MetriKsAmérique

Discussion & Analysis

Prepared for:

The Learning Partnership

Spatial Reasoning Assessment

January 2021

A complete Rasch multi-faceted analysis was performed for the Learning Partnership drafted Spatial Reasoning Assessment. While the psychometric report presents the complete analysis for the examination, this discussion proceeds step-by-step to understand the way in which the analysis proceeded, and the findings therein.

STEP ONE: General Instrument Evaluation

The originally designed instrument consists of 23 items, spanning 5 "cases" and multiple facets (e.g., Claims, Evidence, GIS Principles). Table 1 presents general performance statistics for the instrument as a whole prior to any adjustment. These statistics are considered as the "baseline" performance relative to how well the instrument is able to measure the capacity of students. Future adjustments to the instrument will be verified as appropriate based on how those adjustments positively or negatively impact the holistic item performance. In general, based on the Rasch Strata statistic (4.39) and Rasch Reliability statistic (0.90) it appears that the instrument is performing well. Rasch Strata represents the number of statistically distinct ability groups definable by the instrument. For clarity, Strata levels greater than 2.0 are required, with the rule of thumb that more levels are always better than fewer. Measured Strata above 4.0 is considered excellent. Strata essentially expands the Rasch Reliability statistic, which suggests measure consistency. Shifts in either Strata or Reliability will be used to determine whether future instrument changes were positive.

Table 1: General Instrument Performance Statistics (Baseline)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Num / Student
27.6	46.2	.6	.36	-3.76	.32	1.01	-.1	.94	.4		.60		Mean (Count: 316)
10.9	3.8	.2	.21	1.15	.17	.33	1.3	.87	.6		.14		S.D. (Population)
10.9	3.8	.2	.21	1.16	.17	.33	1.3	.87	.6		.14		S.D. (Sample)

With extremes, Model, Sample: RMSE .36 Adj (True) S.D. 1.10 Separation 3.04 Strata 4.39 Reliability .90
With extremes, Model, Fixed (all same) chi-square: 2307.8 d.f.: 315 significance (probability): .00

Analysis of Item Performance (Baseline)

Item performance is evaluated to determine whether or not each is performing according to expectations (providing unique information in a manner consistent with the holistic instrument. Two observations were noted on item review (Table 2): GIS items were far too difficult for participants and the Item 5 series performed weakly in terms of *fit* and other indicators.

First, all of the items relative to the use of GIS in the student thought process, are too difficult to be correctly addressed by the students. Those five items (noted in red within the table) were answered correctly by almost no students. The total score column represents the raw sum on rating scale points, while total count represents the total number of students multiplied by the number of examiners (2). Note that virtually no students were able to effectively respond to the items. The column "Correlation" presents the "PtMeas" or Point-biserial correlation which effectively demonstrates whether the item is able to differentiate between students who have mastered the concept and those who have not. These items are failing to differentiate student performance largely because they are too difficult.

Second, the Item 5 series, connecting the current instrument to a prior instrument from a different organization, all demonstrated a relative weakness as far as their ability to discriminate consistently between students who have mastered the content versus those who have not.

All other items function well, as demonstrated by Infit MNSQ statistics between 0.60 and 1.4 logits, and by Point Biserial correlations, all of which are positive and most of which are above 0.33.
Based on this review, we recommend removing the Item 5 series from our initial analysis as they are weak in general. While GIS items were too difficult for students, they were not removed because conceptually they represent content imperative for the project.

Table 2: Item Statistics (Baseline)

