

**Using the Many Facet Rasch Model to Explore Item Independence
in a Field Trial of Performance Based Assessments**

Terrence Moore, Ohio Department of Education

Lauren Monowar-Jones, Ohio Department of education

Leah Lyn Walker, Northwest Evaluation Association

The decision to treat items independently or to bundle items together as polytomous items can affect the estimate of examinee ability. This study uses a field trial of a performance based assessment tasks to examine if Rasch model fit statistics have promise for use to materially improve item modelling over alternative approaches in the determination of whether items are functioning independently or contingently on other items on the test form. The context for the study is K-12 school children taking performance based assessments where student work samples are scored by two raters.

Background. Ohio testing programs have used constructed response items that are scored using the Masters' Partial Credit Model (MPCM). That model is described by Masters (1982) as applicable for ordinal ratings that might be classified as either a rating scale or where the examinees' efforts and evidences are contingent and can be ordered. For example, a holistic judgment of, perhaps, a writing sample is scored on a scaled of 1 to 3 points based on the impressions the rater has about the sample in placing it on some ordinal scale. The writing sample could be scored or classified as insufficient, sufficient, and more than sufficient. Masters describes this type of rating with a Likert scale such that higher integer point scores are conferred for higher levels of agreeableness; strongly disagree is worth a rating of zero points, disagree is worth one point, agree is worth two points and strongly agree is worth three points.

Master's ordered effort example is the math problem...

$$\sqrt{\frac{7.5}{0.3} - 16} = ? \quad \text{Eq 1.}$$

The evaluation of this expression (Eq 1) must proceed from first doing the division $\left(\frac{7.5}{3}\right)$, then computing the difference $(25 - 16)$, and finally taking the square root of that difference $(\sqrt{9})$. Because this has three steps that must be done in order and with the correct result (the division of $7.5 / 0.3$ "scaffolds" the computation of the difference), the score for this expression is one point for correctly computing the ratio, another point for correctly taking the difference, and another point for finding the root of the difference. A student getting the division and the difference computations correct would be awarded points of partial credit even if the student failed to provide the square root or did so, but, the root is numerically incorrect. In both the rating scale example and the math problem example the ordinal nature and dependency of the scoring is apparent: each higher score is contingent on getting the lower score first. Moreover, the contingency is illustrated by the lack of independence of the higher score on the lower score.

Ohio tests under ESEA. The reauthorization of the Elementary and Secondary Education Act in 2001 required testing programs and Ohio responded with the Ohio Achievement Test (or OAT later renamed the Ohio Achievement Assessment or OAA) and the Ohio Graduation Test (OGT). Both tests have included a writing section that is scored for writing conventions and applications and are properly scored using a rating scale. Additionally, the Reading Test includes constructed response items scored using a rating scale.

For other subjects, Mathematics, Science, and Social Studies, there are also constructed response items. Some of those items fit the rating scale model and others are presumed to fit the scaffold partial credit model.

In addition to cases that fit the two models offered by Masters, Ohio also uses the MPCM for situations where the contingent progression in scoring or rating work is less obvious. Consider, for example, Table 1 for an item that asks for two reasons and an example of each.

Table 1 – Possible polytomous item scores for two reasons and an example of each reason

Situation	1 st reason	Example for 1 st reason	2 nd reason	Example for 2 nd reason	Total item score
1	0				0
2	1	0	0		1
3	1	1	0		2
4	1	0	1	0	2
5	1	1	1	0	3
6	1	1	1	1	4

