meaningful statement on the level of a student's proficiency and used direct scale anchoring to interpret test results. PIRLS used norm-referenced benchmarks in 2001 but decided on a different set of benchmarks that they have used since then (Martin, Mullis & Kennedy, 2006). Figure 3 shows these benchmarks together with the proficiency categories that PIRLS used in

[3] The text and items of Booklet 16 are available as Appendix H in Martin, Mullis, and Hooper (2017).

[4] We refer to Mullis et al. (2003, 2007, 2017) for information on measurement in PIRLS and the way they have taken care of comparability of test results from different PIRLS studies.

2016. The benchmarks are close to the benchmarks of 2001, but they are not norm-referenced, because they do not refer to a specific standard population.

Proficiency categories by PIRLS 2016		International Frequency
Below low international level	$\Theta_{PIRLS} < 400$	4 %
Low level	$400 \leq \Theta_{PIRLS} < 475$	14 %
Intermediate level	$475 \leq \Theta_{PIRLS} < 550$	35 %
High level	$550 \leq \Theta_{PIRLS} < 700$	37 %
Advanced level	$700 \leq \Theta_{PIRLS}$	10 %

Figure 3.
PIRLS's Benchmarks and proficiency categories used by PIRLS 2016

We regard the benchmarks that PIRLS has used as benchmarks of convenience, but this does not imply that the interpretation of the proficiency categories is an interpretation of convenience. PIRLS has invested time and effort in interpreting and documenting the abilities in reading at different levels in a way that makes sense in subject matter terms and is useful for accountability testing, but probably less useful for formative testing. The following sections will show how to address this issue with smooth estimates of scale-anchored probabilities and criterion-referenced benchmarks.

3. A measurement model for Booklet 16

3.1 The items

The Danish data from PIRLS 2016 is available online[5]. Table 1 shows the Booklet 16 items with different colours, distinguishing between the comprehension processes needed to respond correctly to the items.

Table 1.
Overview of Booklet 16 items.

Comprehension process	MC	Dichotomous	Trinary	Polytomous
INF: Retrieving information	**H01 H07 H10**		**H06**	
	T05	**T08 T10**	**T02**	
IFR: Straightforward inferencing	**H05 H09 H11**	**H03r**		
	T01 T09 T12 T13	**T04 T06**	**T03**	
E&C: Evaluating and critiquing	**H02**	**H08 H16**		
	T15 T16	**T14**		
I&I: Interpreting and integrating	**H12**	**H04 H14 H15**		**H13**
			T07	**T11**

Note: Black = Retrieving information; Green = Straightforward inferencing; Blue = Evaluating & Critiquing; Red = Interpreting & integrating

5 Data can be downloaded here: https://timssandpirls.bc.edu/pirls2016/international-database/index.html.
Danish translations of the texts and items can be downloaded here:
https://dpu.au.dk/forskning/internationaleundersoegelser/pirls/pirls-2016/materialer

Table 2 shows the marginal distributions of responses to items for students with complete responses to Booklet 16 items. Items are presented in order defined by increasing manifest item difficulties, as defined by the average item scores divided by the maximum item score. Table 2 distinguishes between multiple choice (MC) items and items with open responses and uses the same colour codes as Table 1 to discriminate between different comprehension processes.

Table 2.
Distributions of responses to Booklet 16 items for students with responses to all items.

Items			Item scores				Manifest difficulty	
Response	Item	Domain	0	1	2	3	Mean	Mean/max
MC	T01	IFR	0.053	0.947			0.947	0.947
MC	H01	INF	0.084	0.916			0.916	0.916
MC	H11	IFR	0.119	0.881			0.881	0.881
MC	H02	E&C	0.119	0.881			0.881	0.881
OPEN	**T06**	**IFR**	**0.148**	**0.852**			**0.852**	**0.852**
MC	T16	E&C	0.155	0.845			0.845	0.845
MC	T12	IFR	0.164	0.836			0.836	0.836
MC	H10	INF	0.183	0.817			0.817	0.817
OPEN	**H15**	**I&I**	**0.242**	**0.758**			**0.758**	**0.758**
OPEN	**T10**	**INF**	**0.247**	**0.753**			**0.753**	**0.753**
MC	H12	I&I	0.251	0.749			0.749	0.749
OPEN	**H06**	**INF**	**0.116**	**0.274**	**0.610**		**1.493**	**0.747**
MC	H05	IFR	0.276	0.724			0.724	0.724
OPEN	**T02**	**INF**	**0.105**	**0.365**	**0.530**		**1.425**	**0.712**
MC	T15	E&C	0.301	0.699			0.699	0.699
OPEN	**H03**	**IFR**	**0.317**	**0.683**			**0.683**	**0.683**
OPEN	**T08**	**INF**	**0.322**	**0.678**			**0.678**	**0.678**
OPEN	**T04**	**IFR**	**0.354**	**0.646**			**0.646**	**0.646**
MC	T09	IFR	0.384	0.616			0.616	0.616
MC	T05	INF	0.386	0.614			0.614	0.614
OPEN	**T03**	**IFR**	**0.285**	**0.256**	**0.459**		**1.174**	**0.587**
MC	H09	IFR	0.416	0.584			0.584	0.584
MC	T13	IFR	0.416	0.584			0.584	0.584
OPEN	**T14**	**E&C**	**0.443**	**0.557**			**0.557**	**0.557**
OPEN	**H14**	**I&I**	**0.466**	**0.534**			**0.534**	**0.534**
MC	H07	INF	0.505	0.495			0.495	0.495
OPEN	**T07**	**I&I**	**0.372**	**0.333**	**0.295**		**0.922**	**0.461**
OPEN	**T11**	**I&I**	**0.484**	**0.091**	**0.199**	**0.226**	**1.167**	**0.389**
OPEN	**H13**	**I&I**	**0.304**	**0.386**	**0.212**	**0.098**	**1.105**	**0.368**
OPEN	**H08**	**E&C**	**0.689**	**0.311**			**0.311**	**0.311**
OPEN	**H16**	**E&C**	**0.724**	**0.276**			**0.276**	**0.276**
OPEN	**H04**	**I&I**	**0.815**	**0.185**			**0.185**	**0.185**

Black = Retrieve information Green = Straightforward inference Blue = Evaluation & Critique Red = Interpretation & integrating

Table 2 underscores two important differences between MC items and items with open responses. First, MC items are easier than items with open responses and second, that INF and IFR items are easier than E&C and I&I items.

It is not surprising that MC items are relatively easy items. The students are aware that the correct response must be one of the alternatives. Notice, however, that T06 and H15 are particularly easy IFR and E&C items with open responses.

3.2 The measurement model

We intend to interpret PIRLS scores by smooth anchor scaling and therefore need estimates of item parameters of a statistical measurement model. Beaton and Allen (1992) describe scale-anchored distributions where expected frequencies of responses to items are calculated according to a 3PL model. Their argument for this model is that "in many applications of test theory the 3PL has been found to fit test data reasonably well" (p. 201). We understand their argument as an argument for a statistical model of convenience, but we must insist on more than a model of convenience. To avoid bias and confounding of the interpretations of the scale-anchored probabilities of responses to items, we must have a measurement model that fits the data and is both realistic and meaningful. One complicating factor is that Booklet 16 consists of two so-called testlets with 16 items that share the same stimulus material. In models with or for testlets, the assumptions of local independence are neither plausible nor realistic and estimates by conventional models like the 3PL or the Rasch models will be biased. To analyze PIRLS's data, we need a model describing the testlet structure where local dependence has been taken into account.

One possibility is the Rasch testlet model proposed by Wang and Wilson (2005), where local dependence among items within a testlet is generated by a random effect associated with the testlet. However, Wang and Wilson's model would generate a two-dimensional structure with separate Rasch models for the two testlets. We considered these models, but tests of fit rejected the Rasch models for the two testlets, disclosing strong evidence of local dependence within both Rasch models.

Another possibility is the so-called Rasch model for item bundles proposed by Wilson and Adams (1995), where scores summarizing responses to items within bundles define super-items fitting Rasch models for polytomous items. The problem with this model is that it cannot be used for scale-anchoring, because the responses to the separate items are unavailable. Items that fit the Rasch model for item bundles must fit Kelderman's (1984) log-linear Rasch model, where estimates of scale-anchored probabilities are available (Kreiner, 2026). For this reason, we fitted a log-linear Rasch model to the Danish Booklet 16 data and used this model to illustrate criterion-referenced interpretation. Appendix A provides some information on the item analysis by this model. The Rasch model for item bundles and the log-linear Rasch model share one important property of conventional Rasch models. The total raw score over items is sufficient for the person parameter, and the estimation of person parameters proceeds in exactly the same way as in conventional Rasch models. Kreiner and Nielsen (2026) provide one example where the estimates of the person parameters from the two models are close to indistinguishable. The major difference between the models is that the bundle model assumes that all items are locally dependent, whereas the log-linear Rasch model only assumes that there may be local dependence and regards the conventional Rasch model as a special case where all items are locally independent. We refer readers to the seminal papers by Kelderman

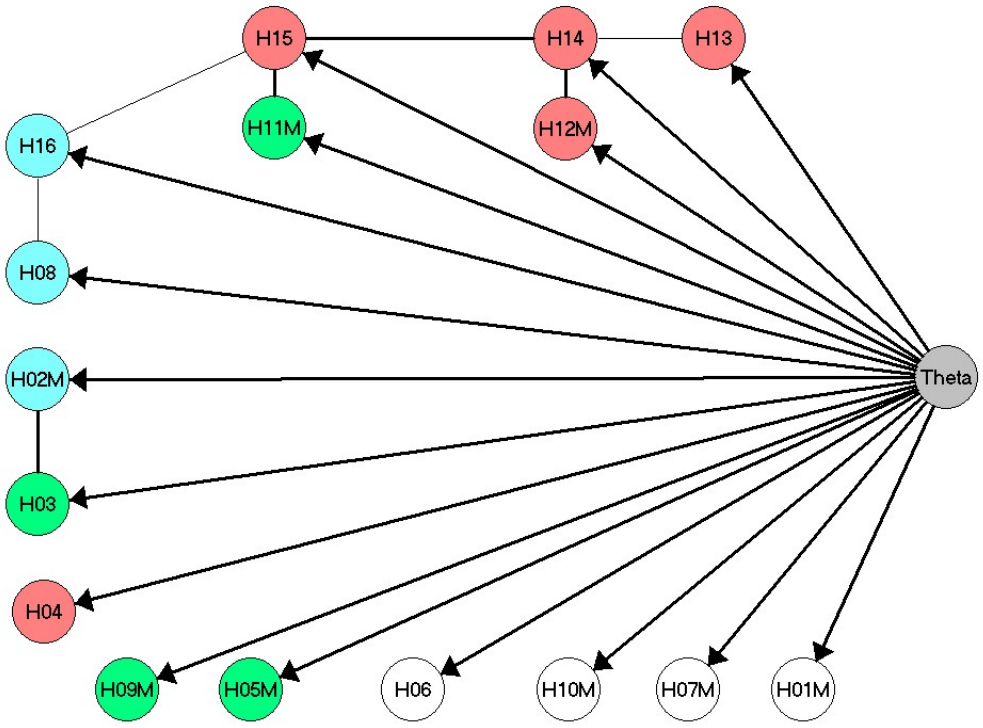
(1984, 1989 & 1992) and to papers of Kreiner and Christensen (2002, 2004, 2007 and 2011a)[6] for information on item analysis by log-linear Rasch models.

Figures 4 and 5 show two so-called chain graphs that visualize the association structure within the two testlets of Booklet 16. In chain graphs of log-linear Rasch models, a missing edge between two items means that they are locally independent[7]. The association structure is complicated, but 13 out of the 32 items are nevertheless locally independent of the other items. To make the association structure interpretable in subject matter terms we use thick edges to indicate very strong association and color codes to distinguish between the different types of items. Retrieving info is white, Inferencing is green, Interpreting is red, and Evaluating is blue. Multiple-choice items have an M attached to the item name.

4. Measurement issues

IRT and Rasch models provide quantitative interval-scaled measurement by estimates of person parameters. In log-linear Rasch models, the estimate of the person parameter $\hat{\theta}$ is a function of the total score over items, because the score is sufficient for θ. From this it follows that $\hat{\theta}$ is a *discrete* variable with a range that consists of 41 real numbers that Rasch models refer to as logits.

Kreiner (2025) compared different estimators of person parameters in Rasch models. In this paper we use Warm's (1989) weighted likelihood (WLE) because it is less biased than the other estimators.



[6] Item analysis by log-linear Rasch models is implemented in the DIGRAM program (Kreiner & Nielsen, 2023).

[7] In chain graph models (Lauritzen, 1996), a missing edge between variables in a so-called interaction graph means that they are conditionally independent.

Figure 4.
IRT graph of the log-linear Rasch model of items related to “Macy’s Red Hen”.

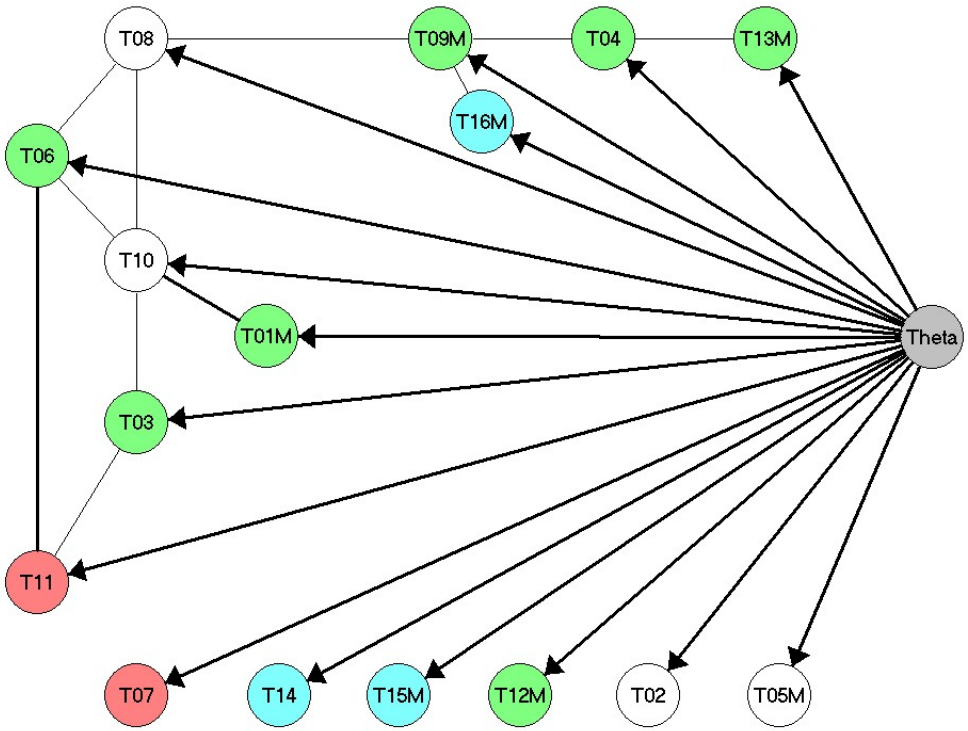


Figure 5.
IRT graph of the log-linear Rasch model of items related to “The Green turtle”.

