JP3I (Jurnal Pengukuran Psikologi dan Pendidikan Indonesia)
DOI: http://dx.doi.org/10.15408/jp3i.v13i1.27975
http://journal.uinjkt.ac.id/index.php/jp3i

The Impact of Sample Size, Test Length, and Person-Item Targeting on The Separation Reliability in Rasch Model: A Simulation Study

Rahmat S. Bintang[1], Suprananto[2]

Faculty of Psychology, UIN Syarif Hidayatullah Jakarta, Indonesia[1]
Faculty of Teacher Training and Education, Universitas Singaperbangsa Karawang, Indonesia[2]

E-mail: rahmatsbintang04@gmail.com

Abstract

This research is a simulation study using resampling methods to see the impact of sample size, test length, and person-item targeting on separation reliability in the Rasch model. Simulation conditions were created with several defined factors, namely sample size with five conditions (200, 500, 1000, 2000, and 4000 people), test length with three conditions (20, 40, and 60 items), and person-item targeting with five conditions (-2, -1, 0, 1, and 2). The total amount of conditions is 75 conditions in which each condition is replicated 50 times to produce a total of 3.750 data. Data is generated using WinGen software. The results of the separation reliability analysis are analyzed using Winsteps software. The separation reliability criteria are defined for Person Separation Reliability (PSR) > 0.80 and Item Separation Reliability (ISR) > 0.90. The results of the study showed that 75 conditions (100%) resulted in an estimate of ISR that met the criteria (> 0.90). For the PSR estimate, 37 conditions (49%) produced PSR assessments that met the criteria (> 0.80) and 38 conditions (51%) produced an estimate of PSR that did not meet the criteria (< 0.80). In addition, the PSR estimate is influenced by the test length and person-item targeting.

Keywords: sample size, test length, person-item targeting, person separation reliability, item separation reliability

Abstrak

Penelitian ini merupakan studi simulasi menggunakan resampling methods untuk melihat pengaruh ukuran sampel, panjang tes, dan person-item targeting terhadap separation reliability dalam Rasch model. Kondisi simulasi diciptakan dengan beberapa faktor yang telah ditetapkan yaitu ukuran sampel dengan lima kondisi (200, 500, 1000, 2000, dan 4000 orang), panjang tes dengan tiga kondisi (20, 40, dan 60 item), dan person-item targeting dengan lima kondisi (-2, -1, 0, 1, dan 2). Jumlah kondisi secara keseluruhan sebanyak 75 kondisi di mana tiap kondisi direplikasi sebanyak 50 kali sehingga dihasilkan total data sebanyak 3.750 data. Data dibangkitkan dengan menggunakan software WinGen. Untuk hasil analisis separation reliability dianalisis dengan menggunakan software Winsteps. Kriteria separation reliability yang ditetapkan yaitu untuk Person Separation Reliability (PSR) > 0.80 dan untuk Item Separation Reliability (ISR) > 0.90. Hasil penelitian menunjukkan bahwa 75 kondisi (100%) menghasilkan estimasi ISR yang memenuhi kriteria (> 0.90). Untuk estimasi PSR, 37 kondisi (49%) menghasilkana estimasi PSR yang memenuhi kriteria (> 0.80) dan 38 kondisi (51%) menghasilkan estimasi PSR yang tidak memenuhi kriteria (< 0.80). Selain itu, estimasi PSR dipengaruhi oleh panjang tes dan person-item targeting.

***Kata kunci**:* ukuran sampel, panjang tes, person-item targeting, person separation reliability, item separation reliability

Introduction

Measurement in psychology is a core discussion in the academic literature (Guyon et al., 2018). One definition of the fundamental measurement is that put forward by Stevens (1946) that measurement in the field of psychology is giving a number to an object or event according to the rules. There are standards for evaluating the psychometric characteristics of a test or instrument that researchers in psychology must meet (APA, AERA, & NCME, 1999). Validity and reliability are important indices in these standards which are the main concepts in measurement (Wright & Stone, 1999). Because validity and reliability are important indices in describing the quality of measurement, they have been widely studied since almost a century ago (Raykov & Marcoulides, 2019).

Over time, test theory was developed which is a general theoretical framework for assessing measurement quality (Wright & Stone, 1979). Three measurement theories are currently developing and are widely used in designing and analyzing a test (Andrich, 2011). The first theory is the classical test theory (CTT; classical test theory) which was developed in 1940 and has been widely used (Brennan, 2011; Hambleton et al., 1991). The second test theory is item response theory (IRT; Lord, 1952) which was developed to overcome the limitations of CTT such as the existence of group-dependent and item-dependent properties (Hambleton et al., 1991). The third test theory is Rasch measurement theory which is a modern test theory in its simple form in terms of the number of parameters in the model (Andrich, 2011).

Rasch Measurement Theory (RMT) is a measurement model developed by George Rasch in 1960 (Embretson & Reise, 2000). There are four purposes of measurement using the Rasch model analysis, namely: (1) placing person's abilities and item difficulty on the same interval scale, (2) single measurement dimension, (3) item calibration does not depend on the distribution of person's ability levels, and (4) the measurement of person does not depend on the distribution of item difficulty level. In general, these four properties are often covered by the term objective measurement (Karabatsos, 2000; Hayat et al., 2020).