As a practical matter, students tend to more frequently acquire point scores 0, 2, and 4 and less frequently get scores of 1 or 3. It is speculated that this happens because a reason and the accompanying example tend to come in pairs. Table 1 shows there to be two different ways to get a score of two but when analyzing the data, no distinction is made between getting the second point by providing an example for the first reason or providing a second reason. The frequency with which a second reason occurs and the frequency with which a second example occurs are different but the model (and the data) make no distinction as to the cause for a particular score (and thereby assume no particular ordering as seen in the prior Masters’ partial credit example). Further, there is scant contingency between providing a first reason, with or without an example, and producing a second reason; the first reason is not much of a scaffold toward determining a second reason and a first example is even less of a scaffold for providing the second reason. The approach of Table 1 seems to differ from the examples in Masters’ description because there are score points that are not contingent – especially the first of the two examples – being modelled as though they scaffold. Still, the connectedness of the points in Table 1 are apparent even if they do not perfectly “scaffold” from the first point through the fourth point: the points all deal with the same topic and tend to probe the topic more completely than would the first question alone.

Alternative models. There are other perspectives for grouping test items into polytomous items including Item Bundles (Rosenbaum, 1988), SOLO superitems (Wilson and Iventosch, 1988), the Saltus model (Wilson, 1989; Draney, and Wilson, 2007, p 119-130), the success or growth model (Verhelst, Glas, and de Vries, 1997, p 123-138), the failure or mastery model (Linacre, 1991) and Testlets (Wainer, Bradlow, and Wang, 2007; p 63). Some of these models have requirements that limit the applicability of the concepts. For example, the SOLO superitem is about a collection of items that deliberately are thought to probe a Piagetian structure of learning where the evidence of the individual items are hierarchical and reflect a progression of human development. The Saltus model is focused on specific developmental stages, classification into those stages, and “leaps” from one stage to the next. Testlets, according to Wainer, Bradlow, and Wang, were originally a group of items sharing the same stimulus (p 44); they go on to suggest that the testlet is a practical collection of items to be invoked together (p 53) under computer adaptive testing and rely on Rosenbaum’s descriptions of Item Bundles for supporting theory (p 54-55). Use of the success model or the failure model are somewhat discouraged in the WINSTEPS (Linacre, 2009b) manual and discouraged in personal communication from Linacre to Kurt Taube in favor of the MPCM even though the WINSTEPS software continues to support both the success model and the failure model.

Research question. Ohio tests use the Masters' Partial Credit Model without apparent regard to the alternative models or to the tight description of the MPCM in Masters' (1982) seminal paper for rating scales and contingent partial credit. Items such as the one shown in Table 1 have a convenience of satisfying a blueprint requirement for a four point constructed response item. Reducing the item of Table 1 from a single 4-point item to a pair of 2-point items also reduces the usefulness item, the two 2-point items seem redundant, and, in the sense that the query coheres as a potential non-redundant list, it would seem impossible to get to the full scope of the task for the item in Table 1 by asking only half the of what was asked (e.g for a single reason and a single example).

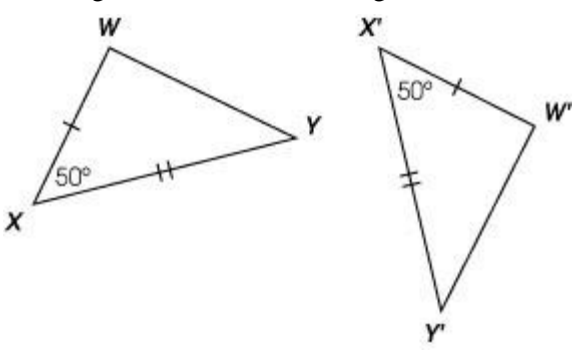
Also important in the use of the item of Table 1 is whether the item writer "got it correct". Often items like the one shown in Table 1 are accepted for use as conceived by the item writer. If the item writer thinks that this is the way to probe a bit of content related student ability, the item writer's judgment is accepted pending receipt of field test data. Alternative structures for probing the content often go unexplored. Nevertheless the purpose of those items is to estimate the ability of the examinee. Therefore the research question that this paper explores is:

Does the assignment of questions as components of a polytomous item materially alter the quality of estimates of examinee ability?

