opción

Revista de Antropología, Ciencias de la Comunicación y de la Información, Filosofía, Lingüística y Semiótica, Problemas del Desarrollo, la Ciencia y la Tecnología

Año 36, 2020, Especial Nº 26

Revista de Ciencias Humanas y Sociales
ISSN 1012-1587/ ISSNe: 2477-9385
Depósito Legal pp 198402ZU45

Universidad del Zulia
Facultad Experimental de Ciencias
Departamento de Ciencias Humanas
Maracaibo - Venezuela

Opción, Año 36, Especial No.26 (2020): 3115-3141
ISSN 1012-1587/ISSNe: 2477-9385

Evaluation of students' Noncognitive Large Scale Assessments measures: Problem Solving Experiences scale as a case study

Hind Hammouri[1]
[1]Department of Educational Psychology, The Hashemite University, Zarqa, Jordan
Email: hind@hu.edu.jo

Mutasem Mohammad Akour[2]
[2]Department of Curricula and Instruction, The Hashemite University, Zarqa, Jordan
Email: mutasem@hu.edu.jo

Saed Sabah[3]
[3]Department of Educational Psychology, The Hashemite University, Zarqa, 13115, Po Box 150459. Jordan
Email: Sabah@hu.edu.jo

Abstract

This study evaluated students' Noncognitive Large Scale Assessments NLSA measures by examining the psychometric properties of the Problem Solving Experiences (PSE) scale as a case study of this kind of measures. We analysed the responses of 2000 students participated in PISA2012 from Germany, Japan, Jordan, and the United States of America to the personality scales of openness and perseverance (student questionnaire) that were assessed under the construct PSE. Results of the Rasch Rating Scale Model RSM analysis revealed that the structure of the PSE-scale differs between countries. There are problems of category disordering. 10% of items demonstrated misfit to the Rasch model. There are differences and inconsistency in items' endorsability by country. The scale is not unidimensional and lacks sufficient sensitivity to discriminate

Recibido: 20-12-2019 •Aceptado: 20-02-2020

individuals with high levels from those with lower levels of PSE. Accordingly, this study provided evidence that it is difficult to adopt the same noncognitive measure in different cultures in order to provide data that enables making decisions regarding intended purposes. Results of this kind of measures should be handled cautiously.

Keywords: Diverse cultures, Program evaluation, Scale Model, Student.

Running head: evaluación de evaluaciones no cognitivas a gran escala

Resumen

Este estudio evaluó las medidas NLSA de las evaluaciones a gran escala no cognitivas de los estudiantes al examinar las propiedades psicométricas de la escala Experiencias de resolución de problemas (PSE) como un estudio de caso de este tipo de medidas. Analizamos las respuestas de 2000 estudiantes que participaron en PISA2012 de Alemania, Japón, Jordania y los Estados Unidos de América a las escalas de personalidad de apertura y perseverancia (cuestionario estudiantil) que se evaluaron bajo el constructo PSE. Los resultados del análisis RSM del Modelo de Escala de Calificación Rasch revelaron que la estructura de la escala PSE difiere entre países. Hay problemas de desorden de categoría. El 10% de los artículos demostraron ser inadecuados para el modelo Rasch. Existen diferencias e inconsistencias en la endosabilidad de los artículos por país. La escala no es unidimensional y carece de la sensibilidad suficiente para discriminar a los individuos con niveles altos de aquellos con niveles más bajos de PSE. En consecuencia, este estudio proporcionó evidencia de que es difícil adoptar la misma medida no cognitiva en diferentes culturas para proporcionar datos que permitan tomar decisiones con respecto a los fines previstos. Los resultados de este tipo de medidas deben manejarse con cautela.

Palabras clave: diversas culturas, evaluación del programa, modelo a escala, estudiante.

1. INTRODUCTION

Noncognitive constructs together with content knowledge are essential requirements for success in school, ability to solve complex problems and situations in life, and proficiency in any subject. They are considered as prerequisites for cognitive learning and goals of education in themselves (Kuger, et al., 2016). Large scale assessments measure students' cognitive and noncognitive constructs across cultures. Noncognitive assessments include context questionnaires that consist of a series of statements about an issue or construct with which a respondent is asked to indicate his/her degree of agreement. These statements are applied to cross cultural participants in order to collect data that are often used to explain variations in students' performance or reported as learning outcomes (Schulz, 2008). Self-reported data that are used in Noncognitive Large Scale Assessments NLSA involve large number of culture-specific variations that might cause inaccurate perceptions and responses because of respondents' interpretations of item content, linguistic and cultural habits, or personal preferences in using response scales (He & Kubacka, 2015).

One of the most challenging requirements for international educational research is the use of comparable measures of variables affecting targeted outcomes. The scaling of noncognitive questionnaires requires a thorough cross-cultural validation of the

underlying constructs (Schulz, 2008). According to Rutkowski and Svetina (2014), equivalence of cognitive scores across countries in the field of international educational surveys has received substantial attention in academic literature; whereas, only a relatively few emphases on scale score equivalence in noncognitive surveys has emerged. The cross-cultural validity of the noncognitive PISA scales has not been thoroughly investigated compared with the cognitive assessments (Braeken & Blömeke, 2016). He, Barrera-Pedemonte, and Buchholz (2019) recommended that to verify that the same construct is being measured across targeted groups, it is important to give evidence on its validity and suitability to different cultures before conducting any cross-cultural comparisons. Otherwise, it is unclear whether the observed differences across countries are due to true differences in the measured construct or to other factors (Cheung & Rensvold, 2002).

Validity can be tested in an item-response theory IRT framework or a structural equation modeling framework (Reise, Widaman, & Pugh, 1993). Embretson and Reise (2000) indicated that using IRT methods provide a more thorough assessment of the measurement properties of a measure. Accordingly, we used in this study IRT framework.

2. METHODOLOGY

The population of PISA2012 consisted of 65 countries and economies. The ages of PISA students range between 15 years 3

months to 16 years 2 months at the time of the assessment; they have completed at least 6 years of formal schooling (OECD, 2014a). The current study is based on PISA2012 data where mathematics was the major domain, and specifically, on the data from the student questionnaire for the personality scales of perseverance and openness to problem solving.