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	Items	
0	634	.0	.00	(7.01	1.83)	Maximum					.00	.00	23	Q5_G	← Q5
1	634	.0	.00	5.80	1.00	1.00	.3	.48	.1	1.01	.04	.02	7	Q2a_G	
2	634	.0	.00	5.11	.71	1.00	.2	1.21	.5	.99	.02	.04	20	Q4bc_G	
4	635	.0	.00	4.41	.50	1.00	.1	.76	.0	1.00	.05	.05	10	Q2bc_G	
7	635	.0	.01	3.85	.38	.99	.1	.66	-.3	1.01	.08	.07	17	Q4a_G	
72	634	.1	.08	1.42	.12	1.08	.7	.89	-.4	1.00	.22	.20	22	Q5_S	← Q5
126	634	.2	.15	.78	.10	1.12	1.4	1.26	1.6	.90	.19	.26	21	Q5_RF	← Q5
205	635	.3	.26	.17	.08	1.11	1.5	.98	-.1	.95	.31	.33	16	Q4a_S	
228	635	.4	.29	.03	.08	1.05	.7	.94	-.5	1.06	.40	.34	14	Q3b_R	
293	635	.5	.39	-.33	.07	1.14	2.4	1.08	.8	.82	.29	.38	15	Q4a_RF	
312	634	.5	.42	-.43	.07	1.17	2.9	1.26	2.8	.90	.40	.39	4	Q1b_R	
315	634	.5	.42	-.44	.07	1.26	4.4	1.46	4.7	.58	.23	.39	6	Q2a_S	
337	634	.5	.46	-.54	.07	1.01	.2	.94	-.7	.97	.38	.40	18	Q4bc_RF	
382	634	.6	.53	-.75	.07	1.06	1.2	1.19	2.3	.75	.30	.42	5	Q2a_RF	
407	634	.6	.57	-.85	.07	.96	-.7	.90	-1.4	1.09	.46	.43	19	Q4bc_S	
459	635	.7	.66	-1.06	.06	.77	-5.1	.82	-2.9	1.23	.47	.45	8	Q2bc_RF	
616	635	1.0	.94	-1.68	.06	.79	-4.8	.84	-3.0	1.19	.50	.50	9	Q2bc_S	
649	633	1.0	1.01	-1.81	.06	1.14	2.7	1.15	2.6	.58	.55	.51	1	Q1a_I	
685	635	1.1	1.07	-1.94	.06	.66	-8.0	.65	-7.2	1.49	.64	.52	13	Q3b_E	
765	635	1.2	1.22	-2.26	.06	1.15	2.9	1.19	3.2	.70	.61	.54	11	Q3a_I	
784	634	1.2	1.26	-2.34	.06	.72	-6.2	.81	-3.5	1.20	.55	.55	3	Q1b_E	
982	635	1.6	1.62	-3.24	.07	1.03	.5	.89	-1.6	1.19	.67	.58	12	Q3b_C	
1091	635	1.7	1.79	-3.89	.08	.93	-.9	.76	-3.1	1.17	.57	.58	2	Q1b_C	
Total Score	Total Count	Obsvd Average	Fair-M Avrage	Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Estim. Discrm	Correlation PtMea	Correlation PtExp	Nu	Items	
379.2	634.4	.6	.57	.30	.25	1.01	-.1	.96	-.3		.34		Mean (Count: 23)		
319.7	.6	.5	.53	2.90	.41	.15	3.1	.24	2.7		.21		S.D. (Population)		
326.8	.6	.5	.54	2.96	.42	.15	3.2	.24	2.7		.21		S.D. (Sample)		

With extremes, Model, Populn: RMSE .48 Adj (True) S.D. 2.85 Separation 5.93 Strata 8.24 Reliability .97
With extremes, Model, Sample: RMSE .48 Adj (True) S.D. 2.92 Separation 6.06 Strata 8.42 Reliability .97

Analysis of the Rating Scale (Rubric) Performance

Rating scales offer another source of measurement error if used incorrectly. In general, the rating scale (in this case from zero to three) should naturally flow in the same direction as the instrument. Zeros represent less content mastery whilst threes represent greater content mastery. To assess rating scale performance, we evaluate the Rasch-Andrich Threshold measures (presented in the middle of Table 3). Rasch-Andrich Thresholds should move from negative to positive, and there should be a “step” or difference of at least 1.4 logits between each rating scale point. In the case of the assessed instrument, the thresholds demonstrate a unidirectional, monotonic exchange from less to more in a manner consistent with a functional scale.