Data. The data for this analysis is an assessment task used for a course in high school geometry called "Exploring Congruent Triangles". The task was part of the Ohio Performance Assessment Pilot Project where an assessment task was preceded by a learning task that was designed to nurture the skills that would be necessary to succeed on the assessment task. All twelve of the responses were constructed responses. The full test form is shown in Appendix A and the second item on the form is shown as Figure 1 to illustrate a typical item.

Figure 1 – An item from the performance assessment task called Congruent Triangles

The diagram below shows triangles WXY and $W'X'Y'$.



(a.) Which Triangle Congruency theorem (SAS, SSS, or ASA) can be used to show that triangles WXY is congruent to triangle $W'X'Y'$? Justify your reasoning. (1 point for the theorem, scored as item 2at, and 2 points for the explanation, scored as item 2ae)

(b.) Which rigid motion or motions can be used to position one triangle onto the other to show congruence? (1 point, scored as item 2b)

While it is not possible to know exactly what the item writer was thinking, there are some unifying features to the score-able tasks in Figure 1. First, all of the scores described in Figure 1 are for the same diagram of the two triangles. If the examinee fails to properly understand the figure then the scores for each part would be depressed. Second, part (a.) for the item in Figure 1 asks the examinee to both recall the appropriate theorem and then to justify the appropriateness of the theorem much like the operational example of Table 1.

The task, Exploring Congruent Triangles, was scored by 19 raters and most of the raters were high school teachers that administered the learning task. Each rater should have provided 15 separate scores for some of the more than 600 student work samples. Nearly 83% of the student work sample was scored twice to link the raters although some of the tests were scored a single time (12%) and other samples were scored three times. Raters were assigned student work to score by an arbitrary process that does not strictly qualify as random but the process was far from completely predictable. Raters scored between 250 and 750 papers each in December of 2013. The fixed form was administered and scored using an on-line system.

Data Analysis. The data were processed to generate several files. The first file was one where the nominal numbering system used by the item writer causes there to be six items and a total possible score of 20. The item writer’s numbering suggested the structure as shown in Table 2 and will be called the Writer’s model. A second file was constructed as a congeneric analysis where each score entry by a rater was an item resulting in 15 items yielding the same 20 raw score points. Fifty additional files were generated using the structure of the first file (Table 2) but randomly assigning the fifteen items into that structure based on the equivalence of the point values: 2-point items were randomly assigned in places for 2-point items and 1-point items were randomly assigned to places for 1-point items.

Table 2 – The combination of discrete scores into items using the item writer’s numbering

Discrete score item	Points for discrete item	Item writer’s conceptualization
1	1	Combine into Item 1
2	1	
3	1	Combine into Item 2 (see figure 1)
4	2	
5	1	
6	1	Combine into Item 3
7	1	
8	2	
9	1	
10	1	Combine into Item 4
11	1	
12	2	Combine into Item 5
13	2	
14	1	
15	2	Combine into Item 6

The method of analysis was the Many Facet Rasch Model (MFRM) implemented through the software package FACETS (Linacre, 2009b). FACETS was chosen because the software accommodates differences in the severity of the raters that would be masked if the study was only of raw scores. Second, the software provides important measures relevant to the research question including both estimates of the

error of an estimate and measures of model fit. Also, FACETS is commercially available, well documented, and has lineage that links to Ohio’s present scoring model software, WINSTEPS.

Operation of the model in FACETS requires the setting of the metric. This was done by setting both the mean item difficulty and the mean rater severity to zero and allowing the mean examinee ability to be determined. Ohio’s customary practice is to set the mean item difficulty to zero because Ohio has some control over the difficulty of items on a test form.

Results.

Structure matters. As shown in Table 3, the mean ability of the examinees is sensitive to the structure of the analysis of the raw scores.

Table 3 – Mean person ability estimated in the 52 models.
(all units logits)

Model:	Writer’s	Congeneric	Fifty random models			
			Mean	SD	Max.	Min.
Mean person difficulty	0.36	0.66	0.38	0.09	0.56	0.24

The data for the three models is completely the same but the organization of the data into items is different for the 52 models in the study. The mean person ability for the 50 random models is about the same as the mean for the Writer’s model and the standard deviation for the 50 random models (0.09 logits) is modest suggesting that the mean for the Congeneric model differs mostly because the structure is different. However, at least one of the random models approached the Congeneric model mean with a maximum mean person ability of 0.56.