Kreiner and Christensen (2013) describe how to estimate the exact distribution of $\hat{\theta}$ including bias and standard error of measurement (SEM). Table 3 presents Warm’s (1989) weighted likelihood (WLE) estimate of θ together with the bias and the standard error of measurement (SEM). $\hat{\theta}$ is virtually unbiased except for extreme values of the raw score.

4.1 Formative testing

Data on classroom applications of Booklet 16 is unavailable. We therefore use simulated data to illustrate how criterion-referenced interpretation may be useful for formative testing.

Table 4 shows the effect of an intensive reading course on the proficiency of a single student. The table includes the true values of the person parameter before and after the course, together with the simulated Booklet 16 scores and the estimates of the person parameters. Since we have an estimate of the distribution of the total score anchored at the person parameters, we have been able to simulate the scores without the responses to the items. The situation, therefore, is similar to a classroom application where responses to items are unavailable and where assessment of proficiency refers to the total scores or the estimates of the person parameters. In this case, the course increased the student’s scores by 9 points and the person parameter by close to 1 logit, $\hat{\theta}_1 - \hat{\theta}_0 = 0.923$. Since the difference between the scores is

marginally significant[8], we conclude that there is evidence of a positive effect of the course on the student's proficiency.

Table 3.
WLE estimates of person parameters in the log-linear Rasch model for Booklet 16.

Score	$\hat{\theta}$	Bias	SEM	Score	$\hat{\theta}$	Bias	SEM
0	-5.038	0.494	.817	21	-.223	0.000	.318
1	-3.874	0.065	.767	22	-.124	0.000	.318
2	-3.298	0.008	.703	23	-.025	-0.001	.318
3	-2.900	0.000	.622	24	.075	-0.001	.320
4	-2.591	-0.001	.556	25	.174	0.000	.323
5	-2.336	-0.001	.506	26	.276	0.000	.327
6	-2.117	-0.001	.468	27	.378	0.000	.333
7	-1.923	-0.001	.439	28	.485	0.001	.340
8	-1.750	-0.001	.416	29	.596	0.001	.349
9	-1.592	-0.001	.397	30	.713	0.002	.361
10	-1.446	-0.001	.381	31	.840	0.002	.375
11	-1.311	-0.001	.369	32	.979	0.002	.394
12	-1.183	-0.001	.358	33	1.133	0.002	.417
13	-1.063	0.000	.349	34	1.308	0.002	.447
14	-.948	0.000	.341	35	1.509	0.002	.487
15	-.837	0.000	.335	36	1.744	0.002	.538
16	-.730	0.000	.330	37	2.032	0.001	.605
17	-.626	0.000	.326	38	2.407	-0.008	.686
18	-.524	0.000	.323	39	2.962	-0.064	.752
19	-.423	0.000	.320	40	4.113	-0.491	.810
20	-.323	0.000	.319				

Table 4.
Measurement of reading before and after an intensive reading course.

	Before	After	Effect
True θ	-1.0	0.0	1.000
Total score	14	23	9
$\hat{\theta}$	-0.948	-0.025	0.923

[8] Under the null-hypothesis of no difference, θ is the same before and after the course. From this it follows that the conditional probability of the first score R_1 given the total score R_1+R_2 does not depend on the common θ. Since $PR(R_1 \leq 14 \mid R_1+R_2 = 37) = 0.034$ we conclude that there is a significant effect of the course.

4.2 Estimates of scale-anchored response distributions

Readers familiar with the Rasch model's logit scale will probably conclude that an effect of this size is important[9], but the problem with this conclusion is that the test results in themselves do not provide insights into what the student could do after the course, that he was unable to do before the course.

Knowing that $\hat{\theta}$ is measured by interval-scaled logits is of considerable interest to measurement specialists, but it is not helpful at all to the teacher who needs concrete information that can help her interpret what θ values equal to -0.948 and -0.025 say about what the student can do after the course that he could not do before. However, calculating estimates of scale-anchored response probabilities by inserting person parameters together with estimates of item parameters in the formulas that define the probabilities of responses to items may be helpful. Table 5 presents estimates of the response probabilities of the 17 items with open responses anchored at the two estimates of θ. We will use these probabilities to interpret the students' abilities in terms of how difficult the items were to the student before the course and how easier they were after the course.

Table 5.
Scale-anchored distributions of response probabilities to items with open responses.

		Before: θ = -0.948				After: θ = -0.025			
Item	Domain	0	1	2	3	0	1	2	3
T06	IFR	.434	.566			.131	.869		
H15	I&I	.558	.442			.251	.749		
T10	INF	.633	.367			.237	.763		
H06	INF	.314	.406	.280		.101	.329	.571	
T02	INF	.273	.507	.220		.093	.433	.474	
H03	IFR	.612	.388			.340	.660		
T08	INF	.721	.279			.343	.657		
T04	IFR	.697	.303			.394	.606		
T03	IFR	.718	.210	.072		.314	.351	.335	
T14	E&C	.710	.290			.493	.507		
H14	I&I	.783	.217			.521	.479		
T07	I&I	.709	.241	.051		.433	.370	.197	
T11	I&I	.924	.052	.020	.004	.648	.120	.149	.082
H13	I&I	.569	.374	.052	.005	.350	.452	.162	.036
H08	E&C	.900	.100			.767	.233		
H16	E&C	.932	.068			.806	.194		
H04	I&I	.942	.058			.867	.133		

Note: Black = Retrieve information; Green = Straightforward inference; Blue = Evaluation & Critique; Red = Interpretation & integrating

Before the course, there were no easy items, and four items were close to unsolvable, with probabilities of incorrect responses larger than or equal to 90 %. After the course, there are

[9] If X_i is a dichotomous item with location equal to θ_0, so that the probability of a correct response to X_i was equal to 0.50 before the course, then the logit effect equal to 0.923 implies that the probability has increased to 0.73 after the course.

three easy items (T06, H15, and T10) with probabilities of correct responses larger than or equal to 75 %, and only three difficult items (H08, H16, and H04) with probabilities of incorrect responses between 75 and 90 %. A closer look at the issues raised by these items[10] may provide insight into what the student was able and unable to do after the course.

5. Criterion-referenced interpretation of Booklet 16 scores.

Criterion-referenced interpretation is supposed to describe the degree to which students can read and understand relevant texts in terms that relate to their abilities, and that requires much more than information on the difficulties of a few items. The rest of this paper will describe one way to provide criterion-referenced interpretation by systematic and careful examination of scale-anchored probabilities interpreted in terms of the degree to which students' proficiencies satisfy the criteria needed to read and understand texts similar to those in Booklet 16.

5.1 Item domains

Popham and Husek (1969, p.6) stated that "those who write criterion-referenced items are more attentive to defining the domain of relevant test results and the situation in which they should be required" and Glaser (1971) underscored that criterion-referenced tests should yield measurement which is interpretable in terms of performance standards defined by "representative samples of tasks" (p.37) drawn from item domains.

We refer to Berk (1980), Haertel (1985) and Reckase (2017) for information on item domains in criterion-referenced tests. In PIRLS, the comprehension processes define four item domains, and Mullis et al. (2017, pp. 85 and 97) describe the comprehension processes that characterize students at high and advanced levels of proficiency. Their summaries provide the criteria that students must meet to be able to read texts like Mary's Red Hen and The Green Turtle. It is these criteria that our attempt to interpret Booklet 16 scores will refer to. The criteria are:

Retrieving and using information in texts:

- When reading literary texts, the student should be able to locate and distinguish significant actions and details embedded across the text.
- When reading informational texts, the student should be able to locate and distinguish information within a dense text or a complex table.

Straightforward inferencing depending on what they have read:

- When reading literary texts, the student should be able to explain relationships between intentions, actions, events, and feelings.
- When reading informational texts, the student should be able to provide explanations and reasons about logical connections.

Interpreting and integrating ideas and information:

- When reading literary texts, the student should be able to interpret story events and character actions to describe reasons, motivations, feelings, and character development.
- When reading informational texts, the student should be able to distinguish and interpret complex information from different parts of the text.

[10] Appendix B provides information on the items with open responses.

Evaluating and criticizing contents:

- When reading literary texts, the student should be able to evaluate the effect of the author's language and style choices on the reader.
- When reading informational texts, the student should be able to evaluate visual and textual elements to consider the author's point of view.

Hambleton (1980) provides an extensive list of requirements that items of criterion-referenced tests must meet in addition to the requirements of items for norm-referenced tests. These requirements are also relevant for criterion-referenced interpretation of measurement by IRT and Rasch models. Even though we assume that all items fit the log-linear Rasch model and therefore provide item information that adds to the precision of the quantitative measurement, it does not follow automatically that all the items are useful for criterion-referenced interpretation. Appendix B provides information on the items that we will use. We exclude multiple-choice items because the difficulty of an MC item is less than what the difficulty of the item would have been if students did not know that the correct response is one of the alternatives provided by the item, and we also exclude two very easy items (T06 and H15) with open responses.

This leaves us with 15 items with open responses that we refer to as the criterion-referencing CR-items. This is less than we would prefer if this had been more than an illustration of methodology, and we acknowledge that it is an open question whether the 15 items constitute a "representative" sample of items from the four domains that satisfy Hambleton's list of requirements. However, it is enough to illustrate the methodology.

5.2 Item-referenced benchmarks

It is intuitively easy to interpret an estimate $\hat{\theta}$ in terms of scale-anchored difficulties of items, but the interpretation will be tied to the specific items that the students responded to, and criterion-referenced interpretation has to look beyond the specific texts and items of Booklet 16. It should attempt to describe the level of development of the student's proficiency and assess whether students at this level are able to or unable to read and understand similar texts. We will address this issue by defining benchmarks in the same way as PIRLS, but we define benchmarks by a stepwise procedure that is fundamentally different from the way that PIRLS selected their benchmarks in 2006. The first step defines item-referenced benchmarks, the second defines domain-referenced benchmarks, and the third and final step defines criterion-referenced benchmarks.

Clifford (2006) characterized stages of development of abilities as either sporadic, emerging, developing or sustained depending on probabilities of incorrect and correct responses to dichotomous items. Table 6 classifies the stages of development in the same way. However, it includes polytomous items where partial credit is neither incorrect nor correct, and it adds two extreme stages that refer to absent and consolidated evidence of development.

Table 6.
Item-referenced classification of abilities and difficulties.
$P_0 = PR(X_i = 0)$ $P_1 = PR(X_i = m_i)$

Ability	Scale-anchored probabilities		Item difficulty
	Incorrect response	Correct	
Absent	90 % $\leq P_0$		Very difficult
Sporadic	75 % $\leq$ P_0 < 90 %		Difficult
Emerging	$P_0 \leq$ 75 %	P_1 <50 %	Challenging
Developing		50 % $\leq P_1$ < 75 %	Less than challenging
Sustained		75 % $\leq P_1$ < 90 %	Easy
Consolidated		90 % $\leq$ P_1	Very easy

Table 7 shows the probabilities of responses to items H06 and T08 anchored at WLE estimates of θ together with the item-referenced benchmarks defined by the lowest value of $\hat{\theta}$ that satisfies the requirements defined in Table 6

In Table 7, probabilities in bold numbers refer to item-referenced benchmarks. Sporadic is gray, emerging is red, developing is red, sustained is blue and consolidated is green.The emerging benchmark of H06 is equal to -2.34, where the probability of an incorrect score is less than 0.75 for the first time, while the emerging benchmark of the more difficult T08 is equal to -0.95.

The item-referenced benchmarks of locally independent dichotomous Rasch items would be consistently ordered at all levels of θ, because Rasch model for dichotomous items satisfy the requirement of invariant item ordering (IIO). Regrettably, Sijtsma and Hemker (1998) show that IIO only applies for polytomous items if they fit the rating scale model defined by Andrich (1978), and Kreiner & Christensen (2004) show that it never applies for locally dependent dichotomous items in log-linear Rasch models.

In Table 7, the developing, sustained and consolidated benchmarks of H06 lie above the benchmarks of T08, and the probability of a completely correct response is larger for the difficult T08 than for the easy H06 if θ > -0.84. This is not unexpected, but it complicates interpretation.

There is no straightforward algebraic way to predict the relationships between the benchmarks of CR-items in log-linear Rasch models with a mixture of dichotomous items and polytomous items. The best we can do is to collect the benchmarks in tables for each of the domains. Tables 8–11 present the benchmarks for each of the four comprehension processes.

5.3 Domain-referenced benchmarks

Identifying item-referenced benchmarks is only the first step. The second step defines domain-referenced benchmarks that encapsulate the information on item-referenced benchmarks within the domains. There is no simple formal way to do this. However, we suggest that students have reached a specific domain-referenced benchmark at the lowest score where most of the items associated with the domain have reached the item-referenced benchmark. Table 12 shows the resultsand Figure 6 provides a visual illustration of the development of proficiency implied by the domain-referenced benchmarks. Retrieving information and straightforward inferencing develop faster than interpreting and evaluating texts, and interpreting develops, at least to begin with, faster than evaluating.