Table 3: Rating Scale Performance (Baseline)

DATA			QUALITY CONTROL			RASCH-ANDRICH		EXPECTATION		MOST	RASCH-	Cat
Category	Counts	Cum.	Avge	Exp.	OUTFIT	Thresholds		Measure at		PROBABLE	THURSTONE	PEAK
Score	Used	%	Meas	Meas	MnSq	Measure	S.E.	Category	-0.5	from	Thresholds	Prob
0	7594	55%	-5.30	-5.32	1.0			(-3.88)		low	low	100%
1	3782	27%	-2.39	-2.29	.9	-2.67	.02	-1.91	-3.03	-2.67	-2.84	52%
2	2410	17%	-.81	-.91	1.0	-1.15	.03	1.32	-.79	-1.15	-.99	86%
3	40	0%	-.46	.30	1.0	3.83	.16	(4.90)	3.83	3.83	3.82	100%
								(Mean)		(Modal)	(Median)	

Analysis of Participant Behavior

A review of the students within the sample did not generate any inconsistencies, and therefore all students in the sample were retained. Examiner performance and student performance will be discussed in greater detail within the next section.

STEP Two: Refined Instrument Evaluation (After Item 5 Series Removal)

Table 4 presents the general performance statistics for the revised instrument. To determine whether our suggested changes were positive (correct) or negative (incorrect) we compare performance of the revised instrument to baseline measures obtained. Because three items were removed and functional statistics are often influenced by N (or items or students) it is natural to expect a slight diminishment in Rasch Reliability and in Rasch Strata. However, if reliability and strata remain relatively the same or increase, it is suggested that our decision is sound. In effect, a consistency suggests that even with fewer items, we retained the performance of the instrument. An increase, of course suggests that a more distinct improvement was made. Baseline Rasch Strata and Rasch Reliability statistics were reported as 4.39 and 0.90 respectively. The current Rasch Strata of 4.36 and 0.90 are functionally identical, representing a net increase in performance. As a result, the elimination of the Item 5 series is verified to be appropriate.

Table 4: General Instrument Performance Statistics (Revised)

Total Score	Total Count	Obsvd Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Correlation PtMea	Correlation PtExp	Num	Student
27.0	40.2	.7	-3.67	.33	1.01	-.1	.93	.5	.60		Mean (N: 316)	
10.6	3.3	.3	1.17	.17	.34	1.3	.91	.7	.14		S.D. (Pop)	
10.6	3.3	.3	1.17	.17	.34	1.3	.91	.7	.14		S.D. (Sample)	

Model, Sample: RMSE .37 Adj (True) S.D. 1.11 Separation 3.02 Strata 4.36 Reliability .90
Model, Fixed (all same) chi-square: 2279.5 d.f.: 315 significance (probability): .00

Once the functionality of the instrument has been assured, we may proceed to review the performance of participants and examiners.

Analysis of the Participants (Final)

No students in the sample were noted to be problematic (appropriate *fit /point biserials*) and all were retained.

Analysis of the Examiners (Final)

Examiner (judge) performance will be evaluated in a manner similar to that items and participants. Table 5 presents examiner performance statistics. Infit and Outfit represent measures of examiner consistency. MNSQ values between 0.6 and 1.4 are indicative of internally consistent examiners, who follow a single pattern for reviewing all students. Internal consistency in multi-faceted analyses is a requirement while equivalent ratings across examiners is not. The Infit and Outfit MNSQ values all fall in the appropriate range. As a result, we may be confident that the examiners are providing clear and consistent ratings.

Internal consistency (a measure of quality) does not guarantee that the examiners always agreed on their ratings, which were made independently. Traditional raw score rating averages are insufficient at evaluating differences in examiner severity because they treat all items and all aspects on the instrument as identical, while Rasch considers factors including item difficulty which almost by design differ from item to item and from aspect to aspect. This schism between raw scores and measures is demonstrated well with the current assessment. The Observed Average (column three in Table 5) displays the raw score averages of ratings, by judge, across students. Notice that there is some difference between severity of ratings; examiner 76 awarded an average rating of 0.60 whilst examiner 75 awarded an average rating of 0.50. At face value, the difference appears minor. Although classical evaluations do not provide error terms to assess the significance of these differences, they appear to be slight. The Rasch Measure (column five in Table 5) presents the severity measure for each examiner which considers item difficulty and student ability in the model. Properly modeled, the differences between examiners become much more striking. Examiner 76 was more severe than Examiner 75, at a level that is statistically significant (difference = 0.50 logits, SE = 0.02). While items in this case do not entirely define examiner differences, the concentration of differences appeared to present themselves relative to the Item 4 series. Because both examiners were internally consistent, this difference is not a concern and instead is simply modeled as part of the multi-faceted equation.