Structure matters not only because of the difference in the means of student scores shown Table 3 but also because there will be a need to produce alternative forms that are thought to be equivalent to a base form. How could the optimum form design be chosen empirically when new forms are expected to produce the same inferences?

Rater analysis. Rater model fit for the 52 models are summarized in Table 4.

Table 4 – Rater model fit estimates for 52 models

Model:	Writer’s	Congeneric	Fifty random models			
			Mean	SD	Max.	Min.
Lowest rater r_{pb}	0.42	0.24	0.45	0.03	0.50	0.24
Infit_{max}	1.29	1.21	1.29	0.06	1.45	1.16
Infit_{min}	0.77	0.77	0.77	0.03	0.84	0.69
Outfit_{max}	1.67	1.49	1.32	0.16	1.89	1.13
Outfit_{min}	0.74	0.64	0.77	0.04	0.88	0.64
#Outfit>1.5	1	0	0.14	0.40	2	0
#Outfit<0.5	0	0	0	0	0	0

None of the models provided strong evidence that the rater scoring was a violation of model fit.

Item analysis. Item model fit is summarized in Table 5.

Table 5 – Item model fit estimates for 52 models

Model:	Writer's	Congeneric	Fifty random models			
			Mean	SD	Max.	Min.
Lowest item r_{pb}	0.25	0.13	0.33	0.10	0.89	0.13
Infit_{max}	1.10	1.20	1.18	0.06	1.27	1.1
Infit_{min}	0.85	0.73	0.80	0.04	0.86	0.73
Outfit_{max}	1.35	1.51	1.26	0.17	1.63	1.13
Outfit_{min}	0.86	0.65	0.79	0.04	0.87	0.65
#Outfit>1.5	0	1	0.02	0.14	1	0
#Outfit<0.5	0	0	0	0	0	0
Item mean	0	0	0	0	0	0
Mean SE	0.04	0.07	0.04	0.00	0.07	0.04
RMSE	0.04	0.08	0.04	0.00	0.08	0.04
Separation	24.62	20.22	20.47	5.58	29.97	9.83
Model fit (χ^2)	2730	5275	2258	1079	5275	599
df for model fit	5	14	5			
Var. ex. Rasch	67.5 %	43.2 %	68.7 %	3.23 (%)	75.2 %	43.2 %

Item analysis is central to the study because the different models are generated by reorganizing the items. Table 5 shows that when compared to the congeneric model, the other models tend to mask the lowest discriminating item. Table 5 also shows that model fit is substantially improved when items are combined compared to the congeneric model. For example, the difference between the congeneric model fit and Writer's model fit is a χ^2 of 2,545 with 9 degrees of freedom – a highly significant difference. One random model (model R36) produced a χ^2 of 599 with 5 degrees of freedom; this model has a substantially better item fit than the writer's model. Random assignment models with a better fit than the Writer's model are not uncommon as shown in Table 6.