Table 7.
Scale-anchored probabilities of responses to items H06 and T08

		H06				T08		
Score	θ	0	1	2	Mean	0	1	Mean
1	-3.87	.933	.065	.002	0.07	.992	.008	0.01
2	-3.30	.884	.109	.007	0.12	.985	.015	0.01
3	-2.90	.832	.153	.015	0.18	.977	.023	0.02
4	-2.59	.779	.195	.026	0.25	.967	.033	0.03
5	-2.34	.725	.235	.040	0.32	.955	.045	0.05
6	-2.12	.672	.270	.058	0.39	.941	.059	0.06
7	-192	.619	.302	.079	0.46	.924	.076	0.08
8	-1.75	.568	.330	.102	0.53	.904	.096	0.10
9	-1.59	.519	.353	.128	0.61	.882	.118	0.12
10	-1.45	.472	.372	.156	0.68	.856	.144	0.14
11	-1.31	.428	.386	.185	0.76	.827	.174	0.17
12	-1.18	.387	.397	.216	0.83	.794	.206	0.21
13	-1.06	.349	.403	.248	0.90	.759	.241	0.24
14	-0.95	.314	.406	.280	0.97	.721	.279	0.20
15	-0.84	.281	.406	.313	1.03	.680	.320	0.32
16	-0.73	.251	.403	.346	1.10	.638	.362	0.36
17	-0.63	.223	.398	.379	1.16	.595	.405	0.41
18	-0.52	.197	.391	.412	1.21	.551	.449	0.45
19	-0.42	.174	.381	.445	1.27	.506	.494	0.49
20	-0.32	.153	.370	.478	1.32	.463	.537	0.54
21	-0.22	.134	.357	.509	1.38	.421	.579	0.58
22	-0.12	.116	.344	.540	1.42	.381	.619	0.62
23	-0.02	.101	.329	.570	1.47	.343	.657	0.66
24	0.07	.087	.313	.600	1.51	.308	.692	0.69
25	0.17	.074	.297	.629	1.55	.275	.725	0.72
26	0.28	.064	.280	.657	1.59	.245	.755	0.75
27	0.38	.054	.263	.683	1.63	.218	.782	0.78
28	0.48	.045	.245	.709	1.66	.193	.807	0.81
29	0.60	.038	.228	.735	1.70	.170	.830	0.83
30	0.71	.031	.209	.760	1.73	.148	.852	0.85
31	0.84	.025	.190	.785	1.76	.128	.872	0.87
32	0.98	.019	.171	.810	1.79	.110	.890	0.89
33	1.13	.014	.151	.834	1.82	.093	.907	0.91
34	1.31	.011	.130	.859	1.85	.076	.924	0.92
35	1.51	.007	.110	.883	1.88	.062	.938	0.94
36	1.74	.005	.089	.906	1.90	.048	.952	0.95
37	2.03	.003	.068	.929	1.93	.035	.965	0.96
38	2.41	.001	.048	.951	1.95	.024	.976	0.98
39	2.96	.000	.028	.971	1.97	.014	.986	0.99

Table 8.
Item-referenced benchmarks of items asking students to retrieve information in texts

Benchmarks	T10	H06	T02	T08
Sporadic	8	2	2	9
Emerging	12	5	4	14
Developing	17	21	24	20
Sustained	23	30	33	26
Consolidated	30	36	38	33

Table 9.
Item-referenced benchmarks of items asking for straightforward inference

Benchmarks	H03	T04	T03
Sporadic	6	7	9
Emerging	11	13	14
Developing	18	20	27
Sustained	27	28	32
Consolidated	35	35	37

Table 10.
Item-referenced benchmarks of items asking for interpretation and integration

Benchmarks	H14	T07	T11	H13	H04
Sporadic	9	6	16	4	20
Emerging	16	13	21	8	31
Developing	24	32	33	37	37
Sustained	33	37	37	39	39
Consolidated	38	39	39	40	40

Table 11.
Item-referenced benchmarks of items asking for evaluation and critique

Benchmarks	T14	H08	H16
Sporadic	6	14	18
Emerging	13	24	26
Developing	23	33	33
Sustained	33	37	38
Consolidated	38	40	40

Table 12.
Domain-referenced benchmarks

Benchmarks	Retrieving info	Inferencing	Interpreting	Evaluating
Sporadic	8	7	9	14
Emerging	12	13	16	24
Developing	21	20	33	33
Sustained	30	28	37	37
Consolidated	36	35	39	40

CR-Stage	Score	$\hat{\theta}$	Retrieving	Inferencing	Interpreting	Evaluating
A	0	-5.04				
	1	-3.87				
	2	-3.30				
	3	-2.90				
	4	-2.59				
	5	-2.34				
	6	-2.12				
	7	-1.92		Sporadic		
B	8	-1.75	Sporadic			
	9	-1.59			Sporadic	
	10	-1.45				
	11	-1.31				
	12	-1.18	Emerging			
	13	-1.06		Emerging		
C	14	-0.95				Sporadic
	15	-0.84				
	16	-0.73			Emerging	
	17	-0.63				
	18	-0.52				
	19	-0.42				
	20	-0.32		Developing		
	21	-0.22	Developing			
D	22	-0.12				
	23	-0.02				
	24	0.07				Emerging
	25	0.17				
	26	0.28				
	27	0.38				
	28	0.48		Sustained		
	29	0.60				
	30	0.71	Sustained			
	31	0.84				
E	32	0.98				
	33	1.13			Developing	Developing
	34	1.31				
	35	1.51		Consolidated		
	36	1.74	Consolidated			
F	37	2.03			Sustained	Sustained
	38	2.41				
	39	2.96			Consolidated	
	40	4.11				Consolidated

Figure 6.
Definition of stages of development defined by domain-referenced benchmarks.
Colorless is absent, Sporadic is gray, emerging is red, developing is yellow, sustained is blue, and consolidated is green.

5.4 Criterion-referenced classification and interpretation

The final step defines criterion-referenced stages of development. The horizontal lines of Figure 6 define five criterion-referenced benchmarks that partition the proficiency scale into six intervals, and Table 13 interprets these intervals as CR-stages of development from absent to consolidated proficiency. The table provides information on the definition of the stages in terms of scores and estimates of person parameters and on the criterion-referenced benchmarks. In addition to interpretation of proficiency at the midpoints of the stages, the table includes information on the distribution of the students across CR-stages and on the CR-items with midpoints in the stages (Appendix B includes Information on scale-anchored response distributions at the midpoints of the stages).

Table 13.
Criterion-referenced classification in stages of development

CR-Stage	Score	Frequency*	Range	Criterion-referenced interpretation at the midpoints of the stages	Items with midpoints within stages
A	0-7	4.8 %	$-5.038 \leq \theta < -1.750$	Clearly inadequate abilities.	
B	8-13	14.1 %	$-1.750 \leq \theta < -0.948$	Information: Sporadic Inferencing: Sporadic Interpreting: Sporadic Evaluating: Absent	INF: T10 IFR: T06**
C	14-21	24.7 %	$-0.948 \leq \theta < -0.124$	Information: Emerging Inferencing: Emerging Interpreting: Sporadic or emerging Evaluating: Sporadic	INF: H06 T02 T08 ITR: H03 T04 I&I: H15**
D	22-31	40.7 %	$-0.124 \leq \theta < 0.979$	Information: Developing Inferencing: Developing. Close to sustained Interpreting: Emerging Evaluating: Emerging	IFR: T03 I&I: H14 T07 T11 H13 E&C: T14
E	32-36	13.9 %	$0.979 \leq \theta < 2.032$	Information: Sustained Inferencing: Consolidated Interpreting: Developing Evaluating: Developing	 I&I: H04 E&C: H08 H16
F	37-40	1.8 %	$2.032 \leq \theta < 4.113$	Information: Consolidated Inferencing: Consolidated Interpreting: Consolidated Evaluating: Sustained. Almost consolidated	

*: Including students with incomplete responses to items
**: Items with open responses that were not included in the criterion-referenced interpretation

The setup of Table 13 is similar to the construct maps proposed by Wilson (2005, Chapter 2), but Wilson's construct maps require coherent and substantive theories of the stepwise *development* of reading proficiency. Since this is only implicitly implied by PIRLS's theoretical framework, and because the six stages of Table 13 are the result of an exploratory analysis, we regard Table 13 as a construct *model* of development of reading ability, and do not claim that the criterion-referenced benchmarks have handed us a *theory* of development of reading.

5.5 Formative testing and criterion-referenced interpretation of test scores

In classroom applications, the teacher must focus on the total test scores and the associated CR-stages because it is impractical to examine the students' responses to the separate items. We regard the CR-stage as an estimate of the true stage. Table 14 uses simulated test scores of students to illustrate how this estimate functions in practice. It includes the true values of θ and the true stages associated with θ together with the probabilities that the estimated CR-stage is equal to the true stage.

Table 14.
Interpretation of simulated scores from the log-linear Rasch model. Information written in bold is available to the teacher. Misclassified CR-stages are written in red.

	True values			**Observed**	Estimates	
Student	θ	Stage	P*	**score**	$\hat{\theta}$	**CR-stage**
1	-3.0	A	99.2 %	**3**	-3.298	**A**
2	-2.0	A	62.2 %	**5**	-2.336	**A**
3	-1.5	B	73.2 %	**10**	-1.446	**B**
4	-1.0	B	46.6 %	**14**	-0.948	**C**
5	-0.5	C	59.3 %	**18**	-0.524	**C**
6	0.0	D	70.5 %	**23**	-0.025	**D**
7	0.5	D	87.0 %	**27**	0.378	**D**
8	1.0	E	54.1 %	**34**	1.308	**E**
9	1.5	E	73.0 %	**37**	2.032	**F**
10	2.0	E	46.4 %	**37**	2.032	**F**
11	2.5	F	82.1 %	**36**	1.744	**E**
12	3.0	F	95.2 %	**38**	2.407	**F**

P*: A measure of accuracy. The probability that the score assigns the student to the true CR-stage.

Table 14 illustrates the main point that we hoped to make when we set up the experiment; namely, that criterion-referenced interpretation of test scores during formative classroom testing is not only possible in principle but it is also easy and practical. All it takes is a table with columns 1, 2, and 5 of Table 13. In addition to this, Table 14 discloses two issues of concern, both of which turn up in connection with the results of Student 4.

The first is that there is a considerable risk of misclassification. Student 4 is at Stage B because the true θ is equal to -1. We refer to the probability that the CR-stage is equal to the true stage as the accuracy of the estimate. In table 13 we can see that the estimate of the CR-stage is equal to B if the total score is larger than or equal to 8 and less than or equal to 13. To assess the accuracy of the estimate of the CR-stage, we therefore calculate the probability that this is true anchored at θ = -1, and since this is equal to 0.466, we say that the accuracy is only 46.6 % for students with θ = -1 and

are not surprised that this is one of three misclassification cases in Table 14. Student 4's score is equal to 14, and the criterion-referenced interpretation is that the CR-stage is C instead of B.

The second is that the interpretation of CR-stages in Table 14 refers to the scale-anchored probabilities at the midpoint of the stages. Users of Table 13 therefore must be aware that interpretation may be biased in cases where θ is close to a benchmark. To see this for Student 4, we compare the interpretations of CR-stages B and C in Table 13, with the response probabilities anchored at the estimate $\hat{\theta} = -0.948$ in Table B9 of Appendix B. The true CR-stage is B, and Table 13 interprets retrieval of info, inferencing and interpreting as sporadic while evaluating is absent. Table B9 of Appendix B, interpreting the stage a little above the true value of θ, disagrees, insisting that retrieval of information and inferencing are emerging.

The only solution to this problem is to include tables with scale-anchored probabilities together with Table 13. Since this is impractical in connection with classroom applications of the test, the user must be aware that there may be a considerable risk of misclassification and bias of interpretations when test scores are equal to or very close to a benchmark and therefore focus on interpretations at both sides of the benchmark.

6. Analysis of criterion validity and accuracy

Interpretation of quantitative measurement of latent traits must assume that the measurement satisfies conventional requirements of psychometric measurement, e.g. the requirements of criterion-related construct validity defined by Rosenbaum (1989). The assumption of local independence implies that items are conditionally independent given the latent trait variable. Log-linear Rasch models with local dependence violate this assumption of local independence, but Kreiner (2007) and Kreiner and Christensen (2007) replace the assumption of zero partial correlation with the assumption that the partial association of items is uniform, defined by the same log-linear interaction parameters for all values of the latent trait.

Validity of criterion-referenced interpretation requires much more than valid quantitative measurement. Hambleton (1980) describes criterion-referenced tests consisting of items from a single item domain so that raw scores may be interpreted as domain scores and elaborates on test score validity and reliability described in terms of classical test theory and on item validity and "representativeness" depending on the degree to which items fit the definitions of the item domain. Hambleton also includes appendices describing how to review item validity. Following these guidelines to assess the item validity in PIRLS from the point of view of Hambleton would be interesting and maybe useful, but it is beyond what we can do in this paper.

The problem is that criterion-related construct validity of quantitative measurement does not imply that the interpretation of test scores is valid and trustworthy. The next subsections describe one way to test criterion-related validity of interpretation by examination of association between alternative interpretations and of trustworthiness or accuracy by calculation of scale-anchored risks of misclassification.

6.1 Criterion validity

The best way to assess the criterion *validity* of the CR-stages would be to examine the association between the CR-stages and PIRLS's proficiency categories. To do this, we must classify Booklet 16 scores in the same way that PIRLS would do it. And to do that, we must estimate the locations of PIRLS's benchmarks on the log-linear Rasch model's proficiency scale.

We assume that the sample of Danish students responding to Booklet 16 is a representative subsample of the complete sample of Danish students that took part in PIRLS 2016 and use the

information on the distribution of the Danish students across the PIRLS-categories provided by Mejding et al. (2017). Comparison of the percentiles of Mejding's distribution and the percentiles of the distribution of the estimates of θ in the sample of students responding to Booklet 16 provides the locations of PIRLS's benchmarks.

We used the 438 students with complete responses to items to estimate the person parameters and to test the fit of the model. However, to estimate the distribution of the complete sample of Danish students, we must include estimates of θ of the 124 students with incomplete responses. Figure 7 shows the distributions of the scores of the students with complete responses and the expected total scores defined by the estimates of θ of the subsample with incomplete responses. The subsample of students with incomplete responses includes a larger frequency of students at lower levels of proficiency than the subsample of students with complete responses.

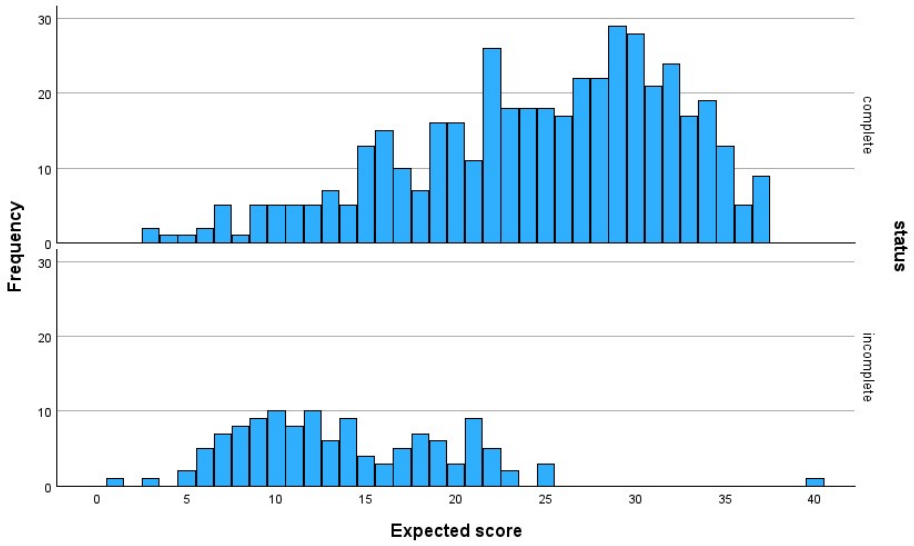


Figure 7.
The distribution of the observed scores for students with complete responses to items and the distribution of the expected total scores of student with incomplete responses

Table 15 shows the PIRLS-categories defined by the 3rd, 15th, 48th, and 89th percentiles of the distribution of the estimates of θ among the complete sample of Danish students, together with the distributions of the sample and subsamples of Danish students with responses to Booklet 16 items.