In the context of the Rasch model, measurement is very important to know whether the instrument used is good enough to sort a person. It is important to place items and persons along a continuum to see where items and persons are located. Separation of the location of items and persons on a continuum is done to produce a good measurement. Therefore, the items must be well separated based on the difficulty of the items and the selection of items in the test must be able to separate a person based on a person's abilities. Separation statistics for items and persons in the Rasch model provide a medium for evaluating the successful use of the test. This statistic is referred to as reliability which has a range of values between 0.00 – 1.00, where the higher the value, the better the separation (Wright & Stone, 1999).

In 1982, Wright and Masters developed a reliability coefficient based on the Rasch measurement model which has a different concept from classical test theory (e.g., Cronbach's alpha) (Suryadi et al., 2020). Separation reliability in the Rasch model includes the separation of persons and the separation of items (Wright & Stone, 1979). For the separation of persons, it is displayed through person separation reliability, and for the separation of items, it is displayed through item separation reliability. Person separation reliability is used to classify a person, while item separation reliability is used to classify item difficulty hierarchies (Linacre, 2018).

Separation reliability is important to know to distinguish objects based on the characteristics or attributes being measured. Low person separation reliability (PSR < 0.8) with a sufficient sample indicates that the instruments used were not good enough to differentiate between groups high and low groups. Therefore, many items are required. Meanwhile, a low item separation reliability (ISR < 0.9) indicates that the number of samples used is not large enough or the sample is not sufficient to be able to differentiate the difficulty hierarchy of items from the instruments used (de Ayala, 2009; Linacre, 2018). For example, the resulting PSR value in the mathematical exam is 0.08. This value indicates that the mathematics instrument does not appear to be doing a good job of distinguishing people. Once seen, the mathematical instrument has only five items (de Ayala, 2009).

There are several factors that affect the separation reliability in the measurement using the Rasch model. These factors include sample size, test length, person-item targeting, and response format (Linacre, 2018). This study aims to determine the interrelationships between these variables in influencing separation reliability by using a simulation study using generation data. The simulation study approach can overcome the limitations of studies with empirical data which is time-consuming and expensive and often the data collected is incomplete. In addition, another limitation of using empirical data is that often the data that has been collected is not sufficient for analysis (Feinberg & Rubright, 2016).

As a solution to these limitations by using empirical data, another method can be used, namely by using a simulation study (Wright & Douglas, 1977). One of the advantages of simulation studies is that it allows researchers to compare parameter estimates against each unknown true parameter when using real data (Feinberg & Rubright, 2016). Studies using data simulations became popular in the late 19th and early 20th centuries which were used in various fields of science (Feinberg & Rubright, 2016). There are various methods for data simulation studies, one of which is resampling methods. Resampling methods use a computer to generate some simulation data and then analyze and summarize the patterns in the data (Casey & Harden, 2014).

Separation Reliability

The use of separation statistics between persons and items used in the measurement of the Rasch model is still foreign to many researchers, especially those who do not study the Rasch model approach (Wright & Stone, 1999). The location of persons and items on the same continuum is needed to see the accuracy of the measurement. The statistics on the separation of persons and items in the Rasch measurement provide an analytical method used to evaluate the success of test development and can be used to see the continued usefulness of the test. Person separation indicates how efficiently a test can separate the person being measured. Item separation indicates how well a sample of a person can separate the items used in the test. This statistic is expressed as reliability, where the value ranges from 0 to 1.

Person Separation Reliability (PSR)

PSR is an estimate of how well the instrument can distinguish respondents on the measured variable. The equations of the PSR are as follows (Wright & Masters, 1982):

$$\mathbf{PSR} = \frac{\mathbf{OV}-\mathbf{EV}}{\mathbf{OV}}$$

Where:

OV = observed variance of a person's ability measures

EV = mean of the squared standard error of a person's ability measures

Item Separation Reliability (ISR)

Reliability coefficients for items in the Rasch model are described in terms of ISR (Wright & Stone, 1999). The ISR provides information to test users on how well the items sort out the person taking the test. The ISR equation is as follows:

$$\mathbf{ISR} = \frac{\mathbf{OV}-\mathbf{EV}}{\mathbf{OV}}$$

Where:

OV = observed variance of item difficulty measures

EV = mean of squared standard errors of item difficulty measures

Dichotomous Rasch Model

Someone who has higher abilities than others should have a greater chance of answering the item of each question correctly, and in the same way, one item that is harder than the other item means that the chance to answer the difficult item requires a higher ability (Rasch, 1960). Instruments that have the type of correct and wrong answer and the correct answer are scored 1 and the wrong answer is scored 0, this is called a dichotomy item (Andrich & Marais, 2019). The simplest response format is having two levels of choice on an item. This model is the simplest model which has one parameter to describe the characteristics of the item and one parameter to describe the characteristics of a person. The equations of the Rasch dichotomous model are as follows:

$$\mathbf{P(x_j = 1 \mid \theta, \delta_i)} = \frac{e^{(\theta-\delta_i)}}{1+e^{(\theta-\delta_i)}}$$

Where p ($x_j = 1|\theta, \delta_i$) is the probability of response 1 ($x_j = 1$), θ is the location of the person, and δ_i is the location of the item i (Wright & Masters, 1982). Hayat (1995) states that based on the above

equation, it can be understood that "the probability of the j-th person getting a score of 1 (answering correctly/agreeing) on the i-th item is determined by the result of the interaction between the location of the person (the ability of the person) and the location of the i-th item (item difficulty)". In simple terms it can be concluded as follows:

1. If θ (person's ability) > δ (item's level of difficulty), then the person's chance of correctly answering the item is above 50%.
2. If θ (person's ability) < δ (item's level of difficulty), then the person's chance of correctly answering the item is below 50%.
3. If θ (person's ability) = δ (item's level of difficulty), then the person's chance of correctly answering the item is equal to 50%.