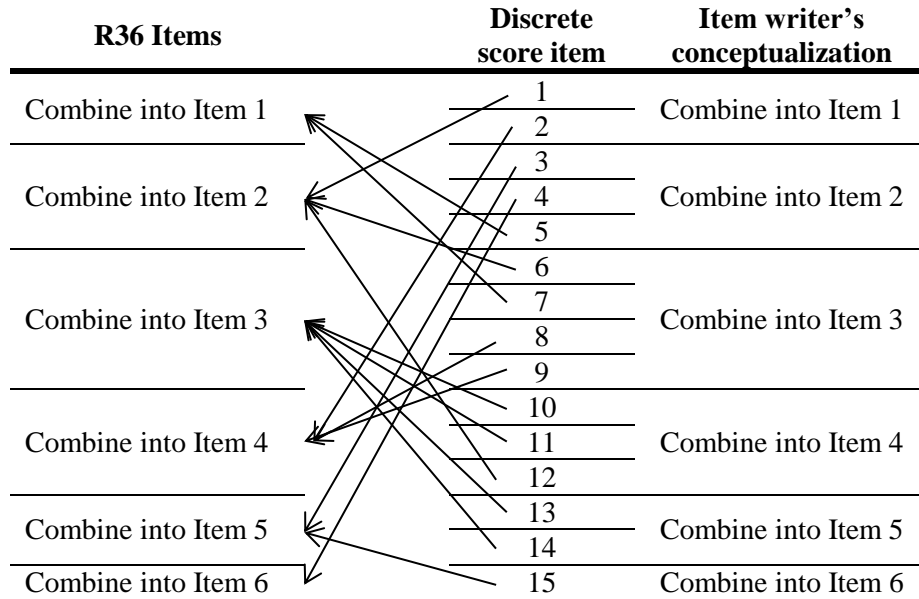
Table 6 – Random Assignment models with a substantially better fit than the Writer's model
(all models have the same number of degrees of freedom)

Model	χ^2
Writer's (reference)	2730
R42	1186
R25	1174
R11	1099
R15	1089
R44	1031
R49	1013
R20	1012
R44	927
R36	599

The best fitting of items for these alternative random models is, by far, model R36.

Comparing the R36 model to the writer's model. The mapping of items in R36 is shown in Table 7.

Table 7 – The combination of discrete scores into items using the item writer's numbering



Efforts were made to identify some pattern in R36 that would suggest a reason that R36 produces good item fit. For example, do the three theorem questions group and the three explanations group to produce R36? No, each of the six R36 items contains either a theorem or an explanation. Are the R36 items of more equal difficulty and that explains superior item fit? No, the standard deviation of item measures does not predict the χ^2 statistic. Is there some relationship of item co-variation that would be revealed through exploratory factor analysis? No, the 15 items produce five factors with the largest having an eigenvalue of almost four and all of the items loading at least 0.2 and often considerably higher on the first factor. None of the remaining factors suggest the combination shown as R36. The reason the items of R36 group to produce a superior item fit was in the examinee response data but it is doubtful that the examinees are consciously aware of that reason and we authors are unable to discern a reason..

Table 8 – Item fit statistics for R36 compared to the writer's model.

Item	Writer's model				R36 model			
	Infit	Outfit	Discrim	r_{pb}	Infit	Outfit	Discrim	r_{pb}
1	1.05	1.35	0.91	0.25	1.11	1.17	0.84	0.36
2	1.03	1.03	0.96	0.46	0.81	0.79	1.23	0.55
3	0.85	0.90	1.15	0.54	1.13	1.13	0.84	0.46
4	0.89	0.86	1.13	0.50	0.79	0.78	1.24	0.55
5	1.06	1.02	0.93	0.43	1.15	1.16	0.88	0.42
6	1.10	1.09	0.92	0.38	1.00	1.06	0.99	0.46
rms	1.00	1.05	1.01	0.44	1.01	1.03	1.02	0.47

The root-mean-square (rms) values for the item characteristics of the Writer's model and the R36 model are not very different; it is not clear how the improved χ^2 model fit is manifest as improvement in the parameters of Table 8.

Having failed to discern the reason as to why R36 produces superior fit, there are no “principles” to test in further research and no guiding principles for the design of items or forms.

How would the organization of R36 alter the inferences made about examinee ability? The objective of the study is to explore the possibility that grouping of items, and some particular grouping reduces the errors of the estimates of student ability. Summary data for the examinee facet is shown in Table 9.