Four countries were selected for this study representing different cultures; each of them is using different language. The countries are: Germany, Japan, Jordan, and the United States of America USA. The numbers of students from these countries who took PISA2012 were: 5001, 6351, 7038, and 4978 respectively. These countries showed a diversity of performance in mathematics literacy, Japan and Germany achieved higher than population mean score in mathematics. Their means were 536 and 514 respectively. USA and Jordan achieved less than population mean score, their means were 481 and 386 respectively (OECD, 2014b). Moreover, Japan, Germany and USA are developed countries whereas, Jordan is a developing one.

To decide on the number of appropriate sample size that provides stable item and person estimates, we followed Linacre's (2002) suggestion that sample size could be as many as 100*(number of categories). Accordingly, random samples of 500 students from each of the four countries were selected to be the data of this study (2000 students).

3. RESULTS AND DISCUSSION

In order to be able to answer the study questions, PTMEA was calculated (as shown in Table 1) aiming to examine item polarity.

Table 1: Item PTMEA Broken by Country

	ST93Q01	ST93Q03	ST93Q04	ST93Q06	ST93Q07	ST94Q05	ST94Q06	ST94Q09	ST94Q10	ST94Q14
Japan	0.36	0.60	0.62	0.64	0.68	0.69	0.68	0.73	0.68	0.19
Germany	0.40	0.55	0.62	0.66	0.57	0.63	0.65	0.66	0.68	0.15
USA	.50	0.58	0.64	0.65	0.66	0.69	0.59	0.69	0.71	0.07
Jordan	0.14	0.56	0.62	0.62	0.58	0.64	0.61	0.66	0.64	0.31

Table 1 reveals that all correlations are positive; accordingly, item polarity is maintained. This result indicates that these items were rescored accurately and properly reflect the rating scale, so we can pursue the analysis. However, PTMEA of item ST94Q14 is less than 0.40 in the four countries, whereas item ST93Q01 has PTMEA less than 0.40 in Jordan and Japan, only.

To answer question (1.1), Response category functioning was evaluated according to the following Linacre's (2002) criteria: a) the shape of the rating-scale distribution is smooth and contained a unimodal progressive increase in the frequency with which each

ordered rating category was chosen; b) average respondent measure associated with each category increases with the increasing values of the categories; c) Outfit/MNSQ is greater than 0.5 and less than 1.5; d) category thresholds indices increase monotonically with the values of categories; and e) adjacent category thresholds are at least 1.4 logits and no more than 5 logits apart. The results are shown in Table 2 which reveals that:

- As for criterion (a)

- Rating-scale distribution is unimodal in Japan, Germany and USA, it is ascending in Jordan.

- All categories are frequently used in the four countries.

- Category 1 has the least observations in Germany, USA and Jordan; whereas category 5 has the least observations in Japan. While Category 4 has the highest observations in both Germany and USA; categories 3 and 5 has the highest observations in Japan and Jordan respectively. So, there are irregularities in the frequencies across categories in the four countries which may indicate abnormal category use.

Table 2: Response Structure of PSE-Scale Broken by Country

	Category Label	% Observed Count	Average measure	Outfit/MNSQ	Threshold calibration
Japan	1	11	-1.37	1.01	-
	2	29	-0.55	1.01	-1.86

	3	33	0.00	0.95	-0.43
	4	18	0.46	0.91	0.52
	5	9	0.95	1.29	1.47
Germany	1	5	-0.81	1.17	-
	2	15	-0.25	1.12	-1.63
	3	29	0.27	0.87	-0.67
	4	34	0.88	0.89	0.43
	5	18	1.62	1.08	1.87
USA	1	4	-0.60	1.10	-
	2	12	-0.03	1.28	-1.40
	3	31	0.35	0.86	-0.77
	4	32	0.89	0.86	0.57
	5	21	1.55	1.05	1.60
Jordan	1	6	-0.40	1.35	-
	2	10	-0.03	0.99	-0.88
	3	16	0.36	0.86	-0.23
	4	31	0.95	0.93	0.09
	5	38	1.54	1.25	1.02

- The four countries met criteria (b) to (d).
- As for criterion (e), Table 2 reveals that although there is the desired monotonic progression from each step calibration to the next in the four countries, some of the calibrations intervals between the steps are smaller than the recommended 1.4 to 5.0 logit, (e.g. steps 3 & 4 in

all countries), indicating that some categories in the four countries are practically inseparable.

The previous results indicate that response categories are not functioning as expected.

To assess response patterns of the items (question 1.2), we:

1) Calculated Outfit/MNSQ for each item in the four countries. Its values range between 0.6 and 1.4 according to Wright & Linacre (1994) criterion. Table 3 reveals that across the four countries the outfit/MNSQ values for item ST94Q14 did not meet the criterion, whereas, item ST93Q01 met the criterion in only the USA, indicating that these items might not tap the same trait as the other items of the instrument (Boone, 2019), These values are even greater than 1.5 logits, so these items have problems because there is more than 50% unexplained randomness (Smith, 1996).