Methods

This research is a simulation study to determine the accuracy of the results connecting the variables used in producing separation reliability (ISR and PSR) in the Rasch model. The method used in generating simulation data is the resampling method. The variables whose conditions have been determined are sample size with five conditions, namely 200, 500, 1000, 2000, and 4000 persons: the test length with three conditions, namely 20, 40, and 60 items; and person-item targeting with five conditions, namely: -2, -1, 0, 1, and 2 logit. The results of the combination of the three variables obtained as many as 75 conditions (5 × 3 × 5).

Regarding person-item targeting, there are two things to consider. First, the distribution of items was made constant, namely normally distributed: mean = 0 and SD = 1 (Bastari, 2000; Setiadi, 1997); second, for the distribution of persons, the normal distribution was generated by SD = 1 (Bastari, 2000), but the mean level of person's abilities varied from 0 (on-target) and -2, -1, 1, and 2 (off-target). The simulated data produces dichotomous responses (0 and 1) Rasch model.

Table 1. Distribution for generating 5 conditions person – item targeting

Targeting	Distribution Item	Distribution Person
2		Normal (mean = 2, SD = 1)
1		Normal (mean = 1, SD = 1)
0	Normal (mean = 0, SD = 1)	Normal (mean = 0, SD = 1)
-1		Normal (mean = -1, SD = 1)
-2		Normal (mean = -2, SD = 1)

Results and Discussion

Results

This study produces data on 75 conditions that have been analyzed to produce an estimate of the average separation reliability (ISR and PSR). For ISR, 75 conditions (100%) resulted in a good ISR (> 0.90). Meanwhile, for PSR, out of 75 conditions, there were 37 conditions (49%) that resulted in a good PSR estimate or met the criteria (> 0.80) and 38 conditions (51%) had PSR that did not meet the criteria (< 0.80). Table 2 shows the results of the average separation reliability estimation (ISR and PSR).

Table 2. The results of the average separation reliability estimation (ISR dan PSR)

No	Sample	Test Length	Targeting	*Separation Reliability*	
				PSR	ISR
1	200	20	2	0.43	0.93
2			1	0.75	0.97
3			0	0.77	0.97
4			-1	0.67	0.96
5			-2	0.53	0.94

6			2	0.39	0.97
7			1	0.73	0.99
8	500	20	0	0.78	0.99
9			-1	0.71	0.99
10			-2	0.52	0.97
11			2	0.41	0.98
12			1	0.73	0.99
13	1000	20	0	0.78	0.99
14			-1	0.73	0.99
15			-2	0.46	0.99
16			2	0.58	1.00
17			1	0.73	1.00
18	2000	20	0	0.78	1.00
19			-1	0.76	1.00
20			-2	0.37	0.99
21			2	0.26	1.00
22			1	0.68	1.00
23	4000	20	0	0.77	1.00
24			-1	0.71	1.00
25			-2	0.50	1.00
26			2	0.74	0.94
27			1	0.85	0.96
28	200	40	0	0.87	0.98
29			-1	0.86	0.96
30			-2	0.72	0.94
31			2	0.71	0.98
32			1	0.83	0.98
33	500	40	0	0.87	0.99
34			-1	0.84	0.99
35			-2	0.72	0.97
36			2	0.72	0.99
37			1	0.86	0.99
38	1000	40	0	0.88	0.99
39			-1	0.84	0.99
40			-2	0.72	0.98
41			2	0.72	0.99
42			1	0.84	1.00
43	2000	40	0	0.88	1.00
44			-1	0.84	1.00
45			-2	0.75	0.99
46			2	0.73	1.00
47			1	0.85	1.00
48	4000	40	0	0.87	1.00
49			-1	0.86	1.00
50			-2	0.77	1.00

51			2	0.82	0.94
52			1	0.88	0.97
53	200	60	0	0.92	0.96
54			-1	0.89	0.96
55			-2	0.83	0.95
56			2	0.83	0.98
57			1	0.90	0.99
58	500	60	0	0.91	0.99
59			-1	0.89	0.98
60			-2	0.82	0.98
61			2	0.81	0.99
62			1	0.89	0.99
63	1000	60	0	0.91	0.99
64			-1	0.89	0.99
65			-2	0.79	0.99
66			2	0.81	0.99
67			1	0.90	1.00
68	2000	60	0	0.91	1.00
69			-1	0.89	1.00
70			-2	0.79	0.99
71			2	0.82	1.00
72			1	0.90	1.00
73	4000	60	0	0.91	1.00
74			-1	090	1.00
75			-2	0.79	1.00

Item Separation Reliability (ISR)

ISR indicates how well the items in the test used can be separated by the person who takes the test (Wright & Stone, 1999). In this study, from 75 conditions it was found that all conditions resulted in an ISR estimate that met the criteria. It means that the test length with 20 items, 40 items, and 60 items using several sample sizes, namely 200, 500, 1000, 2000, and 4000 persons in all targeting conditions, namely -2, -1, 0, 1, and 2 resulted in ISR estimation is good or meets the criteria (> 0.90).