Table 9 – Examinee facet estimates for 52 models

Model:	Writer’s	Congeneric	R36	Fifty random models (R36, inc)			
				Mean	SD	Max.	Min.
Person Mean	0.36	0.66	0.32	0.38	0.09	0.56	0.24
Mean SE	0.43	0.42	0.46	0.45	0.01	0.47	0.43
RMSE	0.44	0.44	0.48	0.46	0.01	0.48	0.44
Separation	2.7	2.68	2.99	2.88	0.08	3.02	2.71
Model fit (χ^2)	4400	4314	5219	4910	232	5322	4387

The data in Table 9 show little difference in the magnitude of the error estimates, in the separation indices, and in the model fit. There are difference in the person mean that appear to be related to both the structure (e.g compare the congeneric model to R36 and the Writer’s model), and to assignment of items (see the range of person means for the random models).

The correlation between examinee ability estimates for three models are shown in Table 10.

Table 10 – Correlations between examinee abilities estimates for three different models

Model	Writer’s	Congeneric	R36
Writer’s	1.27	1.000	1.000
Congeneric	0.999	1.25	1.000
R36	0.997	0.999	1.50

Notes: The diagonal is the standard deviation of examinee abilities (logits)
 Above the diagonal is the Spearman correlation
 Below the diagonal is the Pearson correlation

The unity Spearman correlation coefficients in Table 10 suggest that there is little reordering of the ability estimates no matter which model is chosen. The near unity Pearson correlation coefficients suggest the same; the departure from unity may be due to granularity (rounded 2 decimal) ability estimates or possibly due to re-centering of the examinee ability on the operating curve as well as the difference in dispersion suggested by difference in the standard deviation. The interpretation of one examinee’s score in the context of other examinee scores seems to be mostly un-affected by the model chosen.

Criticisms of the analysis. This is a study of a single data set for a single assessment and may not be generalizable to other assessments in this content/subject area or in other content areas. However, this assessment is attractive for study because (1) the linkage of all but one of the congeneric items is apparent through referring to one of the five figures, (2) there are many items involving each one of the figures, and (3) all of the items are constructed response eliminating guessing because the examinee could not have supplied the keyed response by the prompting found in multiple choice selection.

The study of the Triangle task is not exhaustive of the possible random combinations.

Conclusions and recommendations. This study suggests the following:

1. Using the scoring model as a strong guide in how to organize items as a single polytomous item or as several polytomous items or as a string of dichotomous items seems to be mostly unproductive. In this study the best item fit was found by random chance and, despite a remarkably improved item fit, the best fit neither led to principles to reproduce good item fit in future tests or to remarkably different (and better) examinee ability estimates.
2. Structure matters. Structure is a potent influence on the estimates of examinee ability. For tests to be considered equivalent, having equivalent structure is necessary. As shown in Table 9, despite having the same responses and the same score data, how the score data is organized to fit the structure produced substantially different mean examinee ability estimates. Pick a structure and stay with it.
3. Imposing structure through the test blueprint has great utility. Supposing one could determine the “proper structure” from a scoring model analysis of the items themselves then the future design of the hoped for equivalent forms of a test would be extremely difficult. Further, a test blueprint without an allowance for polytomous items would be encumbered from presenting many complex items that can and should be scored for partial credit or from presenting items that look like lists. It appears that committing to a test blueprint that accommodates some number of polytomous items is sound practice because those provisions in the structure of the test can be used for items that are complex (and worthy of multiple points) or items that may resemble lists that can be used to more fully sample a domain. If the items are all handled as independent items, then it is difficult to put contingent items on test where they are either necessary (a step in a scaffold) or a list (necessary for sufficient completeness).

References.