Table 3: Outfit MNSQ Values for PSE-Scale Broken by Countr

	Japan		Germany		USA		Jordan	
Item	Measure	Outfit MNSQ	Measure	Outfit MNSQ	Measure	Outfit MNSQ	Measure	Outfit MNSQ
ST93Q01	0.46	1.75	0.49	1.55	0.13	1.26	1.37	2.83
ST93Q03	-0.13	0.82	-0.21	0.96	-0.03	0.93	-0.32	0.91
ST93	0.08	0.94	0.24	0.87	-0.03	0.94	-0.31	0.75

Q04								
ST93 Q06	0.53	0.80	1.05	0.90	0.25	0.89	0.04	0.84
ST93 Q07	-0.10	0.73	-0.61	0.78	-0.13	0.74	-0.29	0.73
ST94 Q05	-0.24	0.80	-0.66	0.67	-0.17	0.73	-0.22	0.58
ST94 Q06	-0.27	0.80	-0.63	0.74	-0.38	0.90	-0.24	0.75
ST94 Q09	-0.07	0.67	-0.41	0.72	-0.17	0.69	-0.03	0.64
ST94 Q10	0.55	1.17	0.70	1.08	0.66	1.01	0.49	1.09
ST94 Q14	-0.81	1.89	0.02	1.99	-0.14	2.29	-0.49	1.73

Moreover, in USA and Jordan Outfit/MNSQ are greater than 2 for items ST94Q14 and ST93Q01 respectively, indicating that there is more misinformation than information in the observations (Linacre, 2002). The above mentioned results reveals that 10% of items demonstrated misfit to the Rasch model in the four countries, and another 10% fits the model in only USA. Item misfit indicates, basically, that the respondents rated this item inconsistently in relation to their overall response pattern. The misfiting items are misperforming for targeted students (Linacre, 2012), unpredictable, have too much variations, and too haphazard response pattern (Bond &

Fox, 2007). These results may indicate a general misfit of these items across PISA population; which might be the result of carelessness; response set answering; or item bias.

2) Plotted empirical item category measures for all items in the four countries as depicted in figure 1.

Japan	Germany
-5 -4 -3 -2 -1 0 1 2 3 |-----+-----+-----+-----+-----+-----+----| ITEM | 12 3 4 5 | ST93Q01	-3 -2 -1 0 1 2 3 4 |------+------+------+------+------+------+----+-ITEM | 213 4 5 | ST93Q01
| 1 2 3 4 5 | ST93Q03	| 12 3 4 5 | ST93Q03
| 1 2 3 4 5 | ST93Q04	| 1 2 3 4 5 | ST93Q04
| 1 2 3 4 5 | ST93Q06	| 1 2 3 4 5 | ST93Q06
| 1 2 3 4 5 | ST93Q07	| 1 2 3 4 5 | ST93Q07
| 1 2 3 4 5 | ST94Q05	| 1 2 3 4 5 | ST94Q05
| 1 2 3 4 5 | ST94Q06	| 1 2 3 4 5 | ST94Q06
| 1 2 3 4 5	| 1 2 3 4 5

| ST94Q09	| ST94Q09
| 1 2 34 5 | ST94Q10	| 1 2 3 4 5 | ST94Q10
| 2 435 | ST94Q14 |-----+-----+-----+-----+-----+----- +---- ITEM -5 -4 -3 -2 -1 0 1 2 3	| 432 5 | ST94Q14 ------+------+------+------+------+--- ---+-- ITEM 3 -2 -1 0 1 2 3 4
USA	Jordan
-3 -2 -1 0 1 2 3 |-------+-------+-------+-------+----- --+-- ITEM	-4 -3 -2 -1 0 1 2 3 |------+------+------+------+------+-- ----+- ITEM
| 1 23 4 5 | ST93Q01	| 3241 5 | ST93Q01
| 2 1 3 4 5 | ST93Q03	| 1 2 3 4 5 | ST93Q03
| 1 2 3 4 5 | ST93Q04	| 1 2 3 4 5 | ST93Q04
| 1 2 3 4 5 | ST93Q06	| 1 2 3 4 5 | ST93Q06
| 1 2 3 4 5 | ST93Q07	| 1 2 3 4 5 | ST93Q07
| 1 2 3 4 5	| 1 2 3 4 5

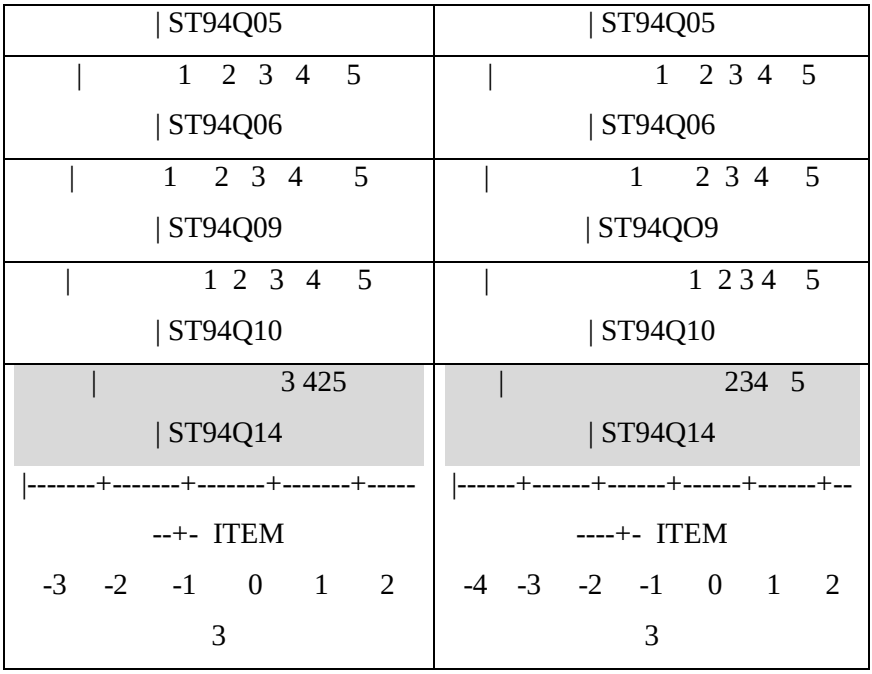


Figure1**:** Average Persons' Measures by Category for PSE-Scale Broken by Country

Figure 1 reveals that:

1) For some items (e.g. ST93Q04), all categories are observed and there are smooth advances everywhere in the categories, for each of the four countries.

2) Responses to some items of PSE-scale (highlighted rows in figure1) do not correspond with the levels of the construct. (e.g. ST93Q01 and ST94Q14 in Japan, Germany and Jordan), as shown in Table4.