Table 15.
Estimates of score intervals that define the Proficiency levels of PIRLS 2016

PIRLS-categories	Distribution of Danish students	Score intervals defining PIRLS-categories	Danish students responding to Booklet16		
			Students with complete responses	Students with incomplete responses	Total
Below low level	3%	0-6	1.4 %	7.3 %	2.7 %
Low level	12 %	7-12	5.9 %	41.9 %	13.8 %
Intermediate	33 %	13-22	28.8 %	46.0 %	32.6 %
High level	41 %	23-32	49.5 %	4.0 %	39.5 %
Advanced level	11 %	33-40	14.4 %	0.8 %	11.4 %

Table 16 shows the joint distribution of the CR-stages and PIRLS' categories among students with complete responses to items. It is not surprising that there is a strong positive association between the two classification systems, because both classification systems define categories by score groups. However, the association is not only strong; it is very strong. Goodman and Kruskal's gamma is equal to 1.00.

Table 16.
The joint distribution of students with complete responses to items according to PIRLS-categories and CR-stages.

CR-stage	PIRLS-category						
	-	Low	intermediate	High	Advanced	Total	%
A	6	5				11	2.5 %
B		21	7			28	6.4 %
C			93			93	21.2 %
D			26	193		219	50.0 %
E				24	54	78	17.8 %
F					9	9	2.1 %
Total	6	26	126	217	63	438	1.000
%	1.4 %	5.9 %	28.8 %	49.5 %	14.4 %	1.000	

Strong association between quantitative scores is usually taken as evidence of criterion validity, but criterion validity of qualitative statements asks for more than strong correlation. We need to compare what the CR-stages and the PIRLS-categories have to say about what students can and cannot do.

Consider, for instance, the students at CR-stage B. PIRLS classifies 75 % of these students at the low level and 25 % at intermediate level of proficiency. CR-stage B interprets abilities of information retrieving, inferencing, and interpretation as sporadic and says that evidence of evaluating is absent.

To us, this implies that students at this stage are probably unable to read and understand literary texts like Mary's Red Hen and informational texts like The Green Turtle. Mullis et al. (2017, p. 60 and 75) describe the reading abilities of the students in more positive terms. At the low level, they claim that:

When reading predominantly simpler Literary Texts, students can:

- Locate and retrieve explicitly stated information, actions, or ideas
- Make straightforward inferences about events and reasons for actions
- Begin to interpret story events and central ideas

When reading predominantly simpler Informational Texts, students can:

- Locate and reproduce explicitly stated information from text and other formats
- Begin to make straightforward inferences about explanations, actions and descriptions

At the intermediate level, Mullis et. al. (2017, p. 71) claim that

When reading a mix of simpler and relatively complex Literary Texts, students can:

- Independently locate, recognise and reproduce explicitly stated actions, events, and feelings
- Make straightforward inferences about the attributes, feelings, and motivations of main characters
- Interpret obvious reasons and causes, recognise evidence, and give examples
- Begin to recognise language choices

When reading a mix of simpler and relatively complex Informational Texts, students can:

- Locate and reproduce two or three pieces of information from the text
- Make straightforward inferences to provide factual explanations
- Begin to interpret and integrate information to order events

Comparison of other CR-stages and PIRLS-categories reaches the same conclusion. PIRLS-categories have a more positive view of the abilities of the students than the corresponding CR-stages. It is clearly not up to us to claim that PIRLS exaggerates the abilities and that CR-grades are closer to the true state, but to claim that CR-stages and/or PIRLS-categories are criterion valid, we must comprehend the discrepancies.

One reason could be that we and PIRLS interpret test results relative to very different criteria. The criteria on which the CR-stages are described in this paper. PIRLS refers to their four comprehension processes, but that does not have to imply that PIRLS agrees with our criteria. A second reason could be that our criteria refer to texts that PIRLS describes as relatively complex texts, whereas the low and intermediate PIRLS-categories refer to simpler texts.

The third, and perhaps the greatest difference is that CR-stages refer to reading without help and therefore disregard MC-items, whereas PIRLS-categories describe reading with the help provided by the relatively easy MC-items. We do not claim that MC-items are inappropriate items and that MC-items facilitate random guessing. MC-items facilitate qualified guessing where students may reject some alternatives before they select the response. But we do claim that reading with the kind of help provided by MC-items is less than realistic reading. Responses to MC-items are not relevant if we want to understand what the students can and cannot do when they read on their own.

Tables 17 and 18 show the response probabilities of all items asking for straightforward inferencing anchored at PIRLS's low and intermediate benchmarks. They may throw concrete light on how the inclusion of MC-items influences interpretation. In Table 16, 80 % of the students in the low proficiency category are at CR-stage B, and our interpretation is that the students are unable to read and understand the texts. Table 17 appears to support this interpretation. Three out of four items with open responses and three out of seven MC-items have probabilities of incorrect responses that are larger than 75 %, and there is only one MC-item where the probability of a correct response is larger than 50 %.

Table 18 describes inferencing close to the top of CR-stage B and at PIRLS's intermediate benchmark. The distributions of the items with open responses provide evidence that inferencing is emerging. At this level, PIRLS can defend their interpretation because they include MC-items. Four out of seven items have probabilities of correct responses that are larger than 50 %, and there are only two items (T09 and T03) where evidence of developing proficiency is sporadic.

It is up to the readers to decide whether they agree or disagree with our decision to disregard MC-items during the interpretation of test results. The main point is that these examples illustrate that criterion-referenced interpretation of quantitative test results is complicated and that in addition to the definitions of criteria, there are several things that must be taken into account. PIRLS-categories consider reading of both simple and complex texts and include responses to multiple-choice items, whereas CR-stages base the interpretation on reading of relatively complex texts and only include responses to items where students must solve the problems without help. For these reasons, interpretations must differ. However, since we can understand why the interpretations are different and can see that both are consistent on their own terms, we claim that the very strong statistical association in Table 16 supports our claims of criterion validity.

Table 17.
Response probabilities of items asking for inferencing anchored at PIRLS's low benchmark Score = 7.

Item	Reponse type	Item scores 0	1	2	Mean	Interpretation
T01	MC	0.356	0.644		0.64	Developing
H11	MC	0.578	0.422		0.37	Emerging
T06	**Open**	**0.750**	**0.250**		**0.25**	**Emerging**
T12	MC	0.584	0.416		0.42	Emerging
H05	MC	0.748	0.252		0.25	Emerging
H03	**Open**	**0.862**	**0.138**		**0.14**	**Sporadic**
T04	**Open**	**0.889**	**0.111**		**0.11**	**Sporadic**
T09	MC	0.938	0.062		0.06	Absent
T03	**Open**	**0.926**	**0.065**	**0.009**	**0.06**	**Absent**
H09	MC	0.858	0.142		0.14	Sporadic
T13	MC	0.886	0.114		0.09	Sporadic

Note: Information on items with open responses in bold numbers.

Table 18.
Response probabilities of items asking for inferencing at PIRLS's intermediate benchmark Score = 13.

Item	Response type	Item scores 0	1	2	Mean	Interpretation
T01	MC	0.159	0.841		0.84	Sustained
H11	MC	0.319	0.681		0.68	Developing
T06	**Open**	**0.485**	**0.515**		**0.52**	**Developing**
T12	MC	0.365	0.635		0.64	Developing
H05	MC	0.548	0.452		0.45	Emerging
H03	**Open**	**0.654**	**0.346**		**0.35**	**Emerging**
T04	**Open**	**0.733**	**0.267**		**0.27**	**Emerging**
T09	MC	0.805	0.195		0.20	Sporadic
T03	**Open**	**0.763**	**0.182**	**0.055**	**0.29**	**Sporadic**
H09	MC	0.712	0.288		0.29	Emerging
T13	MC	0.745	0.255		0.26	Emerging

Note: Information on items with open responses in bold numbers.

6.2 Accuracy

We refer to the probability of correct classification as the accuracy of the interpretation by CR-stages to avoid confusing the degree to which the classification into CR-stages is precise with measurement error and reliability of quantitative measures.

Table 19 shows that accuracy is an important issue and that accuracy depends on the location of the student. If the student is close to a benchmark, the risk of misclassification will be close to 50 %. In this section, we define accuracy by estimates of scale-anchored probabilities of misclassification at the midpoints of the categories. In connection with Rasch models, this is particularly easy, because

we can estimate anchor-scaled distribution of the scores at the midpoints of the stages. Table 19 shows the results for CR-stages B-E, but does not include the extreme A and F stages where there are no natural midpoints of θ.

Table 19.
Scale-anchored probabilities of criterion-referenced classification

Midpoint of true stage	Observed stages defined by score intervals					
	A	B	C	D	E	F
	0-7	8-13	14-21	22-31	32-36	37-40
B	9.3 %	**72.8 %**	17.8 %	0.02 %		
C	0.02 %	7.1 %	**79.8 %**	13.1 %		
D			3.6 %	**89.8 %**	6.6 %	0.02 %
E				13.3 %	**73.9 %**	12.8 %

Table B10 of Appendix B has classification probabilities anchored at WLE estimates of θ for all scores from 1 to 39. There is little risk of misclassification at the extreme CR-stages and for stages B-E, the risk of misclassification is below 35 % at one point above the benchmark. Apart from that, the highlighted probabilities of correct classification in Table 19 provide an honest assessment of the reliability of the CR-stages.

7. Discussion

7.1 Criterion-referenced testing: sixty wasted years.

Even though many refer to Glaser's (1963) paper on criterion-referenced testing as a seminal paper, it appears that the notion of criterion-referenced testing has not lived up to its promise. A few points on the timeline from 1963 until today do tell a discouraging story:

- In 1985, more than twenty years later, Haertel (1985, p 41) claims "that a unified framework for the validation of criterion-referenced test interpretations has not emerged".
- In 2005, more than forty years later, Wilson (2005) refers to "Glaser's seminal paper" in a section on "Next Steps in Measuring" discussing future perspectives.
- In 2014, Popham (2014), who had been seriously involved in criterion-referenced testing, looks back on "Criterion-referenced measurement: Half a century wasted".
- Finally, in 2023, the final chapter of Wilson (2023) repeats the reference to Glaser in a way that suggests that nothing has happened since 2005.

The requirement that educational tests should be designed expressly for the interpretation of a student's performance in terms of what he or she can or cannot do, irrespective of other students, must be above dispute. Therefore, the question is why criterion-referenced testing never took flight despite an explosion of interest in the 1970s, that Berk (1980, p. 4) claims "had not yet abated" in 1980. However, the collection of papers edited by Berk (1980) provides one possible explanation. It describes the state of the art of criterion-referenced testing in 1980, but it appears to disregard what was happening in psychometrics at the same time. In 1980, criterion-referenced testing had not gone beyond the limitations imposed by classical test theory (CTT), and there were no references at all to IRT and Rasch models.

The limitations of CTT define limitations on the kind of criterion-referenced tests that the authors contributing to Berk (1980) described. Criterion-referenced tests were supposed to consist of a homogeneous set of congruent items from a single item domain with interpretable test or domain scores, whereas conventional educational tests consisted of items from several vaguely specified item domains and therefore provided test scores and inaccurate interpretations of abilities.

Under CTT this was the only way forward for criterion-referenced tests, but contemporary psychometrics and IRT and Rasch models provide several alternative possibilities, because IRT models may include items from several item domains so that item analyses may focus on and provide domain-referenced interpretation of responses from single domains. In this paper, we have taken steps towards a methodological framework for domain- and criterion-referenced interpretation by anchor scaling supported by unidimensional IRT and/or Rasch models, but the methodology extends to, and we are also using it for, multi-dimensional mixed Rasch models.

7.2 Scale-anchored probabilities and distributions

Anchor scaling is founded on a single presumption: to understand and interpret educational test results in meaningful and useful terms, we must estimate and examine scale-anchored probabilities of responses to items. Interpreting test results in terms of probabilities is not an original idea. It can be found in Rasch (1960, p.11) where it was made clear "that we can never know *with certainty* how a pupil will react to a problem, but we may say whether *he has a good or poor chance of solving it,* and therefore that *the behavior of a pupil is described by means of a probability that he solves the task"*. Forsyth (1991) and Beaton and Allen (1992) were the first to describe how to interpret tests by scale anchoring in practice, and PIRLS has interpreted test scores by scale anchoring since 2001.

Beaton and Allen (1992) distinguished between direct and smoothing calculation of estimates of scale-anchored probabilities of items. We have adopted PIRLS's terminology, but we calculate scale-anchored probabilities by smoothing, whereas PIRLS's estimates are direct.

Another difference may be that we interpret the anchored probabilities in terms that refer to different stages of development of proficiency, from none to consolidated ability. We are aware that PIRLS invests considerable time in the interpretation of the scale-anchored probabilities at their benchmarks, but we have no information on how they do it in practice. A third difference is that we interpret the proficiency associated with a stage at the midpoint of the score interval defining the stage, whereas PIRLS interprets the stage proficiency at the benchmark that defines the low point of the stage.

These differences are minor technical differences that in themselves may not result in very different outcomes, but the fourth difference is important and may perhaps be our only original contribution in this paper. We define benchmarks by scale anchoring that refer explicitly to criteria that proficiency must meet. Beaton and Allen (1992) and PIRLS select convenient, arbitrary benchmarks, and they do not calculate scale-anchored probabilities before the benchmarks have been selected. We do not claim that PIRLS's interpretation is meaningless and useless, quite the opposite in fact. However, we claim that their interpretation is not criterion-referenced.

7.3 The frames of reference and inference

Criterion-referenced interpretation of test scores is only possible and useful if three requirements are satisfied. The first is that we have a construct theory that refers to a number of domains that characterize aspects of the latent trait and define the criteria that proficiency must satisfy to be consolidated. The second is that we have subsets of items associated with the domains, and that the test provides content valid quantitative measurement, because the set of items covers all the construct-domains and all other secondary construct characteristics.

The third is that we have a plausible and realistic psychometric measurement model. Criterion-referenced interpretation uses this model for two purposes: to provide quantitative measures of proficiency and to provide estimates of scale-anchored distributions that we need to understand what the quantitative test results are telling us. To avoid confounding of the estimates, it is essential that the estimates of the scale-anchored probabilities are asymptotically unbiased and precise. It is therefore essential that the model is more than a model of convenience that can provide estimates. Careful item analysis must support the claims that the model actually fits the data.

7.4 PIRLS

PIRLS is of considerable interest, but this paper only uses PIRLS data to illustrate the methodology. PIRLS is the perfect example for this because it provides the theoretical framework and the item domains that we need, together with extensive documentation on the way PIRLS classifies and interprets test results.

PIRLS defined norm-referenced benchmarks in 2001 but modified them in 2006. Since they are no longer norm-referenced, we regard PIRLS's proficiency categories as defined by benchmarks of convenience. The criterion-referenced stages must be positively correlated, but we did not expect a close to deterministic relationship. We take it for granted that PIRLS's definition of domains is adequate, and we have confidence in the methods leading from item-referenced benchmarks to the benchmarks that define the criterion-referenced stages of development. Experts who are more familiar with the theoretical framework of PIRLS may disagree with some details, but the strong association between our criterion-referenced stages and PIRLS's proficiency categories support our claims that the criterion-referenced interpretation of quantitative measurement by Booklet 16 is valid.