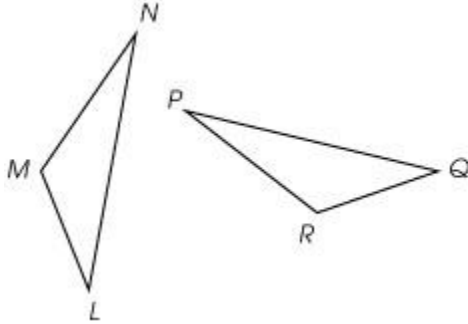
- Draney, K., & Wilson, M., (2007). Application of the Saltus Model to Stagelike Data: Some Applications and Current Developments. In *Multivariate and Mixture Distribution Rasch Models*, van Davier & Carstensen, eds. New York: Springer, 398 pp
- Linacre, J. M. (2009a) Facets Rasch measurement computer program, version 3.66.0. Chicago: Winsteps.com
- Linacre, J. M. (2009b). Winsteps® Rasch measurement computer program User's Guide. Chicago: Winsteps.com
- Linacre, J. (1991). *Structured Rating Scales*. Paper at Sixth International Objective Measurement Workshop, Chicago.
- Masters, G. (1982). *A Rasch Model for Partial Credit Scoring*. Psychometrika 47(2), 149-174.
- Rosenbaum, P. (1988). *Item Bundles*. Psychometrika 53(3) 349-359.
- Verhelst, N., Glas, C., & de Vries, H. (1997). A Steps Model to Analyze Partial Credit. In *Handbook of Modern Item Response Theory*, van der Linden & Hambleton, eds. New York: Springer, 510 pp
- Wainer, H., Bradlow, E., & Wang, X (2007). *Testlet Response Theory and its Applications*. Cambridge: Cambridge University Press. 267 pp.
- Wilson, M. (1989). *Saltus: A psychometric Model of Discontinuity in Cognitive Development*. Psychological Bulletin, 105(2), 276-289.
- Wilson, M. & Iventosch, L. (1988). *Using Partial Credit Model to Investigate Responses to Structured Subtests*. Applied Measurement in Education, 1(4), 319-334.

Appendix A. Test Form for “Congruent Triangles”

Task 1

Question 1a

1. In the diagram below, triangle MNL is congruent to triangle RPQ .



a. List the 3 corresponding pairs of angles between triangle MNL and triangle RPQ .

(Max chars: 10,000)

1 point for all three being correct

0 Count

Question 1b

b. List the 3 corresponding pairs of sides between triangle MNL and triangle RPQ .

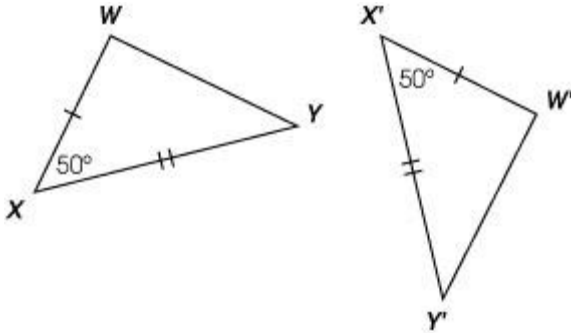
(Max chars: 10,000)

1 point for all three being correct

0 Count

Question 2a

1. The diagram below shows triangles WXY and $W'X'Y'$.



a. Which Triangle Congruency theorem (SAS, SSS, or ASA) can be used to show that triangles WXY is congruent to triangle $W'X'Y'$? Justify your reasoning.

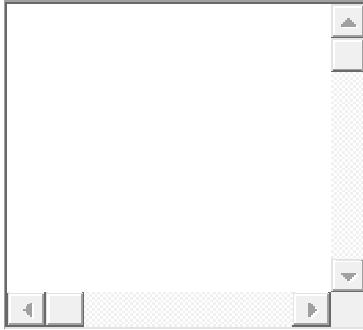
1 point for the theorem and 2 points for the explanation

(Max chars: 10,000)

Count

***Question 2b**

b. Which rigid motion or motions can be used to position one triangle onto the other to show congruence?



1 point

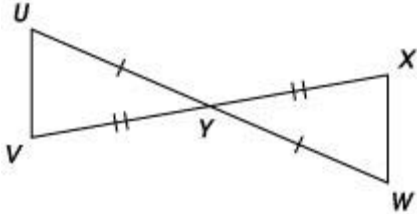
(Max chars: 10,000)

Count

***Question 3a**

3.