Table 4: Average Measure of Problematic Items across Countries

Item		Japan		Germany		USA		Jordan	
	category	Average measure	Problem(s)	Average measure	Problem(s)	Average measure	Problem(s)	Average measure	Problem(s)
ST93Q01	1			0.24	disordering between categories 1, 2			1.07	disordering between categories 1, 2,3, & 4
	2			0.22*				0.87*	
	3			0.40				0.73*	
	4			0.83				0.91*	
	5			1.55				1.84	
ST93Q03	1					-0.06	disordering between categories 1 & 2		
	2					-0.24*			
	3					0.34			
	4					0.83			
	5					1.78			
ST94Q14	1	----	missed category 1	---	missed category 1	----	missed category 1	-----	missed category 1
	2	-		0.52		0.82		0.45	

		0.43							
	3	-0.10	disordering between categories 3 and 4	0.50 *	disordering between categories 2,3, and 4	0.54 *	disordering between categories 2,3, and 4.	0.64	
	4	-0.19 *		0.48 *		0.70 *		0.71	
	5	0.24		0.99		0.93		1.26	

Table 4 is self-explanatory and sheds light on these items to elaborate their problems. it shows that:

- The four countries have problems with at least one item; these problems are not necessarily the same.
- Some items have disordering between some categories (e. g. item ST93Q01 has disordering between all categories in Jordan). Disordering degrades the interpretability of the resulting measures, and can indicate that a category corresponds to a concept is poorly defined in the minds of the respondents and that the item is tapping a misconception (Linacre, 2002).
- Category 1 is not observed in item ST94Q14 in all countries
- Other problems may be combination of more than one (e. g. ST93Q14 in Japan, Germany and USA. However, this item has different problem(s) in the four countries).

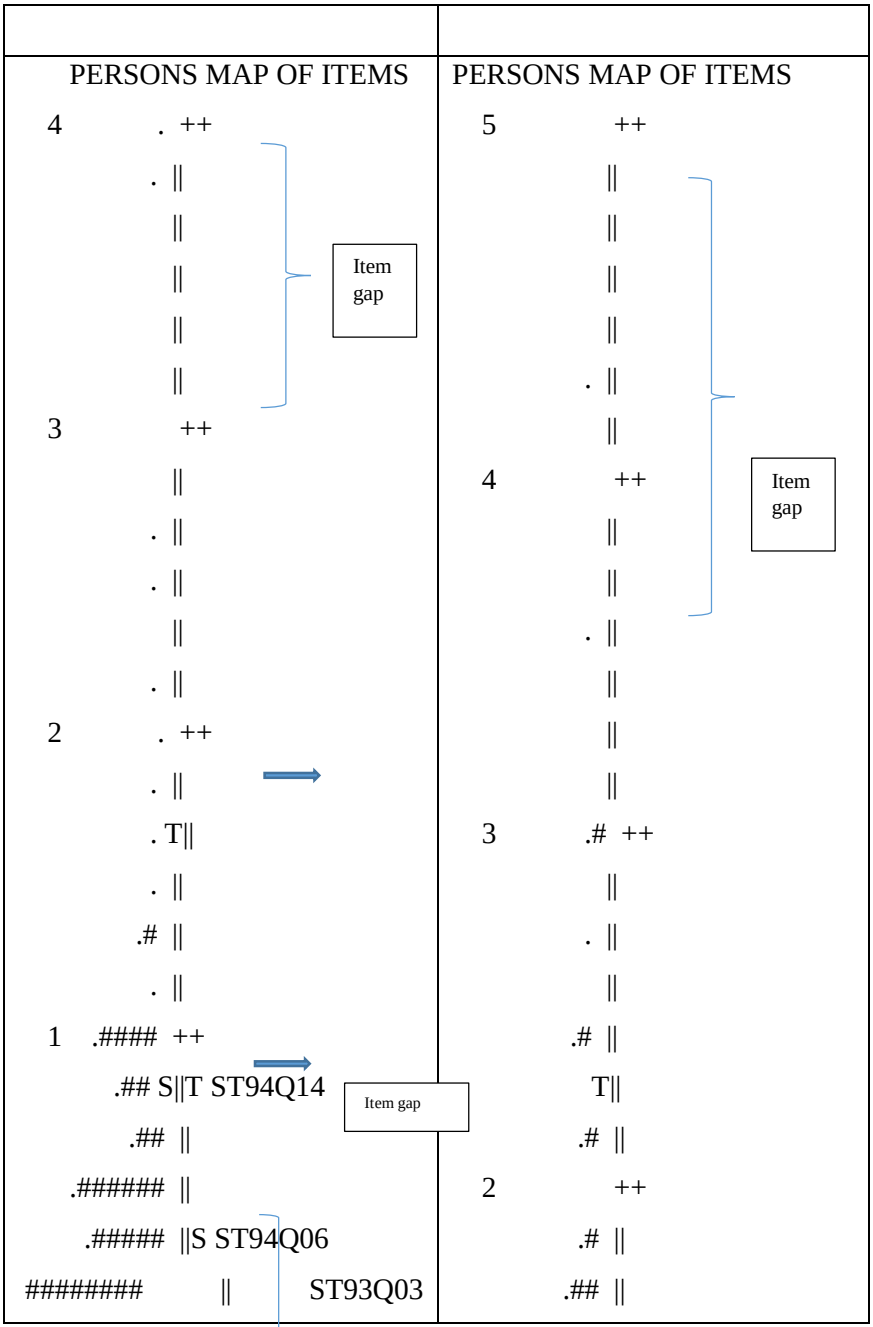
Furthermore, taking into consideration that item ST94Q14 has PTMEA<0.4, in the four countries, and item ST93Q01 has