Despite this, we and PIRLS interpret test results in somewhat different terms. Our interpretation is less optimistic than PIRLS's because we do not only focus on what students can do, but also on what they cannot do when they read texts like those in Booklet 16. Another reason could be that PIRLS is accountability testing. Mullis and Prendergast (2017) state that the PIRLS-categories are meant to "provide as much information as possible for policy and curriculum reform" (p. 13.1), whereas we are more concerned with formative classroom testing. PIRLS's terminology may reflect what PIRLS expects of fourth grade reading. We focus on texts similar to the two texts of Booklet 16 in a way that does not take the grade into account.

7.5 Criterion-referenced interpretation in practice.

Developing criterion-referenced interpretation for practical applications takes much more than what this paper has described. If we had intended to suggest that PIRLS should use our criterion-referenced benchmarks and stages of development, we would tested for differential item functioning across countries or across subpopulations defined by covariates that could have a direct effect on responses to items. If evidence of DIF turns up, it may only be a practical problem if a log-linear Rasch model fits the data, but we would probably have to abandon the notion of criterion-referenced interpretation that applies beyond the current frame of reference if the fit of the log-linear Rasch model is rejected.

To understand how the interpretation functions in practice, we should collect Booklet 16 data at the beginning and the end of the fourth grade to make sure that measurement is invariant and to estimate the effect of the education in the fourth grade. If there is close to no difference between the distribution across stages at the beginning and the end of the fourth grade, we would probably have to admit that the sensitivity of the *interpretation* is inadequate, because it is unable to describe how reading ability develops.

Finally, to test the validity of the stages, we would develop criterion-referenced stages for all the booklets of PIRLS and compare them with the stages defined by Booklet 16 in the same way that we compared the Booklet 16 stages with PIRLS's proficiency categories. This would be the true test of fire of our criterion-referenced stages. If it fails, it indicates that there are item validity issues. We would consult the guidelines for improving items of criterion-referenced tests suggested by Hambleton (1980), and we would ask the subject matter experts responsible for the items to tell us what the problems are.

7.6 Log-linear Rasch models

Log-linear Rasch models with locally dependent items may be unfamiliar to some readers. We therefore stress that our methodology applies to other kinds of IRT models, including multidimensional models where items from different domains depend on different latent traits. Analyses by Rasch and log-linear Rasch models are expedient for many reasons, and it is of interest that conditional maximum likelihood estimates of item parameters– and therefore also the estimates of scale-anchored probabilities and the criterion-referenced interpretation– satisfy the same requirement as specific objective measurement requires (Kreiner, 2007; Kreiner & Christensen, 2007). Interpretation is sample-free and does not depend on or refer to distributions of students.

We decided to abandon the convenient Rasch model that PIRLS uses for national analyses, because Booklet 16 consists of two testlets in which it is inconceivable that all items are locally independent. We used the log-linear Rasch model, which is consistent with the Rasch model for item bundles described by Wilson and Adams (1995), but provides the information on response probabilities that we need to estimate scale-anchored distributions. Baghaei et al. (2025) used the Rasch testlet model of Wang and Wilson (2005) for analyses of PIRLS data from 2016 that included the Danish data on The Green Turtle. The Rasch testlet model describes trait dependence among items, adding random effects to the model. We do not claim that trait dependence plays no role, but the log-linear Rasch model describes response dependence among items that we regard as a more plausible reason for local dependence in testlets. We are relatively confident about the fit of our model. Appendix A provides some, but not all our evidence supporting the model, but we do not claim that additional analyses would not have been able to disclose fit issues. The test of unidimensionality comparing the observed and expected correlations of the subscores defined by Macy's Red Hen and The Green Turtle rejected unidimensionality with a one-sided $p = 0.036$. Since we do not regard p-values of this size as strong evidence against the model, and because we did not disclose significant evidence of a negative association between items from the two testlets, we accepted the model, but we acknowledge that the weak evidence against unidimensionality could be due to trait dependence.

Acknowledgements

The authors acknowledge several helpful comments from the referee.

Funding

The authors received no specific funding for this work from any funding agencies.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability Statements

The Danish data on responses to Booklet 16 can be found on the DIGRAM homepage, https://biotat.ku.dk/digram.

How to Cite

Kreiner, S., Müller, M., & Nielsen, T. (2026). Criterion-referenced interpretation through scale anchoring: illustrated by analysis of Danish results from PIRLS 2016. *Educational Methods & Psychometrics, 4*, 31.

References

Andersen, E. B. (1970). Asymptotic properties of conditional maximum-likelihood estimators. *Journal of the Royal Statistical Society: Series B (Methodological), 32*(2), 283–301. https://doi.org/10.1111/j.2517-6161.1970.tb00842.x

Andersen, E. B. (1973). A goodness of fit test for the Rasch model. *Psychometrika*, *38*(1), 123-140. https://doi.org/10.1007/BF02291180

Andrich, D. (1978). Application of a psychometric rating model to ordered categories which are scored with successive integers. *Applied Psychological Measurement, 2*(4), 581–594. https://doi.org/10.1177/014662167800200413

Baghaei, P., Ravand, H., & Striethолt, R. (2025) Modeling local item dependence in PIRLS 2016: Comparing ePIRLS and paper-based PIRLS using the Rasch testlet model. *Psychological Test and Assessment Modeling*. *67*, 127–140. https//doi.org/102440/001-0021

Beaton, A. E., Allen, N. L. (1992) Interpreting scales through scale anchoring. *Journal of Educational Statistics*, *17*, 191–204. https//doi.org/10.2307/1165169

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological), 57*(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x

Berk, R.A. (Ed.) (1980). *Criterion-referenced measurement: The state of the art*. John Hopkins Press.

Bruggink, M., Swart, N., Van der Lee, A., & Segers, E. (2022). *Putting PIRLS to use in classrooms across the globe: Evidence-based contributions for teaching reading comprehension in a multilingual context*. Springer Nature. https://doi.org/10.1007/978-3-030-95266-2

Clifford, R. (2016). A rationale for criterion-referenced proficiency testing. *Foreign Language Annals*, *49*(2), 224-234. https://doi.org/10.1111/flan.12201

de Vet, H. C. W., Terwee, C. B., Mokkink, L. B., Knol, D. L. (2011). *Measurement in Medicine: A Practical Guide*. Cambridge University Press. https//doi.org/10.1017/CBO9780511996214

Forsyth, R. A. (1991). Do NAEP Scales Yield Valid Criterion-Referenced Interpretations? *Educational Measurement: Issues and Practice, 10*, 3-9. https//doi.org/10.1111/j.1745-3992.1991.tb00197.x

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American psychologist*, *18*(8), 519. https://doi.org/10.1037/h0049294

Glaser, R. (1971). A Criterion-Referenced Test. In Popham, W. J. (Ed.). *Criterion-Referenced Measurement: An Introduction.* Educational Technology Publications. 37-50

Goodman, L. A., & Kruskal, W. H. (1954). Measures of association for cross classifications. *Journal of the American Statistical Association, 49*(268), 732–764. https://doi.org/10.1080/01621459.1954.10501231

Haertel, E. (1985). Construct validity and criterion-referenced testing. *Review of Educational Research*, *55*(1), 23-46. https://doi.org/10.3102/00346543055001023

Hambleton, R.A. (1980). Test Score Validity and Standard-Setting Methods. In Berk R.A (Ed.) (1980) *Criterion-referenced Measurement: The state of the art*. (pp. 80-124), John Hopkins Press.

Hambleton, R. K., Zenisky, A. L. (2013). Score reporting and Interpretation. In Linden, W. J. v d (Ed.) *Handbook of Item Response Theory. Volume Three. Applications*. (pp. 127-142), CRSC Press.

Kelderman, H. (1984). Loglinear Rasch model tests. *Psychometrika*, *49*(2), 223-245. https://doi.org/10.1007/BF02294174

Kelderman, H. (1989). Item bias detection using loglinear IRT. *Psychometrika*, *54*(4), 681-697. https://doi.org/10.1007/BF02296403

Kelderman, H. (1992). Computing maximum likelihood estimates of loglinear models from marginal sums with special attention to loglinear item response theory. *Psychometrika*, *57*(3), 437-450. https://doi.org/10.1007/BF02295431

Kreiner, S. (1993/2006). Validation of index scales for analysis of survey data. I K. Dean (Ed.) (1993) *Population Health Research: Linking Theory and Methods* (s. 116–144). Sage Publications. (Reprinted in Bartolomew, D. J. (Ed.) (2006) *Measurement, Vol. III* (s. 297–328), Sage).

Kreiner S (2007) Validity and objectivity. Reflections on the role and nature of Rasch models. *Nordic Psychology*, *59*, 268-298. https//doi.org/10.1027/1901-2276.59.3.268

Kreiner, S. (2011). A note on item-restscore association in Rasch models. *Applied Psychological Measurement, 35*(7), 557–561. https://doi.org/10.1177/0146621611410227

Kreiner, S (2025). On specific objectivity and measurement by Rasch models: A statistical viewpoint. *Educational Methods & Psychometrics, 3:* 20. https//doi.org/10.61186/emp.2025.7
Kreiner, S. & Nielsen, T. (2026). One or two Rasch models for item bundles. Submitted.
Kreiner, S., & Christensen, K. B. (2002). Graphical Rasch models. In *Statistical methods for quality of life studies: Design, measurements and analysis* (pp. 187-203). Springer US. https://doi.org/10.1007/978-1-4757-3625-0_15
Kreiner, S., & Christensen, K. B. (2004). Analysis of local dependence and multidimensionality in graphical loglinear Rasch models. *Communications in Statistics-Theory and Methods*, *33*(6), 1239-1276. https://doi.org/10.1081/STA-120030148
Kreiner, S., & Christensen, K. B. (2007). Validity and objectivity in health-related summated scales: Analysis by graphical loglinear Rasch models. I M. von Davier & C. H. Carstensen (Eds.), *Multivariate and mixture distribution Rasch models: Extensions and applications* (pp. 329–346). Springer. https://doi.org/10.1007/978-0-387-49839-3_21
Kreiner, S., & Christensen, K. B. (2011a). Item screening in graphical loglinear Rasch models. *Psychometrika*, 76(2), 228-256. https://doi.org/10.1007/s11336-011-9203-y
Kreiner, S., & Christensen, K. B. (2011b). *Exact evaluation of bias in Rasch model residuals. Advances in Mathematics Research, 12*, 19-40.
Kreiner, S., & Christensen, K. B. (2013). Person parameter estimation and measurement in Rasch models. In K. B. Christensen, S. Kreiner, & M. Mesbah (Eds.), *Rasch models in health* (pp. 63-78). ISTE & John Wiley & Sons. https://doi.org/10.1002/9781118574454.ch4
Kreiner, S., & Nielsen, T. (2023). Item analysis in DIGRAM 5.01: Guided tours. Department of Biostatistics, University of Copenhagen. https://biostat.ku.dk/DIGRAM/Item%20analysis%20in%20DIGRAM%205-01%20-%20guided%20tours.pdf
Lauritzen, S. L. (1996). *Graphical models.* Clarendon Press. http://dx.doi.org/10.1093/oso/9780198522195.001.0001
Martin, M. O., Mullis, I. V., & Kennedy, A. M. (2007). *PIRLS 2006 Technical Report.* TIMSS & PIRLS International Study Center, Boston College. Retrieved from: https://timssandpirls.bc.edu/PDF/p06_technical_report.pdf
Martin, M. O., Mullis, I. V. S., & Hooper, M. (Eds.) (2017). *Methods and Procedures in PIRLS 2016.* TIMSS & PIRLS International Study Center, Boston College. Retrieved from: https://timssandpirls.bc.edu/publications/pirls/2016-methods.html
Mejding, J., Neubert, K., & Larsen, R. (2017). *PIRLS 2016. En international undersøgelse om læsekompetence i 3. og 4. klasse: Rapport.* Aarhus Universitetsforlag. https://unipress.dk/udgivelser/p/pirls-2016/
Mullis, I. V. S., Martin, M. O., González, E. J., & Kennedy, A. M. (2003). *PIRLS 2001 International Report.* International Association for the Evaluation of Educational Achievement (IEA), International Study Center, Lynch School of Education, Boston College. Retrieved from: https://timssandpirls.bc.edu/pirls2001i/pdf/p1_IR_book.pdf
Mullis, I. V. S., & Martin, M. O. (Eds.). (2015). *PIRLS 2016 Assessment Framework* (2nd ed.). Retrieved from: https://timssandpirls.bc.edu/pirls2016/framework.html
Mullis, I. V. S., Martin, M. O., Foy, P., & Hooper, M. (2017). *PIRLS 2016 International Results in Reading.* International Association for the Evaluation of Educational Achievement (IEA), TIMSS & PIRLS International Study Center, Boston College. Retrieved from: https://eric.ed.gov/?id=ED580353
Mullis, I. V. S., & Prendergast, C. O. (2017). Using scale anchoring to interpret the PIRLS and ePIRLS 2016 achievement scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and Procedures in PIRLS 2016* (pp. 13.1 – 13.23). International Association for the Evaluation of Educational Achievement (IEA), TIMSS & PIRLS International Study Center, Boston College. Retrieved from: https://timssandpirls.bc.edu/publications/pirls/2016-methods/chapter-13.html
Popham, W. J. (Ed.) (1971). *Criterion-Referenced Measurement: An Introduction.* Educational Technology Publications.
Popham, W. J. (2009). *Unlearned Lessons.* Harvard Education Press.
Popham, W. J. (2014). Criterion-referenced measurement: Half a century wasted? *Educational Leadership, 71*, 62-66.
Popham, W. J., & Husek, T. R. (1969). Implications of criterion-referenced measurement. *Journal of Educational Measurement, 6*, 1-9. https://doi.org/10.1111/j.1745-3984.1969.tb00654.x
Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks pædagogiske Institut.
Reckase, M. D. (2017). A tale of two models: Sources of confusion in achievement testing (ETS Research Report RR-17-44). Educational Testing Service. https://doi.org/10.1002/ets2.12171
Rosenbaum, P. (1989). Criterion-related construct validity. Psychometrika, 54, 625-633. https://doi.org/10.1007/BF02296400
Sijtsma, K., & Hemker, B. T. (1998). Nonparametric polytomous IRT models for invariant item ordering, with results for parametric models. *Psychometrika*, *63*(2), 183-200. https://doi.org/10.1007/BF02294774
Wang, W.-C., & Wilson, M. (2005). The Rasch Testlet Model. *Applied Psychological Measurement*, *29*, 126-149. https://doi.org/10.1177/0146621605276281
Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika*, *54*(3), 427-450. https://doi.org/10.1007/BF02294627
Wilson, M. (2005). *Constructing Measures. An Item Response Approach*., Psychology Press.
Wilson, M. (2023). *Constructing Measures. An Item Response Approach. Second edition* , Psychology Press. https://doi.org/10.4324/9781003286929
Wilson, M., & Adams, R.J. (1995). Rasch models for item bundles. *Psychometrika*, *60*,181-198. https://doi.org/10.1007/BF02301412
Wright, B. D. (1980). Afterword. In Rasch (1960/1980) *Probabilistic Models for some Intelligence and Attainment tests. 50th anniversary edition.* Mesa Press.
Wright, B. D. & Masters G. N. (1982) *Rating Scale Analysis: Rasch Measurement*. Chicago: MESA Press

Manuscript Received: 02 DEC 2025
Final Version Received: 11 MAY 2026
Published Online Date: 30 MAY 2026

Appendix A. Analysis of Booklet 16 items by log-linear Rasch models

This appendix provides information on the log-linear Rasch models that we used during the item analysis of Booklet 16 items. Inference was conditional to avoid using estimates of person parameters during the analysis. We calculated conditional maximum likelihood (CML) estimates of item parameters and log-linear interaction parameters, conditional likelihood ratio (CLR) tests for tests of homogeneity and local independence and assessed the significance of the tests of fit by the conditional distributions of fit statistics given the total scores over all items.