Table 5: Examiner (Judge) Performance Statistics

Total Score	Total Count	Obsvd Average	Fair-M Avrage	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Judge
3826	6336	.6	.26	.25	.02	.87	-7.3	.88	-.5	76
4698	6354	.7	.40	-.25	.02	1.11	5.8	1.02	.1	75

Model, Sample: RMSE .02 Adj (True) S.D. .35
Separation 14.58 Strata 19.78 Reliability 1.00
Model, Fixed (all same) chi-square: 213.6 d.f.: 1 sig (probability): .00

With the understanding that examiners (judges) are scoring in an internally consistent manner, attention may finally be turned once again to item performance. Table 6 presents item performance statistics generated from the revised instrument.

Analysis of the Items (Final)

Table 6: Item Performance Statistics (Final)

Total Score	Total Count	Obsvd Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Correlation PtMea	Correlation PtExp	Nu	Items
1	634	.0	5.93	1.00	1.00	.3	.44	.1	.04	.03	7	Q2a_G
2	634	.0	5.24	.71	1.00	.2	1.21	.5	.02	.04	20	Q4bc_G
4	635	.0	4.54	.50	1.00	.1	.72	.0	.05	.05	10	Q2bc_G
7	635	.0	3.98	.38	.99	.1	.68	-.2	.08	.07	17	Q4a_G
205	635	.3	.29	.08	1.12	1.8	1.01	.1	.30	.33	16	Q4a_S
228	635	.4	.15	.08	1.04	.6	.93	-.5	.41	.35	14	Q3b_R
293	635	.5	-.21	.07	1.16	2.7	1.10	1.1	.29	.38	15	Q4a_RF
312	634	.5	-.31	.07	1.17	2.9	1.26	2.7	.41	.40	4	Q1b_R
315	634	.5	-.33	.07	1.28	4.6	1.47	4.8	.23	.40	6	Q2a_S
337	634	.5	-.43	.07	1.03	.6	.96	-.4	.37	.41	18	Q4bc_RF
382	634	.6	-.64	.07	1.08	1.5	1.20	2.5	.30	.43	5	Q2a_RF
407	634	.6	-.74	.07	.98	-.3	.92	-1.0	.45	.44	19	Q4bc_S
459	635	.7	-.96	.06	.78	-4.8	.83	-2.6	.47	.46	8	Q2bc_RF
616	635	1.0	-1.58	.06	.80	-4.6	.84	-2.8	.50	.51	9	Q2bc_S
649	633	1.0	-1.71	.06	1.13	2.6	1.15	2.6	.56	.52	1	Q1a_I
685	635	1.1	-1.85	.06	.66	-8.0	.65	-7.3	.65	.53	13	Q3b_E
765	635	1.2	-2.17	.06	1.15	2.8	1.18	3.1	.62	.55	11	Q3a_I
784	634	1.2	-2.25	.06	.72	-6.3	.81	-3.5	.55	.55	3	Q1b_E
982	635	1.6	-3.16	.07	1.04	.6	.89	-1.6	.66	.58	12	Q3b_C
1091	635	1.7	-3.81	.08	.93	-.9	.76	-3.1	.57	.58	2	Q1b_C
Total Score	Total Count	Obsvd Average	Model Measure	Model S.E.	Infit MnSq	Infit ZStd	Outfit MnSq	Outfit ZStd	Correlation PtMea	Correlation PtExp	Nu	Items
426.2	634.5	.7	.00	.18	1.00	-.2	.95	-.3	.38		Mean (N: 20)	
316.5	.6	.5	2.69	.25	.16	3.2	.24	2.8	.20		S.D. (Pop)	
324.7	.6	.5	2.76	.26	.16	3.3	.25	2.8	.21		S.D. (Sample)	

Model, Sample: RMSE .31 Adj (True) S.D. 2.74 Separation 8.72 Strata 11.96 Reliability .99
Model, Fixed (all same) chi-square: 4309.5 d.f.: 19 significance (probability): .00

As noted earlier, students were across the board, unable to demonstrate mastery within the GIS facet. Difficulty of the GIS facet is illustrated in the raw score (column one of Table 6) and more precisely in the linear Rasch measure (column four of Table 6). Because of their level of difficulty, the GIS facets do not appear to be particularly functional (as observed through their associated Point Biserial correlations with are near zero). Rasch fit statistics demonstrate a heightened level of consistency across these items because so few students address them correctly. As a pretest this may not be entirely surprising.