The ISR estimation results obtained move from a range of values of 0.93 to 1.00. This means that the items in the test used can be separated well by the person who takes the test. In addition, neither the constant sample size nor the constant test length had an effect on the high and low ISR estimates and did not have a consistent pattern on the ISR estimates generated in each targeting condition (see Table 2).

Person Separation Reliability (PSR)

PSR is an estimate of how well the test used can distinguish the level of the respondent's trait (Wright & Masters, 1982). In this study, it was found that from 75 conditions, the estimated PSR that was good or met the criteria was 37 conditions (49%), while the estimated PSR that did not meet the criteria was 38 conditions (51%). In this study, the estimation of PSR is seen from the number of samples and the length of the test in producing PSR estimates for each target.

Based on Figure 1, it shows in general that in the sample condition of 200, the longer the test length, the estimated PSR value also increases in all targeting conditions. However, not all of the targeting conditions resulted in a good PSR estimate. When a sample size of 200 with a test length of 20 items does not produce a good estimate of the PSR value in all targeting conditions. For example, when the targeting condition is 0 (on-target), the estimated PSR value obtained is only 0.77. (< 0.80). For the 40-item test length, the targeting conditions that resulted in a good PSR estimate were targeting 0, -1, and 1 (> 0.80).

Meanwhile, for the 60-item test length, all targeting conditions resulted in a good PSR value estimate (> 0.80).

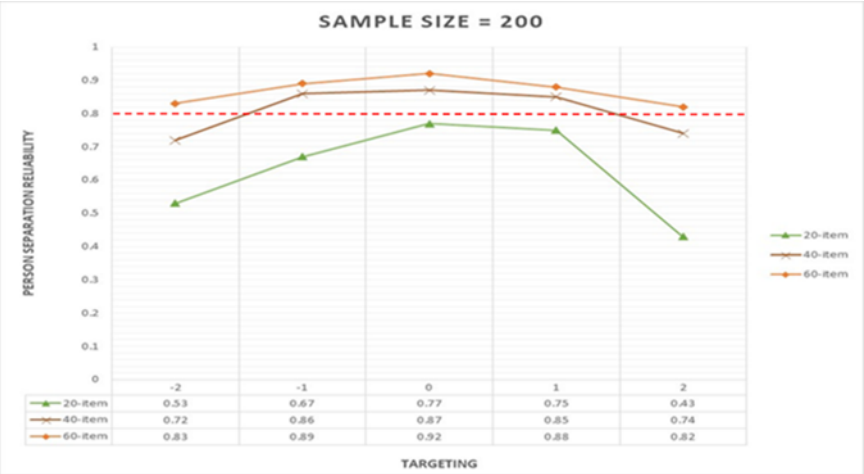


Figure 1. Sample size 200

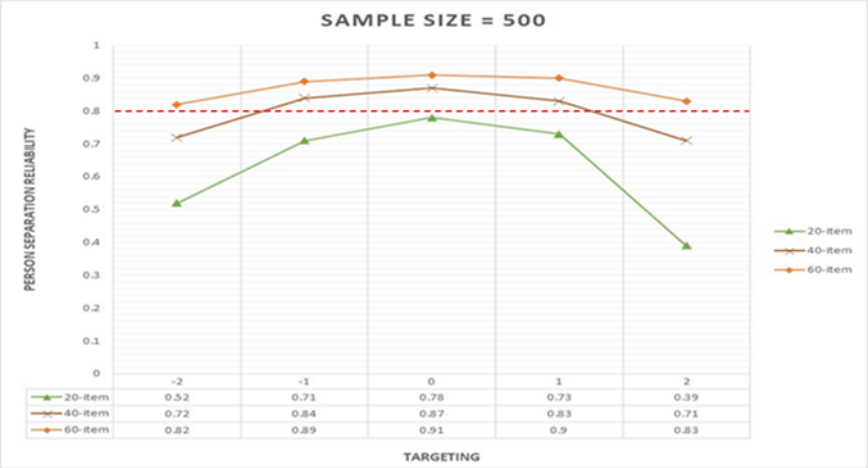


Figure 2. Sample size 500

Figure 2 shows that in the sample condition of 500, the longer the test length, the PSR estimation also increases in all targeting conditions. However, not all conditions produce a good PSR estimate. For example, for a test length of 20 items, there is no targeting condition that results in a good PSR estimate. The maximum PSR estimate obtained when targeting 0 is 0.78 (< 0.80). The results were different when the test length was 40 items, which resulted in good PSR estimates for the three targeting conditions, namely 0, -1, and 1 (> 0.80). On the length of the test with 60 items, the estimated PSR produced is good in all targeting conditions (PSR value > 0.80). In this case, targeting is in the range of -2 to 2 logit.

Based on Figure 3, for a sample size of 1000, it shows that the longer the test length, the PSR estimation also increases in all targeting conditions. However, only a few conditions resulted in a good PSR estimate (> 0.80). For all targeting conditions with a test length of 20 items, it was found that there was no good PSR estimate (the resulting PSR value was < 0.80). The maximum PSR estimate is 0.78 in the targeting 0 conditions (on-target), while the lowest PSR estimate is obtained at 0.41, namely when the targeting condition is 2 or when the average person's ability is above the average test difficulty of 2 logits. For the 40-item test length, the targeting conditions that produce a good PSR estimate (> 0.80) are in three targeting conditions, namely; -1 (0.84), 0 (0.88), and 1 (0.86).