We used the DIGRAM program described by Kreiner and Nielsen (2023) during the analysis. DIGRAM is available free of charge at https://biostat.ku.dk/DIGRAM together with user guides, papers describing item analysis by graphical log-linear Rasch models, and data on responses to the Booklet 16 items.

The appendix consists of discreetly edited DIGRAM output. The initial analysis rejected the Rasch model. For this reason, the subsequent analysis consisted of stepwise model search for a log-linear Rasch model with local dependence among some items. During each step of the analyses, we assessed the adequacy of alternative models by CLR tests together with the same kind of item fit statistics that we apply for tests of conventional Rasch models. This appendix includes the tests of fit of the log-linear Rasch model described in the paper, but provides no information on the steps taken towards the final model.

A.1. Items

In addition to the item names used in the paper, we refer to items by the item labels consisting of single letters, when we communicate with DIGRAM. The list of items on the next page presents the labels together with the average item scores for the items for all Danish students and for the students with complete responses on all items. The results presented in this appendix describe analyses by students with responses to all items.

Notice, that we use labels defined by both capital and lower key letters. Items A – P refer to items from the story about Macy's hen, while items Q – f are items from the text on green Turtles.

complete cases

items	n	mean	mean	item range
A: H01_inf	562	0.907	0.916	0 - 1
B: H02_e&c	562	0.870	0.881	0 - 1
C: H03_ifr	562	0.648	0.683	0 - 1
D: H04_i&i	560	0.177	0.185	0 - 1
E: H05_ifr	560	0.714	0.724	0 - 1
F: H06_inf	560	1.443	1.493	0 - 2
G: H07_inf	559	0.506	0.495	0 - 1
H: H08_e&c	557	0.318	0.311	0 - 1
I: H09_ifr	556	0.577	0.584	0 - 1
J: H10_inf	555	0.791	0.817	0 - 1
K: H11_ifr	552	0.879	0.881	0 - 1
L: H12_i&i	550	0.729	0.749	0 - 1
M: H13_i&i	547	1.007	1.105	0 - 3
N: H14_i&i	526	0.519	0.534	0 - 1
O: H15_i&i	514	0.745	0.758	0 - 1
P: H16_e&c	504	0.268	0.276	0 - 1

Q: T01_ifr	557	0.943	0.947	0 - 1
R: T02_inf	557	1.406	1.425	0 - 2
S: T03_ifr	556	1.133	1.174	0 - 2
T: T04_ifr	556	0.644	0.646	0 - 1
U: T05_inf	556	0.606	0.614	0 - 1
V: T06_ifr	555	0.823	0.852	0 - 1
W: T07_i&i	553	0.893	0.922	0 - 2
X: T08_inf	551	0.628	0.678	0 - 1
Y: T09_ifr	546	0.592	0.616	0 - 1
Z: T10_inf	543	0.718	0.753	0 - 1
a: T11_i&i	541	1.079	1.167	0 - 3
b: T12_ifr	535	0.774	0.836	0 - 1
c: T13_ifr	513	0.561	0.584	0 - 1
d: T14_e&c	503	0.545	0.557	0 - 1
e: T15_e&c	499	0.643	0.699	0 - 1
f: T16_e&c	465	0.826	0.845	0 - 1

A.2 The log-linear Rasch model

Log-linear Rasch models assume that the joint conditional distribution of items given the person parameter θ and a set of categorical covariates is log-linear. Since our model does not include exogenous covariates, our analysis did not include tests of DIF. The log-linear model Rasch model is defined by three sets of parameters:

(1) The person parameter θ.

(2) Item parameters that together with θ define the main effects of the model.

(3) Interaction parameters describing local dependence between items.

Had exogenous covariates been included, the model could also include

(3) Interaction parameters describing DIF-interaction between items and covariates.

Log-linear Rasch models assume that interactions are uniform with the interaction parameters that are constant across all values of θ. The total score over items is sufficient for θ, the item margins are sufficient for item parameters and tables counting the joint distributions of items and exogenous covariates are sufficient for LD and DIF parameters. It is for this reason, that inference by log-linear Rasch models may be conditional in exactly the same way as inference by conventional Rasch models.

Log-linear models are models for nominal variables and interaction parameters consist of tables with several parameters for each interaction. Since items are either dichotomous or ordinal, it is useful to have a measure of the strength and direction of local dependencies. To address this issue, we measure the strength of the local dependence defined by the interaction parameters among items by Goodman and Kruskal's (1954) Gamma coefficient γ and we regard a standardized version of this γ as a measure of the partial correlation of the two items given the value of the person parameter.

The odds-ratio (OR) is a natural measure of association between dichotomous items. Since there is a simple relationship between the well-known odds-ratios and the Gamma coefficients in 2-by-2 tables, $\gamma = (OR-1)/(OR+1)$ & $OR = (1+\gamma)/(1-\gamma)$, it follows that the standardized γ is a measure of local dependence that applies for both dichotomous and polytomous items.

During our analyses we distinguish between

strong local dependence where $\gamma \geqq 0.30$ (corresponding to OR $\geqq 1.86$),

moderate local dependence if $0.20 \leqq \gamma < 0.30$ ($1.5 \leqq$ OR < 1.86),

and weak if $\gamma < 0.20$ (OR < 1.5).

Readers are welcome to disagree. We are not proposing a rule of thumb that everybody should use.

Table A.1 shows the estimates of the thresholds and locations of the Partial Credit Model of the locally independent items.

Table A.1
CML estimates of PCM thresholds of locally independent items

item		1	2	3	Location
A:	H01_inf	-2.472	....	....	-2.472
D:	H04_i&i	1.848	....	....	1.848
E:	H05_ifr	-0.893	....	....	-0.893
F:	H06_inf	-1.207	-0.577	....	-0.892
G:	H07_inf	0.228	....	....	0.228
I:	H09_ifr	-0.182	....	....	-0.182
J:	H10_inf	-1.498	....	....	-1.498
R:	T02_inf	-1.567	-0.114	....	-0.841
U:	T05_inf	-0.324	....	....	-0.324
W:	T07_i&i	0.132	0.606	....	0.369
b:	T12_ifr	-1.641	....	....	-1.641
d:	T14_e&c	-0.054	....	....	-0.054
e:	T15_e&c	-0.754	....	....	-0.754

To connect the locally dependent items to the partial credit model, we use a result by Kreiner & Christensen (2007). They refer to subsets of items that are directly or indirectly dependent in log-linear Rasch models as complete item components, if you cannot include more items without adding independent items, and they show that the distributions of component scores over items in *complete* components are like the distributions of polytomous Rasch items. Table A.2 shows the PCM thresholds and locations of the three components defined by the log-linear Rasch model shown in Figures 4 and 5 of the paper.

Table A.2
CML estimates of PCM thresholds of component scores

Component	Range	Thresholds	Location
B+C	0-2	-1.481 -0.745	-1.113
H+K+L+M+N+O+P	0-9	-2.276 -1.435 -0.860 -0.421 0.203 0.684 1.127 1.624 2.308	0.106
Q+S+T+V+X+Y+Z+a+c+f	0-13	-3.063 -1.970 -1.332 -0.969 -0.720 -0.491 -0.261 -0.017 0.240 0.463 0.578 0.732 1.526	-0.406

We are aware that much more needs to be said about the log-linear Rasch models, but have to refer to Kreiner & Nielsen (2023) for additional information and references about these models. The rest of this appendix provides the most important results that support our model.

A.3 Conditional likelihood ratio tests of homogeneity

We used the conditional likelihood ratio (CLR) test of Andersen (1973) to test the homogeneity of items across different score groups. Since the log-linear Rasch model claims that items from different texts are locally independent, we calculated CLR tests of homogeneity for each text separately and for the two texts taken together.

Table A.3 shows the results for the conventional Rasch model and for the log-linear Rasch model that we claim provides an adequate fit to the items of Booklet 16. The CLR tests of the separate texts compare estimates of item parameters in low and high score groups, whereas the CLR tests for the complete set of items compare item parameters in score groups defined by the criterion-referenced stages described in Table 13 of the paper.

Table A.3.
CLR tests of homogeneity

	Rasch model			Log-linear Rasch model		
Text	CLR	df	p	CLR	df	p
Macy's Hen	55.8	18	<0.00005	31.0	27	0.272
The Green Turtle	91.4	20	<0.00005	59.2	39	0.020
Booklet 16^{+}	1550.8	156	<0.00005	308.6	368	0.044

$^{+}$: the test of homogeneity compares item parameters in five score groups: 1-7, 8-13, 14-21, 22-31, 32-39

Table A.3 provides very strong evidence against the conventional Rasch model. Since CLR tests of local independence provided overwhelming evidence of local dependence, we used the item screening procedure of Kreiner & Christensen (2011) to define an initial log-linear Rasch model for stepwise model search among log-linear Rasch models. During this analysis, we calculated three types of tests and added or deleted log-linear interactions if some of these tests rejected the models that we were considering.

The three types of tests were

Andersen's CLR test of homogeneity.

The CLR tests of local independence proposed by Kelderman (1984).

Four different item fit statistics similar to those that we use during tests of fit of conventional Rasch models.

At the end of the analysis, all tests accepted fit to the log-linear Rasch model that we have used in the paper. Sections A.4 and A.5 present the tests of fit that supports the model.

A.4 Tests of local dependence and independence

We distinguish between CLR tests of local *dependence* and CLR tests of local *independence*.

Tests of local dependence refer to pairs of items that the model claims are locally dependent and tests of local independence refer to pairs of items that are locally independent in the model. The stepwise procedure eliminated interaction terms if tests of local dependence were insignificant and added interaction terms if tests of local independence were significant. The stepwise procedure was not automatic. Decisions concerning elimination and/or inclusion of interaction terms did not only depend on p-values, but also on assessment of the strength of the dependencies and on the degree to which the local dependence made sense.

Table A.4 shows the CLR tests that support our claims of local dependence among items. In this table, the gamma coefficients are the standardized gamma coefficients defined by the log-linear interaction parameters. The majority of the gamma coefficients convey evidence of strong local dependence. In some cases, we only claim that the evidence of local dependence provided by the p-values is relatively weak. In these cases, interactions were nevertheless included because of the measure of the strength of the dependency and because they were needed to take care of issues of item misfit and/or lack of homogeneity.

Table A.4 consists of two parts. One with the tests of local dependence in Mary's Hen and the other with the tests of local dependence in The Green Turtle. It is of interest that local dependence in Macy's Red Hen in 6 out of 7 cases involves interpreting and evaluating, whereas local dependence in The Green Turtle involves items asking for straightforward inferencing in 10 of 11 cases and retaining information in 6 cases. We assume that these differences are due to the differences between reading of literary texts and reading informational texts.

Table A.4
CLR tests of local dependence. The Gamma coefficients are standardized gamma coefficients measuring the strength of the local dependence among items

Items	CLR	df	p	Gamma
BC: H02_e&c & H03_ifr	23.17	1	0.0000	0.65
HP: H08_e&c & H16_e&c	6.41	1	0.0113	0.28
KO: H11_ifr & H15_i&i	9.16	1	0.0025	0.44
LN: H12_i&i & H14_i&i	24.00	1	0.0000	0.54
MN: H13_i&i & H14_i&i	13.26	3	0.0041	-0.22
NO: H14_i&i & H15_i&i	16.40	1	0.0001	0.47
OP: H15_i&i & H16_e&c	5.31	1	0.0212	0.36
QZ: T01_ifr & T10_inf	6.22	1	0.0126	0.52
SZ: T03_ifr & T10_inf	19.48	2	0.0001	0.26
Sa: T03_ifr & T11_i&i	19.39	6	0.0036	0.26
TY: T04_ifr & T09_ifr	13.00	1	0.0003	0.37
Tc: T04_ifr & T13_ifr	5.49	1	0.0192	0.24
VX: T06_ifr & T08_inf	5.68	1	0.0172	0.36
VZ: T06_ifr & T10_inf	4.45	1	0.0349	0.32
Va: T06_ifr & T11_i&i	11.55	3	0.0091	0.56
XY: T08_inf & T09_ifr	12.78	1	0.0004	0.38
XZ: T08_inf & T10_inf	9.22	1	0.0024	0.38
Yf: T09_ifr & T16_e&c	4.91	1	0.0267	0.30

Table A.5 shows the cases of significant CLR tests of local *independence*. Benjamini & Hochberg (1995) reject at 0.00046. In the majority of these cases, the local dependencies were weak or at best moderate. Significant evidence of dependence of items from different texts related to the same comprehension process only appeared in two cases.

Table A.5 consists of three parts. One with the tests of local independence between items from Mary's Red Hen. The second with tests of local independence between pairs of items from both texts, and the third with the tests from The Green Turtle. The tests indirectly provide support for the assumption of unidimensionality. If the responses to items need two latent traits, one for literary reading and another for informational reading, we would expect evidence of positive local dependence within text and negative local dependence between texts.

Taken together, the CLR tests of Tables A.4 and A.5 support the fit of the model. However, CLR tests of log-linear interactions is only one part of the analysis by log-linear Rasch model. Section A.5 provides tests of item fit and Section A.6 adds two tests of unidimensionality.