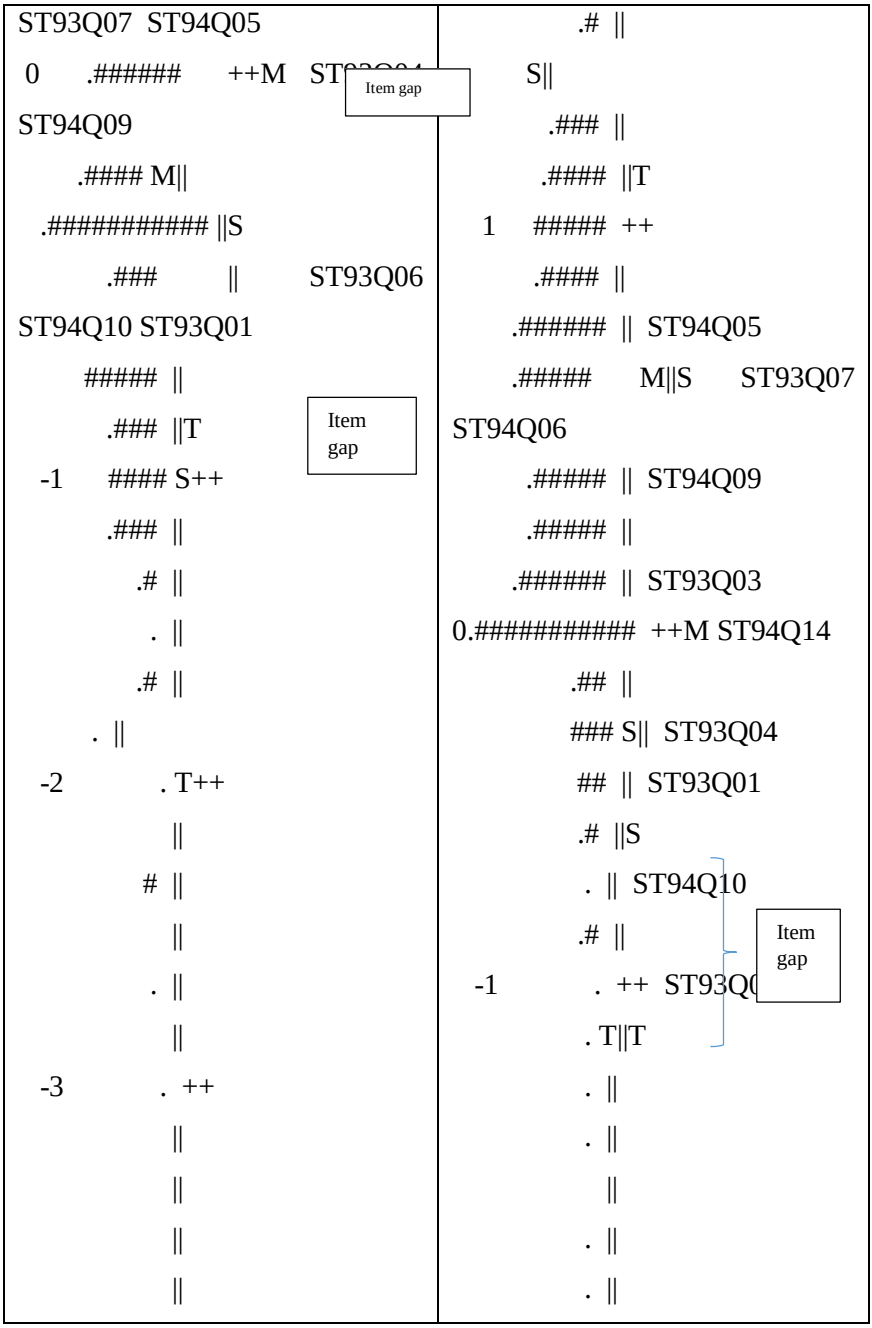
PTMEA<0.40 in Jordan and Japan, this result requires further investigation regarding the relationship between PTMEA and outfit/MNSQ. To allow research results to be communicated in a useful manner, Wright maps are displayed for each of the four countries as depicted in figures 2 and 3. The use of a Wright map enables one to document the hierarchy of items as expressed by the targeted respondents (Boone, Staver, & Yale, 2014). This can be done through visualize the targeting of the test to the sample, as well as the targeting of individual items to persons. (Planinic, et al., 2019). Although there were no a priori hypotheses about the item hierarchy, figures 2 and 3 indicate that item hierarchy is not stable for the four countries. They also reveal that:

1. The persons' mean is higher than the items' mean, except in Japan. This result indicates that, as a group, the PSE items were relatively easy to endorse for the three countries.

2. Items are not distributed evenly across the PSE levels. There are item redundancies in each of the countries. Meaning that there are more than one item measuring similar portions of the trait, (e.g. three items: ST93Q03, ST93Q07 and ST94Q05 with the same endorsability targeted the same PSE-level in Japan, whereas, no items targeted the same group level in Germany). Accordingly, within these groups of items, individual items can be removed with little measurement precision lost (Boone, Staver, & Yale, 2014).

Japan	Germany


PERSONS MAP OF ITEMS
4
3
2
1
Item gap
ST94Q14
Item gap
ST94Q06
ST93Q03
PERSONS MAP OF ITEMS
5
4
3
2
Item gap


ST93Q07 ST94Q05
0 .###### ++M ST
Item gap
ST94Q09
.#### M||
.########### ||S
.### || ST93Q06
ST94Q10 ST93Q01
||
.### ||T
-1 #### S++
.### ||
.# ||
. ||
.# ||
. ||
-2 . T++
||
||
||
. ||
||
-3 . ++
||
||
||
||
Item gap
.# ||
S||
.### ||
.#### ||T
1 ##### ++
.#### ||
.###### || ST94Q05
.##### M||S ST93Q07
ST94Q06
.##### || ST94Q09
.##### ||
.###### || ST93Q03
0.########### ++M ST94Q14
.## ||
S|| ST93Q04
|| ST93Q01
.# ||S
. || ST94Q10
.# ||
-1 . ++ ST93Q
. T||T
. ||
. ||
||
. ||
. ||
Item gap