All of the remaining items and facets function well to discriminate between those students who do and do not understand the concepts (fit statistics between 0.6 and 1.4; point-biserial correlations positive, tending over 0.33). Question 2A_S appears to be the weakest of the items, with a lower point-biserial correlation and partial problems with fit statistics (meeting our expectations as an item). Its performance was not weak enough to be dropped but should be reviewed in future editions of the assessment.

One of the benefits of Rasch analyses is the separation of items and students relative to performance analyses. In the case of items, as a group they are performing admirably. They appear to define a consistent and clear construct (Rasch reliability = .99, Rasch Strata = 11.96). The strata in particular suggest that the construct of “spatial reasoning” as operationalized within the instrument proceeds in a developmental fashion from easier to more difficult to master. While there are nearly 12 statistically significant strata, the meaningfulness of those 12 strata is likely in smaller increments. Our

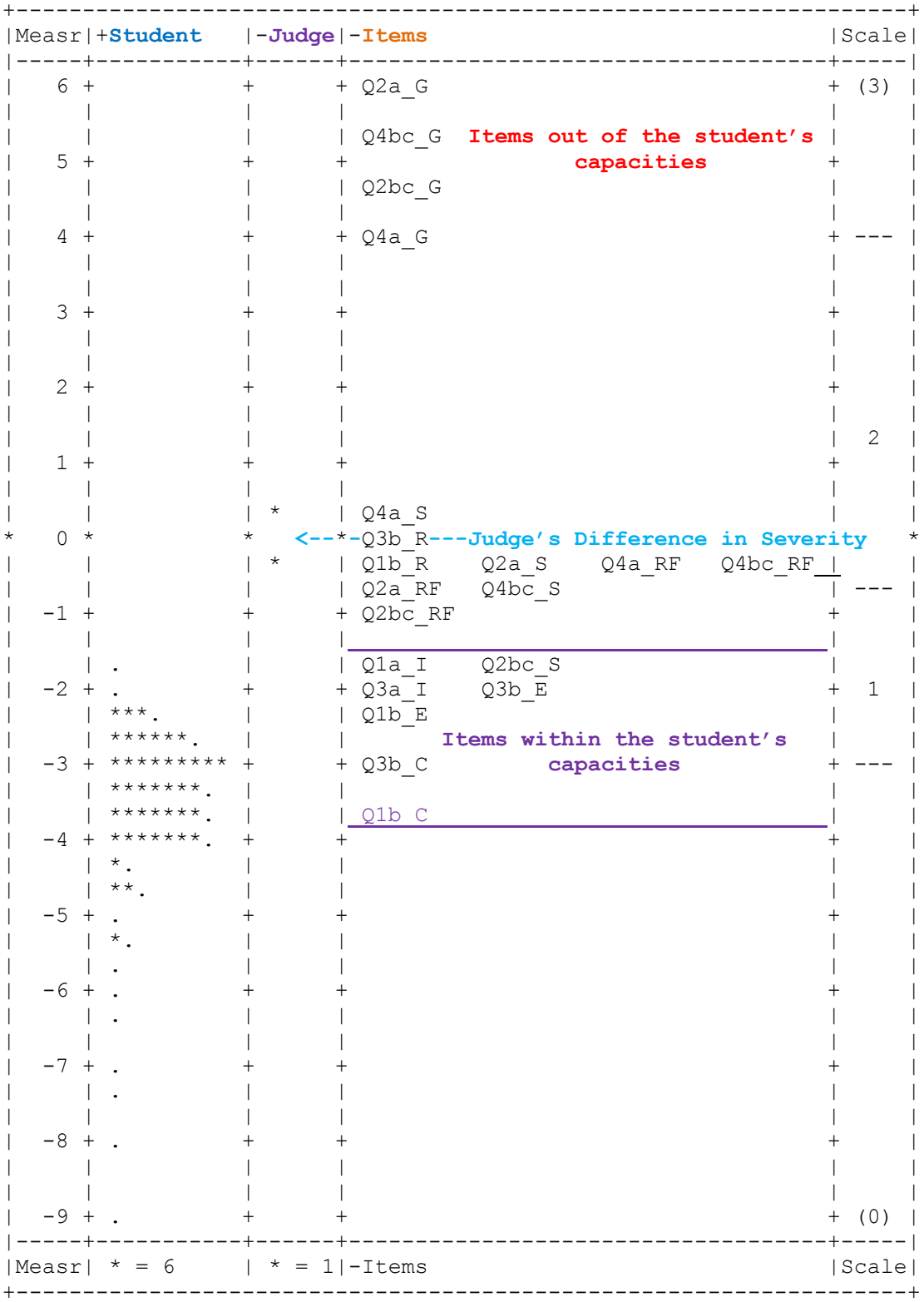
interpretation divides the items and facets into four meaningful developmental steps. Those four steps are presented in Table 6 by horizontal dividing lines in the last column. The easiest concepts for students to master relate to "Claims" (what we are calling Level One within the developmental mastery process). Students were marked at an average of approximately 1.6 raw score rating points (midway between 1 and 2) which represents a reasonable grasp of the content presented. Students also appeared to be functional in the areas of the use of "Evidence" and "Reasoning" (Level Two within the developmental mastery process). With a raw score average of about 1.0 students in the sample have begun their journey within the content of Spatial Reasoning but have not yet reached what we might consider proficiency. Level Three, inclusive of "Relevant Factors" and their general "Spatial" orientation, is much more challenging. Students are struggling to reach even basic mastery. Only a few students are able to affirmatively address these items even at the most basic level of understanding. Finally, Level Four includes items assessing GIS use/reasoning. Virtually none of the sampled students are meeting this GIS challenge. This arrangement is likely not entirely unexpected in part because of the preliminary nature of this assessment, but also because as the construct of reasoning proceeds from easier to more difficult, it also proceeds from general reasoning to more specific reasoning, using the ideas presented (e.g., GIS). Use of any particular methodological framework is almost always more difficult and complex than a broader general use of, for example, reasoning through claims. Ultimately this sort of analytic tool can assist in both understanding the development of Spatial Reasoning and provide an excellent framework that might be used to drive instruction – or the steps within instruction relative to the topic.