Table A.5
CLR tests of local independence. The Gamma coefficients are partial gamma coefficients calculated during the initial item screening

Items	CLR	df	p	Gamma
AC: H01M_inf & H03_ifr	5.11	1	0.0238	0.35
CO: H03_ifr & H15_i&i	6.27	1	0.0123	0.04
FJ: H06_inf & H10M_inf	6.67	2	0.0356	-0.18
FO: H06_inf & H15_i&i	7.75	2	0.0207	0.11
GM: H07M_inf & H13_i&i	8.05	3	0.0450	-0.19
JO: H10M_inf & H15_i&i	5.71	1	0.0169	0.17
AS: H01M_inf & T03_ifr	6.95	2	0.0310	0.23
CZ: H03_ifr & T10_inf	4.41	1	0.0356	0.04
Ca: H03_ifr & T11_i&i	9.36	3	0.0249	-0.12
DU: H04_i&i & T05M_inf	6.61	1	0.0102	-0.22
DZ: H04_i&i & T10_inf	4.46	1	0.0346	0.39
ER: H05M_ifr & T02_inf	8.43	2	0.0148	-0.15
ET: H05M_ifr & T04_ifr	8.95	1	0.0028	-0.37
Ed: H05M_ifr & T14_e&c	5.24	1	0.0221	-0.17
Ee: H05M_ifr & T15M_e&c	7.17	1	0.0074	-0.28
GW: H07M_inf & T07_i&i	6.20	2	0.0451	-0.11
Ga: H07M_inf & T11_i&i	7.97	3	0.0466	0.00
IW: H09M_ifr & T07_i&i	6.99	2	0.0303	-0.13
IX: H09M_ifr & T08_inf	4.87	1	0.0273	-0.16
IY: H09M_ifr & T09M_ifr	5.71	1	0.0169	-0.28
JQ: H10M_inf & T01M_ifr	5.23	1	0.0222	0.39
Jb: H10M_inf & T12M_ifr	4.69	1	0.0303	0.27
Le: H12M_i&i & T15M_e&c	5.53	1	0.0187	-0.18
Md: H13_i&i & T14_e&c	8.14	3	0.0433	-0.18
Ne: H14_i&i & T15M_e&c	5.28	1	0.0216	-0.20
OU: H15_i&i & T05M_inf	5.06	1	0.0245	-0.29
PU: H16_e&c & T05M_inf	4.00	1	0.0455	-0.18
SX: T03_ifr & T08_inf	8.60	2	0.0135	0.36
Ua: T05M_inf & T11_i&i	8.07	3	0.0446	-0.10
Ue: T05M_inf & T15M_e&c	10.78	1	0.0010	-0.17
Wf: T07_i&i & T16M_e&c	6.41	2	0.0405	0.03
YZ: T09M_ifr & T10_inf	3.86	1	0.0495	0.21
Ya: T09M_ifr & T11_i&i	7.97	3	0.0467	0.13

A.5 Tests of item fit

Table A.6
Tests of item fit comparing observed and expected correlations between single items and rests-cores without the items. Correlations are measured by Goodman & Kruskal's gamma.

Item	observed	expected	sd	p
A - H01_inf	0.519	0.435	0.093	0.36495
B - H02_e&c	0.659	0.559	0.072	0.16184
C - H03_ifr	0.556	0.453	0.054	0.06004
D - H04_i&i	0.283	0.352	0.065	0.28811
E - H05_ifr	0.276	0.399	0.058	0.03544
F - H06_inf	0.494	0.450	0.046	0.33613
G - H07_inf	0.324	0.375	0.052	0.32730
H - H08_e&c	0.429	0.405	0.054	0.65533
I - H09_ifr	0.255	0.383	0.053	0.01538
J - H10_inf	0.518	0.414	0.067	0.12164
K - H11_ifr	0.542	0.527	0.075	0.84137
L - H12_i&i	0.522	0.500	0.056	0.69107
M - H13_i&i	0.355	0.405	0.038	0.18818
N - H14_i&i	0.462	0.446	0.050	0.75361
O - H15_i&i	0.663	0.564	0.053	0.06206
P - H16_e&c	0.465	0.442	0.054	0.67746
Q - T01_ifr	0.682	0.593	0.102	0.38166
R - T02_inf	0.422	0.426	0.045	0.91483
S - T03_ifr	0.607	0.575	0.036	0.37342
T - T04_ifr	0.557	0.507	0.050	0.31825
U - T05_inf	0.234	0.386	0.054	0.00458
V - T06_ifr	0.783	0.703	0.050	0.10531
W - T07_i&i	0.385	0.436	0.040	0.20050
X - T08_inf	0.603	0.581	0.049	0.63972
Y - T09_ifr	0.652	0.547	0.048	0.02916
Z - T10_inf	0.707	0.627	0.051	0.11319
a - T11_i&i	0.590	0.590	0.033	0.98817
b - T12_ifr	0.473	0.417	0.070	0.42693
c - T13_ifr	0.491	0.429	0.052	0.23266
d - T14_e&c	0.285	0.380	0.053	0.07122
e - T15_e&c	0.266	0.396	0.057	0.02315
f - T16_e&c	0.590	0.492	0.068	0.14798

	gamma			
Component	observed	expected	sd	p
BC	0.558	0.448	0.049	0.0242
HKLMNOP	0.554	0.545	0.027	0.7304
QSTVXYZacf	0.609	0.589	0.024	0.4176

Table A.6 measures correlations by Goodman and Kruskal's gamma coefficient because items and rest-scores are ordinal discrete variables.

We could have included the three other fit statistics that are routinely calculated for tests of item fit, but the item–restscore correlation is the best test if we want to test whether the model has problems with item discrimination. Significant test results do turn up, but adjustment for multiple testing by the Benjamini & Hochberg (1995) procedure dismiss all the results.

A.6 Tests of unidimensionality

PIRLS reports subscores for the two texts together with the total score over all items and it is not clear whether the PIRLs studies regard measurement as unidimensional or bi-dimensional. Since this is essential for criterion-referenced classification, we include results from the analyses that addressed this issue.

To test the hypothesis that Booklet 16 is unidimensional we compare the observed and the expected joint distribution of the two subscores under the unidimensional model. Figures A.1 shows the observed distribution. MRH is the sub score over items from Mary's Red Hen, TGT is the sub score over items from The Green Turtle, and R = MRH+TRT.

We refer to the set of subscores with the same total score as an orbit. In Figure A.1, the observed counts on the orbit with MRH+TRT = 16 are printed in red.

									MRH									
TGT	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
0				1															
1			1																
2			1		1														
3					3	1													
4		1		2		2	1		1	1		1							
5		1			1		1	1		2	1	2							
6				2	2	3	1	1	1		3	1	1						
7				2	1		2	1	3	4	1		1	1					
8					2	2	1	1	2	1	2	4	3	2	3		1		
9					1		5	1	4	1	3	4	2						
10						1	2	3	2	4	1	2	5	1	3				
11					1			1	3	2	3	4	3	2	3				
12					1				3		6	5	4	6	3				
13							1	1		1	5	2		4	3	2	1	1	
14								2	7	1	3	3	6	7	3	5	2		
15						1			2	3	5	2	6	5	4	3	1		
16								1		1	1	4	5	12	4	4			
17											2	5	6	9	7	11	5	5	1
18									1		2	1	5	6	9	4	2	1	
19												4	3	2	6	6	3	2	4
20													1	2	5	4	3	4	
21												1		1	4		1		

Figure A.1.
Observed counts. MRH = Mary's Red Hen. TGT = The Green Turtle.

To calculate the expected counts of (MRH, TRT) we estimate the conditional orbit probabilities of MRH,TRT given R = MRH+TRT and multiply these probabilities with the number of student on the orbit. Figure A.2 shows the orbit distributions and Figure A.3 show the expected counts.

									MRH									
TGT	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0				0.113	0.050	0.022														
1			0.411	0.268	0.158	0.086	0.043	0.020												
2		0.384	0.418	0.361	0.269	0.179	0.109	0.061	0.031	0.014										
3	0.092	0.227	0.327	0.358	0.327	0.262	0.188	0.122	0.071	0.037	0.017									
4	0.036	0.119	0.217	0.291	0.320	0.302	0.252	0.188	0.125	0.075	0.039	0.018								
5	0.014	0.057	0.128	0.206	0.267	0.294	0.282	0.240	0.181	0.121	0.071	0.036	0.016							
6		0.026	0.069	0.132	0.199	0.252	0.276	0.266	0.226	0.169	0.111	0.063	0.030	0.012						
7		0.011	0.036	0.079	0.137	0.198	0.245	0.265	0.252	0.211	0.153	0.096	0.051	0.023						
8			0.018	0.045	0.089	0.145	0.202	0.244	0.259	0.240	0.194	0.134	0.079	0.039	0.016					
9				0.024	0.054	0.100	0.156	0.211	0.248	0.255	0.227	0.174	0.113	0.062	0.028	0.010				
10				0.013	0.032	0.065	0.114	0.171	0.223	0.253	0.250	0.212	0.152	0.091	0.045	0.018				
11					0.018	0.041	0.078	0.130	0.188	0.236	0.258	0.241	0.192	0.127	0.070	0.031	0.011			
12						0.024	0.051	0.093	0.149	0.207	0.249	0.258	0.227	0.166	0.101	0.050	0.020			
13						0.013	0.031	0.063	0.111	0.170	0.226	0.259	0.251	0.204	0.137	0.076	0.034	0.012		
14							0.018	0.040	0.077	0.131	0.193	0.245	0.262	0.235	0.175	0.108	0.054	0.021		
15								0.024	0.051	0.096	0.156	0.219	0.260	0.258	0.214	0.147	0.083	0.037	0.012	
16								0.014	0.032	0.067	0.120	0.186	0.245	0.271	0.251	0.195	0.125	0.064	0.024	
17									0.019	0.044	0.087	0.150	0.220	0.272	0.285	0.252	0.188	0.114	0.051	0.013
18									0.010	0.026	0.058	0.110	0.181	0.252	0.302	0.310	0.272	0.199	0.111	0.036
19										0.013	0.032	0.069	0.127	0.203	0.281	0.341	0.361	0.327	0.236	
20											0.013	0.032	0.067	0.125	0.204	0.298	0.393	0.461		
21													0.020	0.044	0.086	0.156	0.267			

Figure A.2.
Conditional orbit distributions. MRH = Mary's Red Hen. TGT = The Green Turtle.

									MRH									
TGT	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0				0.2	0.1	0.0	0.0	0.0												
1			0.8	0.3	0.2	0.2	0.2	0.0	0.0	0.0										
2		0.8	0.4	0.4	0.5	0.9	0.1	0.3	0.2	0.1	0.0	0.0								
3		0.2	0.2	0.3	0.7	1.6	0.3	0.9	0.6	0.4	0.2	0.1	0.0	0.0	0.0					
4		0.0	0.1	0.4	1.5	0.3	1.5	1.3	0.9	0.6	0.5	0.2	0.2	0.1	0.0					
5		0.0	0.1	0.6	0.2	1.3	1.5	1.4	1.2	1.3	0.6	0.9	0.5	0.2	0.0	0.0				
6		0.0	0.1	0.1	0.7	1.0	1.3	1.4	1.9	1.1	2.2	1.7	0.6	0.2	0.2	0.1	0.0			
7		0.0	0.2	0.4	0.7	1.0	1.7	1.3	3.3	3.2	1.5	0.7	0.8	0.4	0.1	0.1	0.0			
8		0.0	0.1	0.2	0.4	1.0	1.0	3.2	3.9	2.4	1.4	2.2	1.3	0.4	0.4	0.1	0.0			
9			0.0	0.1	0.4	0.5	2.0	3.2	2.5	1.8	3.6	2.8	1.2	1.6	0.5	0.2	0.1	0.0		
10			0.0	0.1	0.2	0.9	1.7	1.7	1.6	4.1	4.0	2.3	4.0	1.6	0.8	0.3	0.1	0.0		
11			0.0	0.0	0.2	0.6	0.8	0.9	3.0	3.8	2.8	6.3	3.4	2.3	1.3	0.5	0.2	0.1	0.0	
12				0.0	0.1	0.2	0.4	1.5	2.4	2.3	6.5	4.6	4.1	3.0	1.7	1.1	0.4	0.2	0.0	
13				0.0	0.0	0.1	0.5	1.0	1.2	4.4	4.1	4.7	4.5	3.5	3.0	1.7	1.0	0.3	0.1	
14					0.0	0.1	0.3	0.4	2.0	2.4	3.5	4.4	4.5	5.2	3.9	3.1	1.5	0.4	0.1	0.0
15					0.0	0.1	0.1	0.6	0.9	1.7	2.8	3.7	5.7	5.7	6.2	4.1	1.7	0.9	0.2	0.0
16						0.0	0.1	0.2	0.6	1.2	2.0	4.1	5.4	7.9	7.0	4.1	3.0	1.1	0.5	0.1
17						0.0	0.0	0.1	0.3	0.7	1.9	3.3	6.4	7.6	6.0	6.1	3.2	2.2	0.7	0.1
18							0.0	0.1	0.2	0.6	1.3	3.2	5.1	5.3	7.2	5.3	5.2	2.6	0.6	0.3
19								0.0	0.1	0.3	0.9	1.9	2.7	4.9	4.8	6.5	4.7	1.6	2.1	
20									0.0	0.1	0.4	0.7	1.6	2.1	3.9	3.9	2.0	4.1		
21										0.0	0.1	0.2	0.3	0.8	1.1	0.8	2.4			

Figure A.3.
Expected counts of Mary's Red Hen (MGH) and The Green Turtle (TGT). Expected counts less than 0.01 are not printed.

The two-way table summarizing the joint distribution of the two subscores is a large sparse table. Therefore we estimate p-values by parametric bootstrapping.

```
Expected Gamma =   0.588  s.e. =  0.0242
Observed Gamma =   0.557    p  =  0.0350  (One-sided)
```

The one-sided p-value of 0.035 provides weak evidence against unidimensionality.

Assessment at person level, whether the observed MRH and TRT departs from what the unidimensional model expects. Table A.7 shows the conditional orbit distribution with MHR+TRT=16 including the cumulative probabilities that are needed to assess whether one of the subscores are smaller or larger than expected.

Table A.7.
Conditional orbit distribution given MRH+TGT = 16

			one-sided p-values	
MRH	TGT	prob	MRH<TGT	MRH>TRT
0	16	0.0000	0.0000	1.0000
1	15	0.0000	0.0000	1.0000
2	14	0.0001	0.0001	1.0000
3	13	0.0013	0.0014	0.9999
4	12	0.0093	0.0107	0.9986
5	11	0.0406	0.0513	0.9893
6	10	0.1139	0.1652	0.9487
7	9	0.2106	0.3758	0.8348
8	8	0.2589	0.6347	0.6242
9	7	0.2105	0.8452	0.3653
10	6	0.1108	0.9560	0.1548
11	5	0.0363	0.9922	0.0440
12	4	0.0070	0.9992	0.0078
13	3	0.0007	1.0000	0.0008
14	2	0.0000	1.0000	0.0000
15	1	0.0000	1.0000	0.0000
16	0	0.0000	1.0000	0.0000

Figure A.1 includes three student with different departures from what the model expects. One student where MRH is significantly lower than expected (MRH = 4, TRT =12 and p = 0.0107) and two student where MHR is too high (MRH = 11, TRT =15 and p = 0.044).

Since the distribution is discrete with a range consisting of 17 different outcomes, it is not possible to construct a test of significance with $\alpha = 0.05$. In this case, it is close because the two "one-sided p-values" less than 0.05 adds up to a .test with $\alpha = 0.0107 + 0.0457$. To assess whether the count of cases with $p < 0.05$ provide significant evidence against unidimensionality, we have to take that into account when we calculate the expected number of Type I errors. Figure A.4 shows the distribution of the significant cases.