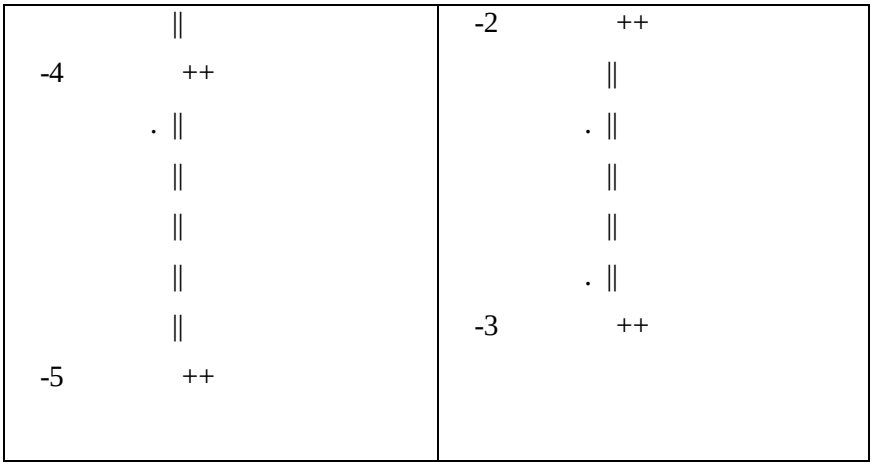
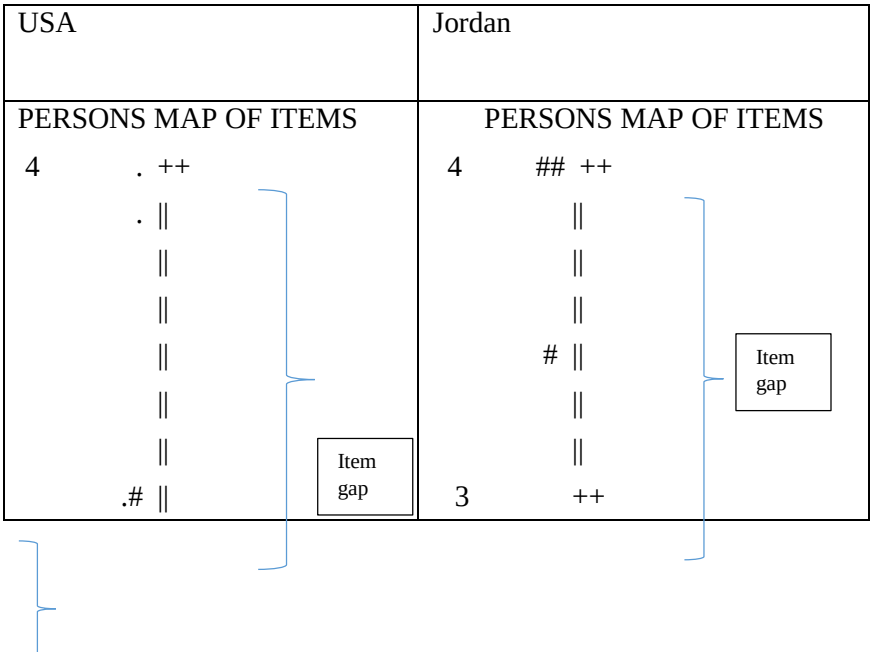
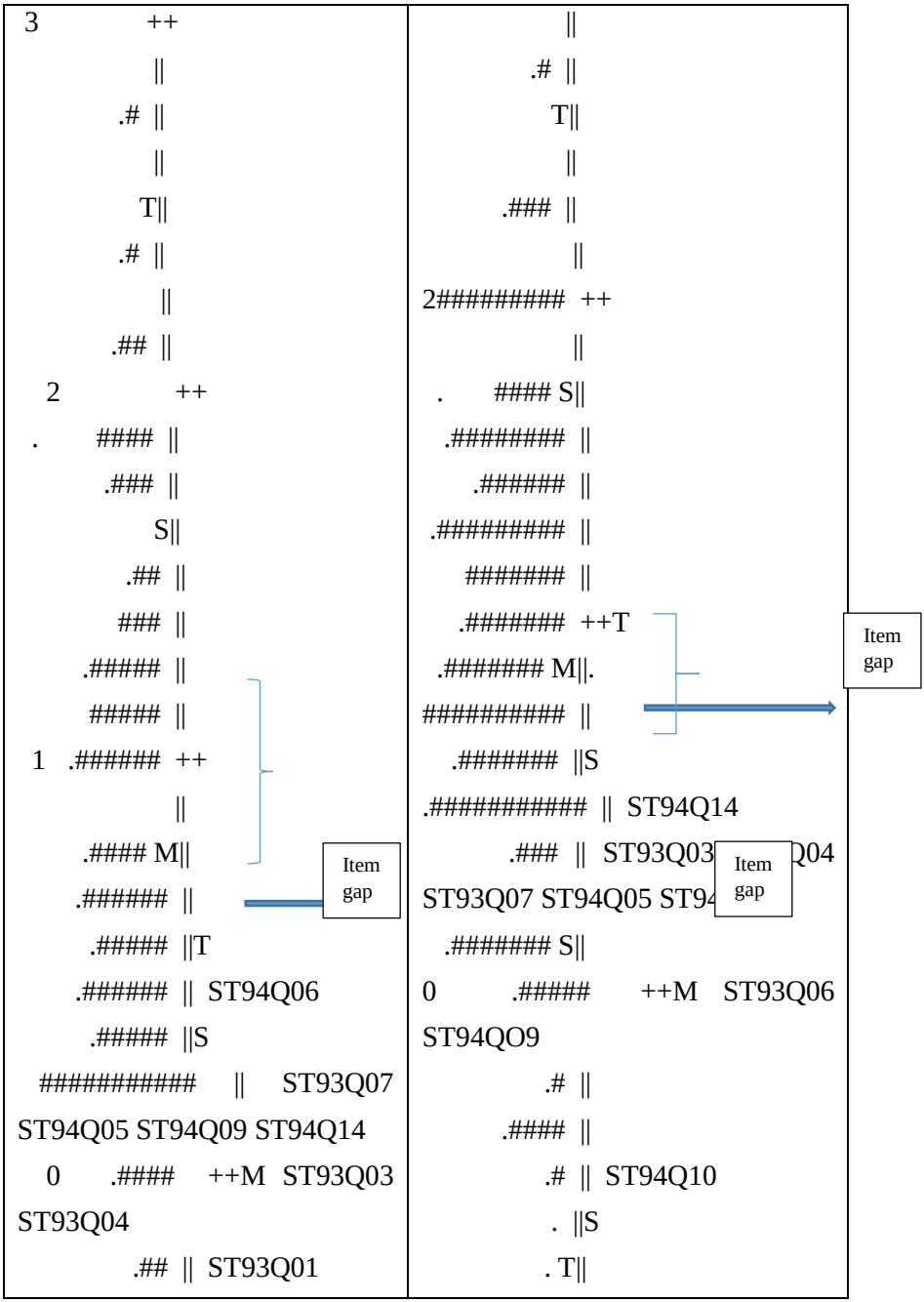