Figure 4 with the 2000 sample condition generally shows that the longer the test length, the more PSR estimation also increases in all targeting conditions. However, not all targeting conditions produce good PSR estimates. It can be seen that for the 20-item test length, the estimated PSR produced is not good (< 0.80) in all targeting conditions. The highest PSR estimation was obtained when the targeting condition was 0, which was 0.78. The length of the test is 40 items, resulting in PSR estimates that meet the criteria for three targeting conditions, namely -1 (0.84), 0 (0.88), and 1 (0.84). Meanwhile, when the length of the

test is 60 items, a good PSR estimate is produced in the four targeting conditions. The PSR estimates that meet the criteria are for targeting 0 (0.91), -1 (0.89), 1 (0.90), and 2 (0.81).

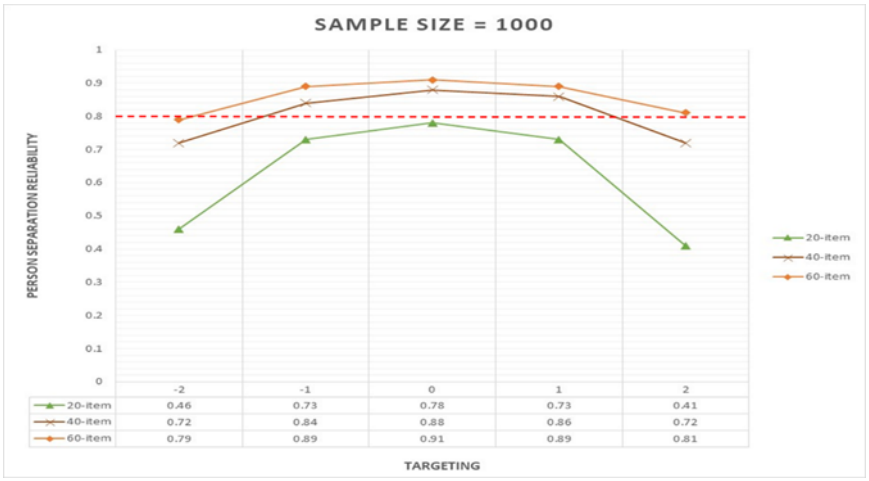


Figure 3. Sample size 1000

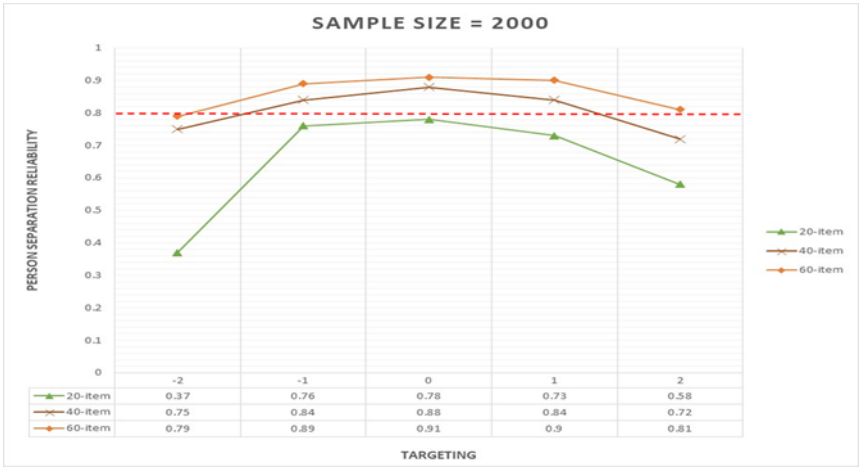


Figure 4. Sample size 2000

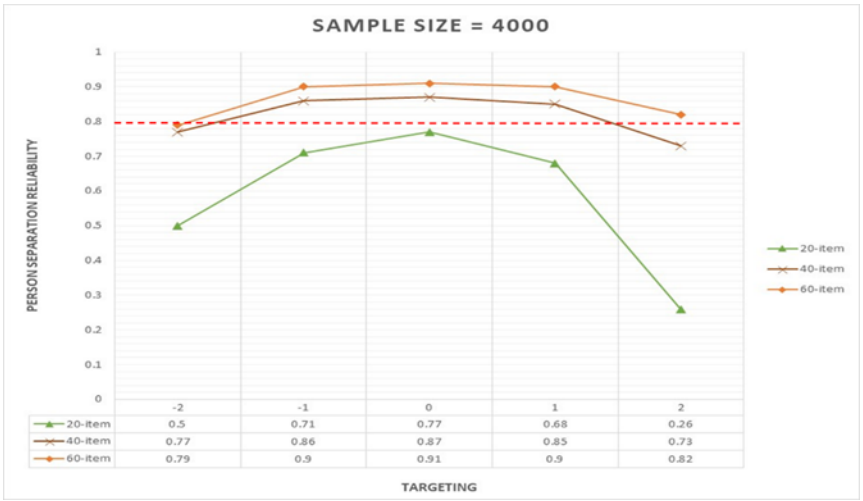
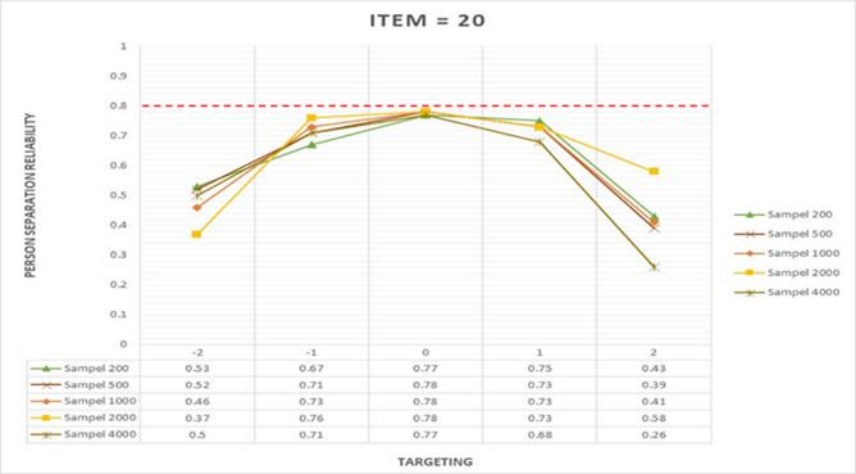