	MRH																		
TGT	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
0																				
1																				
2																				
3																				
4													1							
5			1										2							
6													1	1						
7														1	1					
8															2	3		1		
9						1														
10																3				
11						1														
12						1														
13								1										1	1	
14									2									2		
15							1			2										
16									1											
17																				1
18										1		2								
19													4							
20														1						
21													1		1	4				

Figure A.4.
Observed counts of significant cases.

The results do not disagree with the hypotheses of unidimensionality. The expected number of student with too low and too high MRH scores are 19.9 and 19.7. The observed numbers of significant cases are respectively 25 and 21. None of these are statistically significant. We therefore accept the unidimensional log-linear Rasch model.

Appendix B. Criterion-referenced stages of development

This appendix describes the stages of development of reading ability in terms of the scale-anchored responses to items at the midpoint of stages.

B.1 Items with open responses

The first section presents the items with open responses of the four item domains in order defined by the manifest difficulty defined by table 2. The responses to the two easy items (T06 and H15) marked with a '*' were not utilized during the definition of the item-referenced benchmarks.

Retrieve and use information in the text

T10	2 categories	Why does a sea turtle's body fat become green?
H06	3 categories	Macy wants the red hen to go into the cage. What are two things Macy does that **do not** work?
T02	3 categories	"One of the baby sea turtles begins to stir and hatch from her egg." Write the first two things the hatchling does next.
T08	2 categories	When does a sea turtle hold its breath for up to 5 hours?

Draw straightforward inferences depending on what you have read

T06*	2 categories	According to the article, what is one way people have made the sea more dangerous for turtles?
H03	3 categories	Why does Macy's mother feel sorry for the hen at the bottom of the pecking order?
T04	2 categories	Why is the hatchling's journey to the water a "race for survival"? Use the text to explain your answer.
T03	3 categories	When the hatchling reaches the surface of the sand, what helps her go the right way?

Evaluate and criticize the content.

T14	2 categories	A diagram from the article is shown below+. What does this diagram help you to understand?
H08	2 categories	Dad says the next hen in the pecking order would just take the red hen's place. What does he mean?
H16	2 categories	Why would "Macy Finds a Way" be good as a different title for this story?

Interpret and integrate ideas and information

H15*	2 categories	What do you think the red hen will do next time Macy puts the hens in their cage?
H14	2 categories	Why is Macy at the top of the pecking order at the end of the story? Use the information from the story to explain your answer.
T07	3 categories	The color of a hatchling's shell protects it from predators. Give a way it is protected from birds.
T11	4 categories	What information does the article provide about the sea turtle's size and food at each stage of life? Complete the table+ below. Three have been done for you.
H13	4 categories	You learn what Macy is like from the things she does. Describe what Macy is like and give two examples from the story that shows this.
H04	2 categories	Why does the red hen play tricks on Macy?

B.2 The stages of development

Table B.1 presents the ranges of θ values and the midpoints of the six stages of development defined in Table 13 of the paper.

Table B.1.
The midpoints of the stages of development.

Stage	Range	Midpoint
A	$-5.038 \leqq \theta < -1.750$	-3.39
B	$-1.750 \leqq \theta < -0.948$	-1.35
C	$-0.948 \leqq \theta < -0.124$	-0.54
D	$-0.124 \leqq \theta < 0.979$	0.43
E	$0.979 \leqq \theta < 2.032$	1.51
F	$2.032 \leqq \theta < 4.113$	2.93

Tables B.2 – 7 shows the distributions of items anchored at the midpoints of the stages and Table B.8 shows our interpretation of the stages of development through the scale anchored distributions of responses to items at the midpoints of the stages.

Table B.2.
Scale-anchored item distributions at the midpoint θ = -3.39 of Stage A.

Item		Item scores						
Domain	Item	0	1	2	3	mean	Mean/max	Difficulty
	T10	0.988	0.012			0.012	0.012	Very dif.
Retrieve	H06	0.893	0.101	0.006		0.113	0.056	Difficult
information	T02	0.856	0.138	0.005		0.149	0.074	Difficult
	T08	0.987	0.013			0.013	0.013	Very dif.
	H03	0.976	0.024			0.024	0.024	Very dif.
Inferencing	T04	0.973	0.027			0.027	0.027	Very dif.
	T03	0.985	0.014	0.001		0.015	0.008	Very dif.
Evaluate	T14	0.966	0.034			0.034	0.034	Very dif.
&	H08	0.991	0.009			0.009	0.004	Very dif.
critique	H16	0.996	0.004			0.004	0.004	Very dif.
	H14	0.989	0.011			0.011	0.011	Very dif.
Interpretation	T07	0.971	0.029	0.001		0.030	0.015	Very dif.
&	T11	0.996	0.004	0.000	0.000	0.004	0.001	Very dif.
integrating	H13	0.936	0.063	0.001	0.000	0.064	0.021	Very dif.
	H04	0.995	0.005			0.005	0.005	Very dif.

Table B.3.
Scale-anchored item distributions at the midpoint θ = -1.35 l of Stage B.

Item		Item scores						
Domain	Item	0	1	2	3	mean	Mean/max	Difficulty
	T10	0.780	0.220			0.220	0.220	Difficult
Retrieve	H06	0.441	0.382	0.176		0.735	0.368	
information	T02	0.384	0.477	0.139		0.755	0.377	
	T08	0.835	0.165			0.165	0.165	Difficult
	H03	0.727	0.273			0.273	0.273	
Inferencing	T04	0.793	0.207			0.207	0.203	Difficult
	T03	0.833	0.135	0.032		0.200	0.100	Difficult
Evaluate	T14	0.785	0.215			0.215	0.215	Difficult
&	H08	0.932	0.068			0.068	0.068	Very dif.
critique	H16	0.959	0.041			0.041	0.041	Very dif.
	H14	0.863	0.137			0.137	0.137	Difficult
Interpretation	T07	0.794	0.180	0.026		0.231	0.116	Difficult
&	T11	0.959	0.033	0.008	0.001	0.051	0.017	Very dif.
integrating	H13	0.659	0.310	0.029	0.002	0.373	0.124	
	H04	0.961	0.039			0.039	0.039	Very dif.

Table B.4.
Scale-anchored item distributions at the midpoint θ = -0.54 of Stage C.

Item		Item scores						
Domain	Item	0	1	2	3	mean	Mean/max	Difficulty
	T10	0.446	0.554			0.554	0.554	
Retrieve	H06	0.201	0.392	0.407		1.205	0.603	
information	T02	0.178	0.497	0.325		1.147	0.573	
	T08	0.558	0.442			0.442	0.442	
	H03	0.487	0.513			0.513	0.513	
Inferencing	T04	0.570	0.430			0.430	0.430	
	T03	0.550	0.298	0.152		0.603	0.301	
Evaluate	T14	0.619	0.381			0.381	0.381	
&	H08	0.853	0.147			0.147	0.147	Difficult
critique	H16	0.890	0.110			0.111	0.111	Difficult
	H14	0.675	0.325			0.325	0.325	
Interpretation	T07	0.598	0.309	0.097		0.499	0.250	
&	T11	0.850	0.080	0.053	0.017	0.237	0.079	Difficult
integrating	H13	0.474	0.424	0.090	0.017	0.640	0.213	
	H04	0.916	0.084			0.084	0.084	Very dif.

Table B.5.
Scale-anchored item distributions at the midpoint θ = 0.43 of Stage D.

Item		Item scores						
Domain	Item	0	1	2	3	mean	Mean/max	Difficulty
	T10	0.127	0.873			0.873	0.873	Easy
Retrieve	H06	0.050	0.254	0.693		1.647	0.823	
information	T02	0.047	0.350	0.603		1.555	0.778	
	T08	0.205	0.795			0.795	0.795	Easy
	H03	0.235	0.765			0.765	0.765	Easy
Inferencing	T04	0.260	0.740			0.740	0.740	
	T03	0.152	0.299	0.549		1.397	0.698	
Evaluate	T14	0.381	0.619			0.619	0.619	
&	H08	0.664	0.336			0.336	0.336	
critique	H16	0.701	0.299			0.299	0.299	
	H14	0.390	0.610			0.610	0.610	
Interpretation	T07	0.288	0.387	0.325		1.037	0.519	
&	T11	0.378	0.125	0.267	0.230	1.350	0.450	
integrating	H13	0.239	0.430	0.246	0.085	1.177	0.392	
	H04	0.805	0.195			0.195	0.195	Difficult

Table B.6.
Scale-anchored item distributions at the midpoint θ = 1.51 of Stage E.

Item		Item scores						
Domain	Item	0	1	2	3	mean	Mean/max	Difficulty
	T10	0.036	0.964			0.964	0.964	Very easy
Retrieve	H06	0.007	0.110	0.883		1.876	0.938	Easy
information	T02	0.008	0.163	0.829		1.822	0.911	Easy
	T08	0.062	0.938			0.938	0.938	Very easy
	H03	0.088	0.912			0.912	0.912	Very easy
Inferencing	T04	0.085	0.915			0.915	0.915	Very easy
	T03	0.013	0.119	0.863		1.845	0.922	Easy
Evaluate	T14	0.173	0.827			0.827	0.827	Easy
&	H08	0.364	0.636			0.636	0.636	
critique	H16	0.390	0.610			0.610	0.610	
	H14	0.164	0.836			0.836		Easy
Interpretation	T07	0.068	0.269	0.699		1.596	0.798	
&	T11	0.031	0.036	0.272	0.660	2.561	0.854	
integrating	H13	0.048	0.214	0.368	0.369	2.059	0.686	
	H04	0.805	0.195			0.195	0.195	Difficult

Table B.7.
Scale-anchored item distributions at the midpoint θ = 2.93 of Stage F.

Item		Item scores						
Domain	Item	0	1	2	3	mean	Mean/max	Difficulty
	T10	0.009	0.991			0.991	0.991	Very easy
Retrieve	H06	0.000	0.029	0.970		1.970	0.985	Very easy
information	T02	0.001	0.045	0.954		1.954	0.977	Very easy
	T08	0.014	0.986			0.986	0.986	Very easy
	H03	0.022	0.978			0.978	0.978	Very easy
Inferencing	T04	0.085	0.915			0.915	0.915	Very easy
	T03	0.018	0.119	0.863		1.845	0.922	Very easy
Evaluate	T14	0.048	0.952			0.952	0.952	Very easy
&	H08	0.106	0.894			0.894	0.894	Easy
critique	H16	0.116	0.884			0.884	0.884	Easy
	H14	0.043	0.957			0.957	0.957	Very easy
Interpretation	T07	0.005	0.089	0.906		1.901	0.950	Very easy
&	T11	0.001	0.003	0.093	0.904	2.900	0.967	Very easy
integrating	H13	0.002	0.026	0.190	0.782	2.753	0.918	Easy
	H04	0.253	0.747			0.747	0.747	

Table B.8 shows how we recapitulate the information in Tables B.2-B.7 in terms of interpretations at the midpoints of the stages.

Table B.8.
Interpretation at midpoints of evidence of development of proficiencies at midpoints.

Stage	Retrieving information	Inferencing	Evaluating & Critiquing	Interpreting & integrating
A	Absent	Absent	Absent	Absent
B	Sporadic	Sporadic	Absent	Sporadic
C	Emerging	Emerging	Sporadic	Sporadic. Perhaps emerging
D	Developing	Developing. Close to sustained	Emerging	Emerging
E	Sustained	Consolidated	Developing	Developing
F	Consolidated	Consolidated	Sustained. Almost consolidated	Consolidated

Table B.9.
Scale-anchored item distributions at the benchmark θ = -0.948 of Stage C.

Item		Item scores						
Domain	Item	0	1	2	3	mean	Mean/max	Interpretation
	T10	0.633	0.367			0.367	0.367	Emerging
Retrieve	H06	0.314	0.406	0.280		0.967	0.483	Emerging
information	T02	0.273	0.507	0.220		0.947	0.474	Emerging
	T08	0.721	0.279			0.279	0.279	Emerging
	H03	0.612	0.388			0.388	0.388	Emerging
Inferencing	T04	0.697	0.303			0.303	0.303	Emerging
	T03	0.718	0.210	0.072		0.354	0.177	Emerging
Evaluate	T14	0.710	0.290			0.290	0.290	Emerging
&	H08	0.900	0.100			0.100	0.100	Absent
critique	H16	0.932	0.068			0.068	0.068	Absent
	H14	0.783	0.217			0.217	0.217	Sporadic
Interpretation	T07	0.709	0.241	0.051		0.342	0.171	Emerging
&	T11	0.924	0.052	0.020	0.004	0.105	0.035	Absent
integrating	H13	0.569	0.374	0.052	0.005	0.493	0.164	Emerging
	H04	0.942	0.058			0.058	0.058	Absent

Finally, Table B.10 presents the classification probabilities anchored at the values of θ defined by the WLE estimates of person parameters.

Table B.10.
Estimates of classification probabilities anchored at WLE estimates of θ. The stages are defined by score intervals. The probabilities of correct classification are written with bold numbers.

			A	B	C	D	E	F
CR-stage	Score	θ	0-7	8-13	14-21	22-31	32-36	37-40
	1	-5.038	**1**					
	2	-3,874	**1**					
	3	-3.298	**.998**	.002				
A	4	-2.900	**.987**	.013				
	5	-2.591	**.946**	.054				
	6	-2.336	**.856**	.143	.001			
	7	-2.117	**.718**	.280	.002			
	8	-1.750	.388	**.591**	.021			
	9	-1.592	.250	**.697**	.053			
B	10	-1.446	.148	**.740**	.112			
	11	-1.311	.081	**.718**	.200			
	12	-1.183	.041	**.642**	.316	.001		
	13	-1.063	.019	**.532**	.446	.003		
	14	-0.948	.009	.409	**.574**	.008		
	15	-0.837	.003	.292	**.686**	.019		
	16	-0.730	.001	.194	**.765**	.040		
C	17	-0.626		.120	**.803**	.077		
	18	-0.524		.068	**.757**	.135		
	19	-0.423		.036	**.748**	.216		
	20	-0.323		.018	**.664**	.318		
	21	-0.223		.008	**.554**	.438		
	22	-0.124		.003	.434	**.561**	.001	
	23	-0.025		.001	318	**.678**	.003	
	24	0.075			.216	**.777**	.007	
	25	0.174			.136	**.849**	.015	
D	26	0.276			.078	**.891**	.031	
	27	0.378			.041	**.900**	.058	
	28	0.485			.019	**.876**	.104	.001
	29	0.596			.008	**.818**	.173	.001
	30	0.713			.003	**.726**	.267	.004
	31	0.840			.001	**.603**	.388	.009
	32	0.979				.457	**.522**	.021
	33	1.133				.307	**.646**	.047
E	34	1.308				.174	**.727**	.099
	35	1.509				.078	**.728**	.193
	36	1.744				.026	**.630**	.344
	37	2.032				.005	.442	**.553**
F	38	2.407				.005	.220	**.779**
	39	2.962					.054	**.946**