Figure 2**:** Wright Map of PSE for Japan and Germany

3. There are item gaps -students of different PSE-levels were not targeted with any item- (e.g. students with PSE-level less than half standard deviation from mean in Japan are not targeted with any item).




3 ++
||
.# ||
||
T||
.# ||
||
.## ||
2 ++
. #### ||
.### ||
S||
.## ||
||
.##### ||
||
1 .###### ++
||
.#### M||
.###### ||
.##### ||T
.###### || ST94Q06
.##### ||S
########## || ST93Q07 ST94Q05 ST94Q09 ST94Q14
0 .#### ++M ST93Q03 ST93Q04
.## || ST93Q01
||
.# ||
T||
||
.### ||
||
2######### ++
||
. #### S||
.######## ||
.###### ||
.######### ||
####### ||
.####### ++T
.####### M||.
########## ||
.####### ||S
########## || ST94Q14
.### || ST93Q03 Q04 ST93Q07 ST94Q05 ST94
.####### S||
0 .##### ++M ST93Q06 ST94QO9
.# ||
.#### ||
.# || ST94Q10
. ||S
. T||
Item gap
Item gap
Item gap

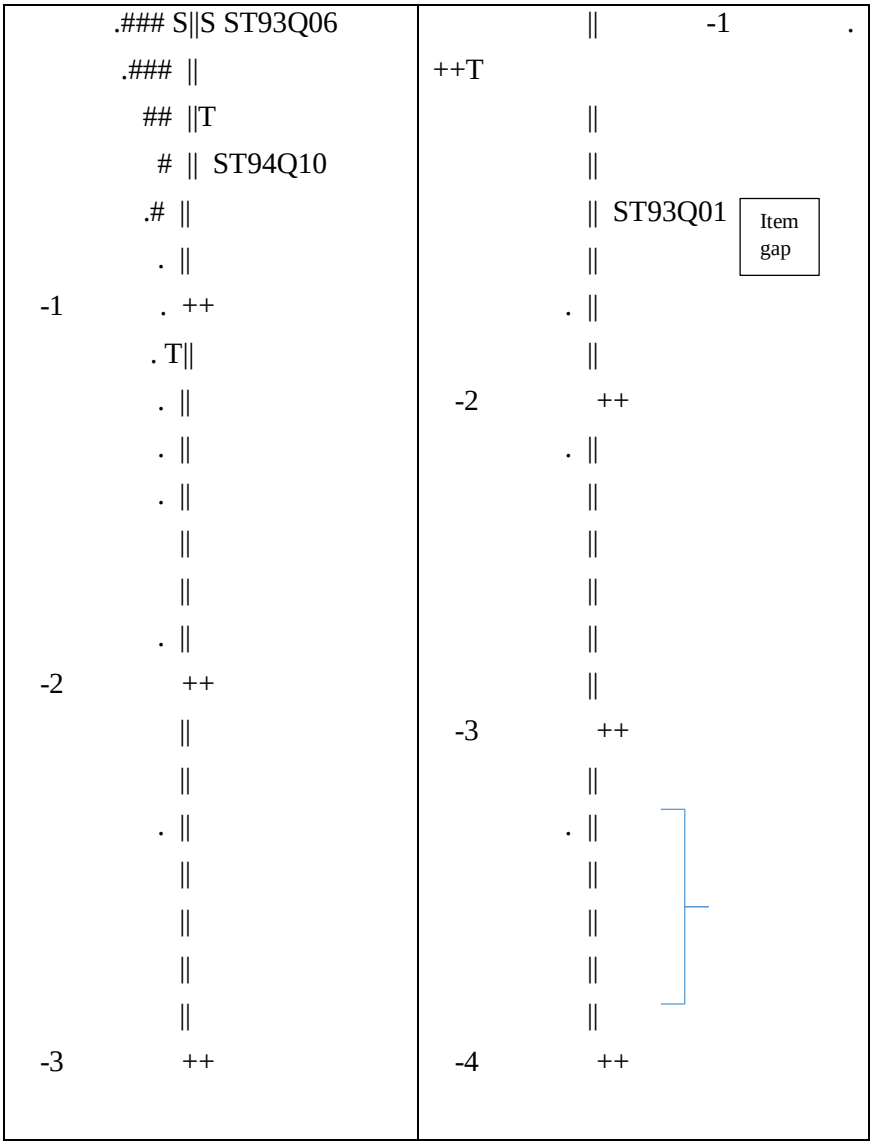


Figure 3. PSE Wright Map for USA and Jordan

The presence of large gaps between the item endorsabilites in the scale means that persons within those gaps cannot be measured

precisely enough due to lack of close items near their PSE-level. A person of PSE-level B_n is best measured with items of difficulties to endorse within ±1 logit from B_n. To fix such gaps in a scale, new items of appropriate endorsability need to be added to it. Since a well-constructed instrument should match the width of PSE-levels distribution of the target population with the width of the distribution of test items (Planinic, et al., 2019).

4. The level of item's endorsability differs by country, (e.g. item ST93Q14 was the most difficult to endorse in both Japan and Jordan, it was of a mean endorsability in Germany and nearly of the same level in USA). The differences and inconsistency in items' endorsability by country could be attributed to culture and personal attributes that may influence the degree to which respondents endorse certain item statements.