Figure 5. Sample size 4000

Based on figure 5, it shows that in the 4000 sample condition, the longer the test length, the PSR estimation also increases in all targeting conditions. However, not all conditions produce a good PSR

estimate. The targeting condition that produces a good PSR estimate is the 0-item test length in three targeting conditions, namely 0 (0.87), -1 (0.86), and 1 (0.85). Meanwhile, for the targeting condition with a test length of 60 items, a good PSR estimate is found in four targeting conditions, namely: -1 (0.90), 0 (0.91); 1 (0.90) and 2 (0.82).

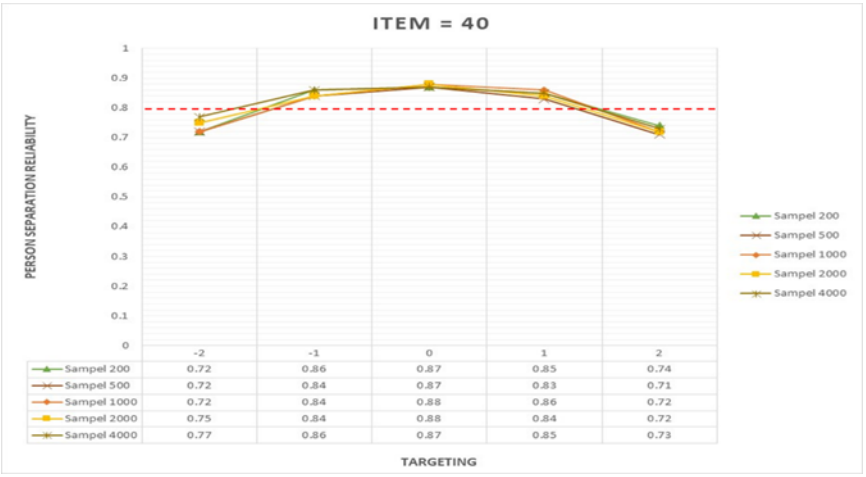
PSR estimation is seen from the constant test length:



	-2	-1	0	1	2
Sampel 200	0.53	0.67	0.77	0.75	0.43
Sampel 500	0.52	0.71	0.78	0.73	0.39
Sampel 1000	0.46	0.73	0.78	0.73	0.41
Sampel 2000	0.37	0.76	0.78	0.73	0.58
Sampel 4000	0.5	0.71	0.77	0.68	0.26

Figure 6. Test length 20 item

Based on Figure 6, the results of the study using a test length of 20 items show that increasing sample size is not followed by increasing PSR estimates in all targeting conditions. This means that a large sample size does not always result in a high PSR estimate compared to a smaller sample size. Another finding shows that sample size has a consistent pattern in generating PSR estimates. This can be seen in the lines that are in the graph crossing each other.



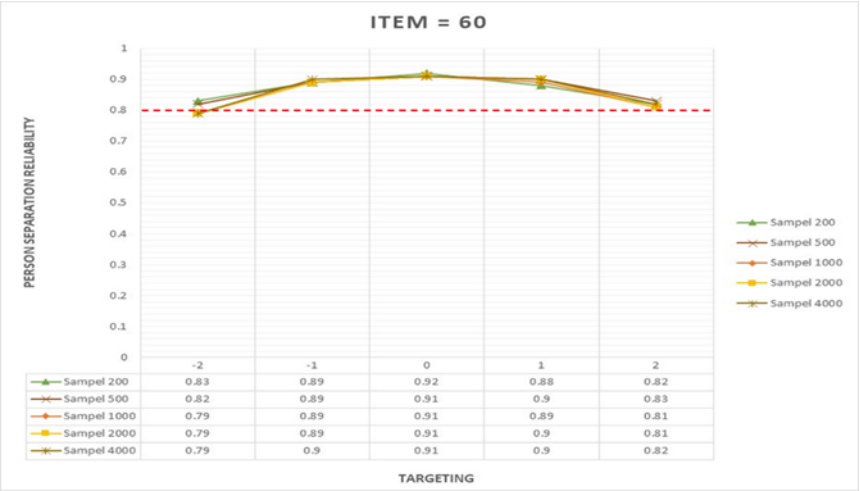
	-2	-1	0	1	2
Sampel 200	0.72	0.86	0.87	0.85	0.74
Sampel 500	0.72	0.84	0.87	0.83	0.71
Sampel 1000	0.72	0.84	0.88	0.86	0.72
Sampel 2000	0.75	0.84	0.88	0.84	0.72
Sampel 4000	0.77	0.86	0.87	0.85	0.73