Two triangles are shown in the diagram below. Line segment UW intersects line segment XV at point Y . Side UY is congruent to side WY . Side VY is congruent to side XY .



a. Using the letters of the vertices in the diagram, name the two triangles that are congruent to each other.

(Max chars: 10,000)

1 point for both

Count

***Question 3b**

b. Which Triangle Congruency theorem (SAS, SSS, or ASA) can be used to show that the two triangles are congruent to each other? Justify your reasoning.

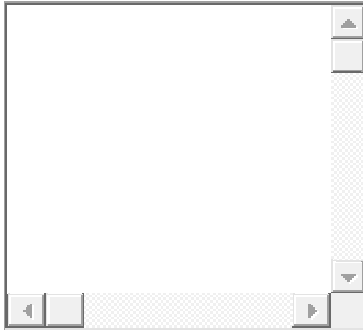
(Max chars: 10,000)

1 point for theorem and 2 points for explanation

Count

***Question 3c**

c. Which rigid motion(s) can be used to position one triangle onto the other to show congruence?



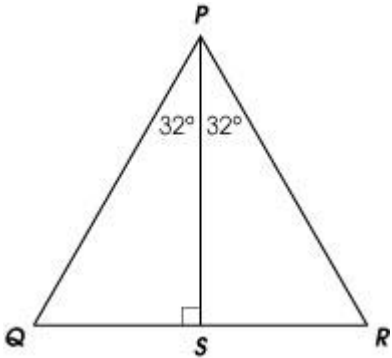
1 point

(Max chars: 10,000)

Count

Question 4a

4. Triangle PQR is shown in the diagram below.



a. Using the letters of the vertices in the diagram, name the two triangles that are congruent to each other.

An empty text input box with a scroll bar on the right and navigation buttons at the bottom.

(Max chars: 10,000)

1 point

Count

Question 4b

b. Which Triangle Congruency theorem (SAS, SSS, or ASA) can be used to show that the two triangles are congruent to each other? Justify your reasoning.

An empty text input box with a scroll bar on the right and navigation buttons at the bottom.

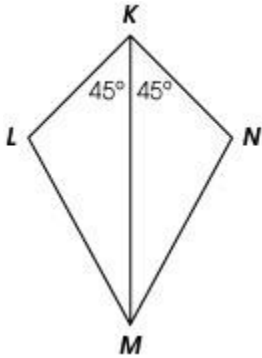
(Max chars: 10,000)

1 point for theorem and 2 points for explanation

Count

Question 5a

5. The diagram below shows triangles KLM and KNM .



a. What other information is necessary to show that triangle KLM is congruent to triangle KNM ? Justify your reasoning.

▲

▼

◀ ▶

(Max chars: 10,000)

2 points

Count

Question 5b

b. What one rigid motion can be used above to position one triangle onto the other to show congruence?

▲

▼

◀ ▶

(Max chars: 10,000)

1 point

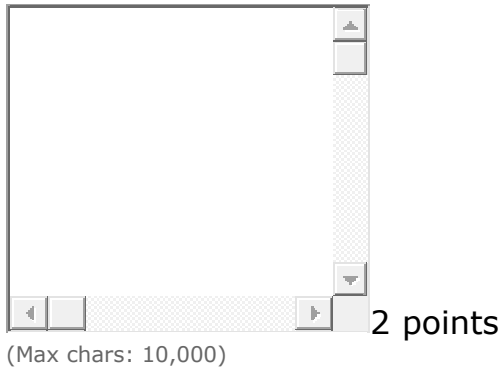
Count

***Question 6**

6. A student stated that triangle ACE is congruent to triangle FHJ because of the following:

- Side AC is congruent to side FH .
- Side CE is congruent to side HJ .
- Angle A is congruent to angle F .

Explain why the student cannot prove that triangle ACE and triangle FHJ are congruent to each other based on the given information.



2 points

(Max chars: 10,000)