As discussed, the construct of Spatial Reasoning is well-elaborated (operationalized) by the instrument under investigation. The final, and quite visual representation of this construct and its relationship to the students and examiners, is useful in a more holistic understanding. Figure 1 presents the construct map, inclusive of students, examiners, and items in an intuitively useful manner for explanation.

Within Figure 1 are presented: (a) the distribution of persons in the sample (column 2) with more able students at the top and less able at the bottom; (b) the distribution of examiners (judges) used to score the assessments (column 3) with more severe examiners at the top; and (c) the distribution of items on the assessments (column 4) with the most difficult items towards the top and the least difficult items towards the bottom. These three distributional placements make interpretation quite intuitive. For example, as noted on the Figure, only a handful of items are within the majority of the sample's capacity as students. Items at or below the level of a student are considered solvable, while items above the level of a student are considered too difficult. Roughly 7 of the 20 items on the assessment worked well to help define student ability. The cluster above the initial group (roughly 9 items) are potentially within the grasp of some students and likely others, with additional instruction. Finally, the four GIS related items are well outside of students' current capacities and do not, at present, contribute to measurement.

Similarly, you will note the two examiners in the middle of the Figure. Their location spread across the lower end of the rating of two, and across a number of items as indicated before. While their positioning is relatively high compared to the students and promotes the idea that in general, examiners did not find most students were coming close to mastering the content being assessed.