Figure 7. Test Length 40 item

Based on Figure 7, the estimated PSR value generated with a test length of 40 items shows that increasing sample size is not followed by increasing PSR estimates in all targeting conditions. In addition, it was also found that the sample size did not have a consistent pattern in each targeting condition in producing PSR estimates. This is indicated by the lines in the graph crossing each other. However, a good PSR estimate (> 0.80) was obtained for all sample sizes with targeting conditions in the range of -1 to 1 logit.

Based on Figure 8, the length of the test with 60 items shows that as the sample size increases, the estimated PSR in all targeting conditions does not increase. In addition, it was also found that the sample size did not have a consistent pattern in producing PSR estimates. This can be seen in a graph where the lines intersect and even overlap at the same point. However, overall with a test length of 60 items, it was

able to produce good PSR estimates in all targeting conditions except for sample sizes of 1000, 2000, and 4000 in targeting conditions of -2, resulting in an estimated PSR of 0.79 (< 0.80).



	-2	-1	0	1	2
Sampel 200	0.83	0.89	0.92	0.88	0.82
Sampel 500	0.82	0.89	0.91	0.9	0.83
Sampel 1000	0.79	0.89	0.91	0.89	0.81
Sampel 2000	0.79	0.89	0.91	0.9	0.81
Sampel 4000	0.79	0.9	0.91	0.9	0.82

Figure 8. Test length 60 item

Discussion

The study was conducted to know the influence of sample size, test length, and person-item targeting on the high level of low person and item separation reliability (PSR and ISR) in the Rasch model, using simulation studies that resulted in 75 conditions (5 sample size × 3 test length × 5 person-item targeting). Based on the results of data analysis it was found that the size of the sample did not affect the elevation of the low separation reliability (ISR dan PSR). This means that a larger sample size will not immediately result in a higher PSR estimate compared to a smaller sample. This finding is consistent with the findings of previous studies that found that the sample size does not affect reliability estimates if the data fits the Rasch model, where the item discrimination for each item was fixed at 1. Whereas in the context of Cronbach's alpha, the highest reliability is affected by the lowest discriminatory power. In sum, the sample size does not have a direct impact on PSR and ISR (Bonett, 2002; Linacre, 2018).

In addition, it was found that the sample size of 200 respondents was able to produce a good estimate of separation reliability when the on-target test was conducted. This finding is in line with the opinion expressed by Wright (1977) which states that the minimum sample size in the Rasch model is 200 persons in order to produce a reliable and optimal standard error. This opinion is reinforced by the findings of another study which found that sample sizes below 200 would bias the results of parameter estimates in the Rasch model and produce unreliable estimates (Chen et al., 2013; Linacre, 1994; O'Neill et al., 2020). Thus, researchers can use the minimum sample size criteria of 200 persons as a rule of thumb if the sampling design used is non-probability as required in the Journal Article Reporting Standards (JARS) quantitative research (Appelbaum et al., 2018).

In addition to findings about sample size, other findings relate to test length. It was found that the length of the test affects the level of reliability estimation (Kopalle & Lehmann, 1997). In the Rasch Model, the recommended minimum test length is 20 items (Wright, 1977), but in this study, the 20-item test length has not been able to produce a good PSR estimate even though it has exceeded the sufficient category. This is influenced by the condition of the criteria used, which is 0.80 (Linacre, 2018) which is higher than Tennant and Conaghan (2007) which is 0.70.

Last but not least, the finding of this study is expected to be used as a reference for researchers or practitioners in determining sample size and test length in using the Rasch model. Researchers and practitioners must also pay attention to or know the level of ability of the test takers who will be given the test so that researchers or practitioners can design an on-target test so that the resulting PSR estimation will be good in sorting test takers. In addition, in designing a test, because PSR estimation is affected by the test length (many items) in sorting test takers, researchers, and practitioners it is also very necessary to pay attention to the availability of time for all test takers in general during the test. Further studies are also expected researchers add conditions to explore other factors that affect the high and low separation reliability (ISR and PSR). The limitations of this study are the number of replications for each condition

of the research data which only uses 50 replications. Researchers in the future can increase the number of replications to 1000 times like research conducted by Putra et al., (2017).

Conclusion

Based on the research results and discussion, it can be concluded that person separation reliability (PSR) is influenced by test length and person-item targeting. Meanwhile, the sample size does not affect the estimated PSR value. The longer the test (many items), the higher the estimated PSR value. Likewise, with person-item targeting, the more on-target the estimated PSR value is the higher. Even under the same conditions (sample size and test length), the factor that most influences the estimated PSR value is person-item targeting.

Meanwhile, for the ISR, neither sample size, test length nor person-item targeting had any effect on the high or low estimated value of ISR. However, the recommended test length in the Rasch model to produce good separation reliability (ISR and PSR) is a minimum of 20 items. In addition, the recommended sample size is a minimum of 200 people.

Conflict of Interest

The authors declare that there are no conflicts of interest with respect to the authorship of this paper.