The above mentioned results reveal that the structure of the PSE-scale differs between countries. To answer the second question, Rasch PCAR was performed. The following Linacre's (2012) criteria were used to judge the unidimensionality: a) at least 50% of the total variance should be explained by the first latent dimension; and b) any additional factor should explain less than 5% of the remaining variance, after removal of the first latent dimension. Table 5 shows the results for each of the four countries.

Table 5: Standardized Residual Variance Broken by Country

	Japan		Germany		USA		Jordan	
	Eige	%	Eige	%	Eige	%	Eige	%

	n-value	variance	n-value	variance	n-value	Variance	n-value	variance
Variance explained by measures	11.3	53.0%	11.1	52.7%	9.0	47.3%	13.1	56.8%
Unexplained variance (total)	10.0	47.0%	10.0	47.3%	10.0	52.7%	10.0	43.2%
Unexplained variance explained by: 1st factor	2.4	11.3%	2.2	10.5%	2.5	13.2%	2.1	8.9%
2nd Factor	1.4	6.5%	1.4	6.7%	1.4	7.3%	1.6	7.1%
3rd Factor	1.3	6.0%	1.3	6.2%	1.2	6.3%	1.3	5.6%

Table 5 reveals that:

- PSE-scale met the first criterion, in Germany, Japan, and Jordan, only.

-The first, second and third secondary dimensions explain more than 5% of the unexplained variance in the four countries. Accordingly, we can say that this scale is not unidimensional across the four countries. Since Rasch PCAR provides an estimation of internal construct validity by examining not only the hypothesized construct but also the error left over from extracting the construct from the data (Waugh & Chapman, 2005), so combining "perseverance" and "openness to problem solving" constructs into PSE is not supported by Racsh PCAR. Even the two constructs are not supported by the model in any of the for countries. To answer the third question, the Rasch person reliability and separation indices were produced in table 6. These indices provide descriptive information showing the number of different groups within the sample, and how well the persons are separated on a linear continuum. Person separation is used to classify people. This index refers to the precision of measurement and indicates how well one can differentiate examinees levels with a scale (Bond & Fox, 2007). According to Linacre (2012), this index should be at least 2.

Rasch reliability of person responses reflects the true score variance to observed score variance, which is based on the estimated locations of persons along the measurement continuum (Fisher, 1992). The lowest person reliability for any decision making involving students' abilities is 0.8 (Linacre,2012).

Table 6: Person Separation and Reliability Broken by Country

	Japan	Germany	USA	Jordan
Person Separation	1.81	1.61	1.74	1.42
Person Reliability	0.77	0.72	0.75	0.67

Table 6 reveals that person separation indices are less than 2.0 in the four countries. Low person separation implies that the instrument lacks sufficient sensitivity to discriminate individuals with high levels from those with lower levels of PSE. More items may be needed (Linacre, 2012). Person reliability indices are less than 0.80 for the four countries; this suggests that persons' responses are inconsistent (Bond & Fox, 2007). In light of the previous results, it could be said that this rating scale is not functioning effectively. The items of the PSE-scale were revisited and carefully reviewed by the study researchers. The following remarks were observed in its items which may negatively affect the scale validity.

1. the use of negative sounding phrases such as: ''gives up easily when confronted with a problem'', may invite a negative mindset or may initiate a defensive attitude in respondents and interpret the item as if it is affirmative.

2. the phrase "solve complex problems" may have different meanings for these adolescents according to their cultures. Are they mathematical problems or real life ones?

3. the choice of scaling in which "very much like me" is assigned 1 and "not at all like me" assigned 5 is a bit contrary to the common sense expectation for many respondents.

4. the use of five rating scale steps can confuse respondents and degrade the quality of data collected.

5. the use of positive sounding phrases such as “seeks explanations for things” combined with the adopted choice of scaling can create complicated problems for the respondents.

6. the use of modifiers such as those in item ST94Q05 ‘‘lot of information’’ and item ST93Q06 ‘‘everything is perfect’’ could add to respondent confusion—what is ‘‘lot” and who does “everything perfect’’?

REFERENCES

Beaty, R, Nusbaum, E, & Silvia, P. (2014). Does insight problem solving predict real-world creativity? Psychology of Aesthetics, Creativity, and the Arts,8(3), 287-292.

Bond, T., &Fox, C. (2007). Applying the Rasch Model: Fundamental measurement in the human sciences. 2nd edition. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Boone, W. (2019). Rasch Methods for Beginners, Paedagogisk Psykologisk Tidsskrift, Nr 1, 82-95.

Boone, W., Staver, J., & Yale, M. (2014). Rasch Analysis in the Human Sciences. Springer Science + Business Media Dordrecht Heidelberg New York London. DOI 10.1007/978-94-007-6857-4.

Braeken, J., & Blömeke, S. (2016). Comparing future teachers’ beliefs across countries: Approximate measurement invariance with Bayesian elastic constraints for local item dependence and differential item

functioning. Assessment & Evaluation in Higher Education, 41, 733–749. doi:10.1080/02602938.2016.1161005.

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

Embretson, S., & Reise, S. (2000). Item response theory for psychologists. Mahwah, NJ: Lawrence Erlbaum Associates.

Fernandez-Cano, A. (2016). A Methodological Critique of the PISA Evaluations. RELIEVE,22(1), art. M15. DOI: http://dx.doi.org/10.7203/relieve.22.1.8806.

Fisher, W. (1992). Reliability, separation strata statistics. Rasch Measurement Transaction,6, 238.

Harris, M., Gibson, S., &Mick, T. (2009). Examining the Relationship between Personality and Entrepreneurial Attitudes: Evidence from U.S. College Students. Small Business Institute Journal,3, 21–53.

UNIVERSIDAD
DEL ZULIA

opción
Revista de Ciencias Humanas y Sociales

Año 36, N° 26, (2020)

Esta revista fue editada en formato digital por el personal de la Oficina de Publicaciones Científicas de la Facultad Experimental de Ciencias, Universidad del Zulia.
Maracaibo - Venezuela

www.luz.edu.ve

www.serbi.luz.edu.ve

produccioncientifica.luz.edu.ve