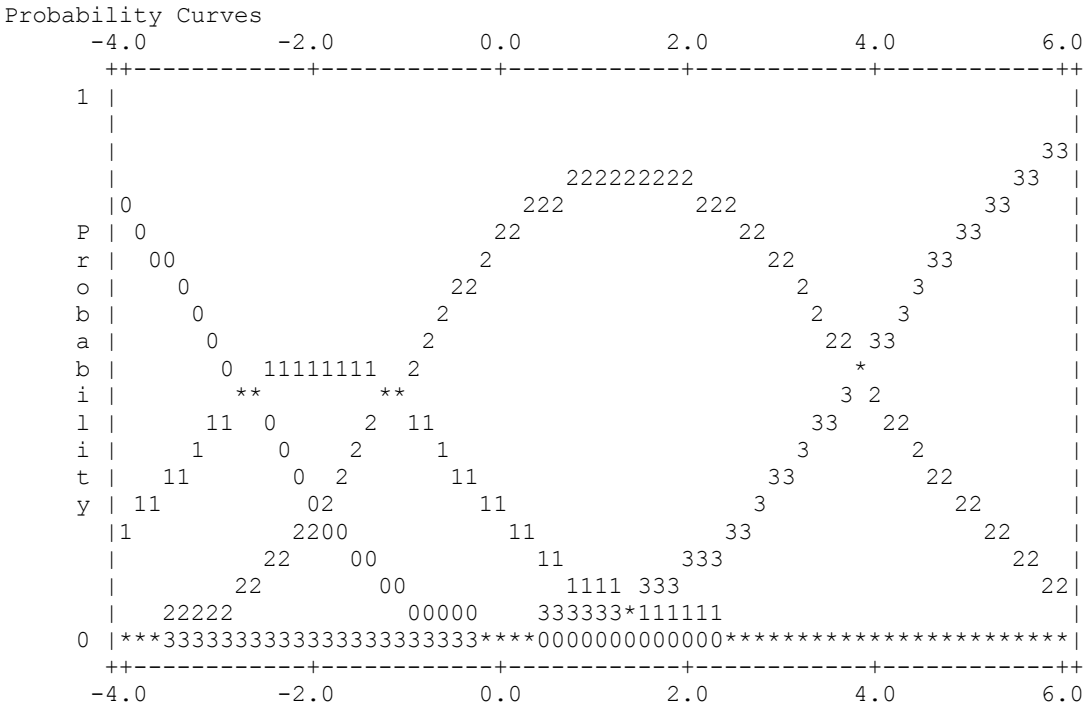
Figure 1: Spatial Reasoning Construct Map



Analysis of the Rating Scale (Rubric) (Final)

As a final piece of the assessment, it is important to review the rating scale deployed to ensure it is effective across the instrument and across the ratings. Figure 2 presents information on the performance of the rating scale on the final instrument, both graphically and numerically. The rating scale on the final evaluation, performed well again. One suggestion for improvement relates to the descriptions for a rating of "1" versus "2". It would not be a mistake on the part of the test-makers to review how 1 and 2 differ, clearly and precisely. They are functional on the current assessment, but the closeness of rating scale point 1 in particular to not achieving unique probability on its own, suggests that perhaps when a larger sample or a more diverse sample is tested, use of that point may encounter some difficulties in terms of scale overlap. This is usually addressable by ensuring that the descriptions with the rubrics are clearly distinct. At this time, however, no immediate concerns were noted.

Figure 2: Rating Scale Performance



DATA			QUALITY CONTROL			RASCH-ANDRICH	
Category Counts		Cum.	Avge	Exp.	OUTFIT	Thresholds	
Score	Used	%	Meas	Meas	MnSq	Measure	S.E.
0	6517	52%	-5.39	-5.42	1.0		
1	3622	29%	-2.33	-2.22	.9	-2.70	.03
2	2391	19%	-.78	-.89	1.0	-1.14	.03
3	40	0%	-.43	.34	1.0	3.85	.16

STEP Three: Final Comments

Holistically the instrument performed admirably. As a pretest, it is likely that students were not expected to demonstrate certain reasoning skills (e.g., GIS) as indeed they did not. The rating scale functions well to capture the examiner judgement, with the comments noted above. Overall, the instrument works together as a functional assessment, capturing the general construct of Spatial Reasoning. Only the Item 5 series is recommended for removal, but because it was generated from outside the current project, it is perhaps neither unexpected nor problematic.