Authors Contribution

RSB contributed to conception, methodology, software, formal analysis, investigation, writing original draft preparation, writing-review and editing. S contributed to conception, methodology, investigation, writing-review and editing. All authors have read and agreed to the published version of the manuscript.

References

Andrich, D. (2011). Rating scales and Rasch measurement. *Expert Review Pharmacoeconomics Outcomes Research, 11*(5), 571-585.

Andrich, D., & Marais, I. (2019). *A course in Rasch measurement theory: Measurement in the educational, social and health science*. Springer Nature Singapore.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (1999). *Standards for Educational and Psychological Testing*. American Educational Research Association.

Appelbaum, M., Cooper, H., Kline, R. B., Mayo-Wilson, E., Nezu, A. M., & Rao, S. M. (2018). Journal article reporting standards for quantitative research in psychology: The APA Publications and Communications Board task force report. *American Psychologist, 73*(1), 3–25. https://doi.org/10.1037/amp0000191

Bastari, B. (2000). *Linking multiple-choice and constructed-response items to a common proficiency scale*. Unpublished doctoral dissertation. University of Massachusetts Amherst.

Bonett, D. G. (2002). Sample size requirements for testing and estimating coefficient alpha. *Journal of Educational and Behavioral Statistics, 27*(4), 335–340. https://doi.org/10.3102/10769986027004335

Casey, T. M., & Harden, J. J. (2014). *Monte carlo simulation and resampling methods for social science*. SAGE Publications, Inc.

Chen, W-H., Lenderking, W., Jin, Y., Wyrwich, K. W., Gelhorn, H., Revicki, D. A. (2013). Is Rasch model analysis applicable in small sample size pilot studies for assesing item characteristics? an

example using promis pain behavior item bank data. *Quality Life Research*. https://doi.org/10.1007/s11136-013-0487-5

de Ayala, R. J. (2009). *The theory and practice of item response theory*. Guilford Press.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates, Inc.

Feinberg, R. A., & Rubright, D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice, 35*(2), 36-49.

Guyon, H., Kop, J-L., Juhel, J., & Falissard, B. (2018). Measurement, ontology, and epistemology: Psychology needs pragmatism-realism. *Theory & Psychology*, *28*(2), 149–171.

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. SAGE Publication, Inc.

Hayat, B. (1995). *Pengantar model Rasch*. Kemendikbud Publishing.

Hayat, B., Putra, M. D. K., Suryadi, B. (2020). Comparing item parameter estimates and fit statistics of the Rasch model from three different traditions. *Jurnal Penelitian dan Evaluasi Pendidikan, 24*(1), 39-50. https://doi.org/10.21831/pep.v24i1.29871

Karabatsos, G. (2000). A critique of Rasch residual fit statistics. *Journal of Applied Measurement*,*1*(2), 152–176.

Kopalle, P. K., & Lehmann, D. R. (1997). Alpha inflation? The impact of eliminating scale items on Cronbach's alpha. *Organizational Behavior and Human Decision Processes, 70*(3), 189–197. https://doi.org/10.1006/obhd.1997.2702

Linacre, J. M. (1994). https://www.rasch.org/rmt/rmt74m.htm

Linacre, J. M. (2018). *Winsteps® Rasch measurement computer program user's guide*. Winsteps.com.

O'Neill, T. R., Gregg, J. L., & Peabody, M. R. (2020). Effect of sample size on common item equating using the dichotomous Rasch model. *Applied Measurement in Education, 33*(1), 1-23. https://doi.org/10.1080/08957347.2019.1674309

Putra, M. D. K., Umar, J., Hayat, B., & Utomo, A. P. (2017). Pengaruh ukuran sampel, dan intraclass correlation coefficients (ICC) terhadap bias estimasi parameter multilevel latent variable modeling: Studi dengan simulasi monte carlo. *Jurnal Penelitian dan Evaluasi Pendidikan*, *21*(1), 34-50.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Raykov, T., & Marcoulides, G. A. (2019). Thanks coefficient alpha, we still need you! *Educational and Psychological Measurement, 79*(1), 200–210. https://doi.org/10.1177/0013164417725127

Setiadi, H. (1997). Small sample IRT item parameter estimates. (Unpublished doctoral dissertation). University of Massachusetts Amherst.

Suryadi, B., Hayat, B., & Putra, M. D. K. (2020). Evaluating psychometric properties of the Muslim Daily Religiosity Assessment Scale (MUDRAS) in Indonesia samples using the Rasch model. *Mental Health, Religion & Culture*. https://doi.org/10.1080/13674676.2020.1795822

Tennant, A., & Conaghan, P. G. (2007). The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? *Arthritis Care & Research*, *57* (8), 1358-1362. https://doi.org.10.1002/art.23108

Wright, B. D. (1977). Solving measurement problems with the Rasch model. *Journal of Educational Measurement, 14*(2), 97–116. https://doi.org/10.1111/j.1745-3984.1977.tb00031.x

Wright, B. D., & Douglas, G. A. (1977). Best procedures for sample-free item analysis. *Applied Psychological Measurement, 1*(2), 281–295. https://doi.org/10.1177/014662167700100216

Wright, B. D., & Stone, M. H. (1979). *Best test design*. MESA Press.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. MESA Press.

Wright, B. D., & Stone, M. (1999). *Measurement essentials* (2nd ed.). Wide Range, Inc.