Anima Indonesian Psychological Journal
2017, Vol. 32, No. 2, 99-108
http://dx.doi.org/10.24123/aipj.v32i2.588
0215-0158

Comparing *t* Test, r_{it} Significance Test, and r_{it} Criteria for Item Selection Method: A Simulation Study

Agung Santoso
Universitas Sanata Dharma

Three criteria of items selection have been widely used despite of their limitations without any empirical evidence to support its practice. Current study examined the three criteria to determine which of the three criteria were the best among the others. Those criteria were the item total correlation, its significance by *t*-test and significance of r_{it}. Simulations were conducted to demonstrate which of the three criteria provided the least errors in both excluding good items and including bad items in the scale. The author manipulate four conditions in conducting simulation study: (a) number of items in a scale; (b) value of r_{it} in population; (c) sample sizes; and (d) criteria in including or excluding items in a scale. The results showed that criteria of $r_{it} \geq .30$ provided the least errors of including bad items and excluding good items, particularly when $n \geq 200$. The two criteria based on significance test provided the largest errors therefore were not recommended in future practice.

Keywords: item-total correlation, item discrimination, item selection

Tiga kriteria seleksi butir telah banyak digunakan meskipun memiliki kelemahan-kelemahan. Penggunaan tiga kriteria ini juga tidak didukung oleh temuan empiris mengenai performansi ketiganya dalam melakukan seleksi butir. Penelitian ini dilakukan untuk menentukan kriteria yang memberikan hasil seleksi butir terbaik. Ketiga kriteria tersebut adalah signifikansi uji *t*, signifikansi nilai r_{it}, dan nilai r_{it}. Simulasi dilakukan untuk menunjukkan kriteria yang menyebabkan kesalahan terkecil dalam menggugurkan butir yang baik dan mempertahankan butir yang buruk. Peneliti memanipulasi empat kondisi dalam simulasi ini: (a) jumlah butir dalam skala; (b) nilai r_{it} di populasi; (c) besarnya sampel; dan (d) kriteria dalam mengugurkan atau mempertahankan butir. Hasil simulasi menunjukkan bahwa kriteria $r_{it} \geq .30$ menghasilkan kesalahan terkecil dalam mempertahankan atau menggugurkan butir, khususnya ketika $n \geq 200$. Dua kriteria lain yang didasarkan pada uji signifikansi menghasilkan kesalahan terbesar sehingga tidak disarankan untuk digunakan.

Kata kunci: korelasi butir-total, daya diskriminasi butir, seleksi butir

Measurement quality is an important aspect in data collection either for research or diagnostic purposes. Measurement quality determines the credibility of the conclusion from the research or diagnosis. Therefore, good measurement quality is a priority in the data collection. One of the determinants of measurement quality is the quality of items in the instrument. If the measurement consists of high-quality items, the measurement tends to generate credible data.

One of the criteria of item quality frequently used is item discrimination power. Items with high discrimination power can distinguish subjects with high scores from subjects with low scores on the specific attribute measured (Anastasi & Urbina, 1997; G. Domino & Domino, 2006). High discrimination power items are used for measurement, while low discrimination power items are discarded.

There are at least three methods commonly used for item selection. The first is the significance of the t-test conducted on the item score (G. Domino & Domino, 2006; Edwards, 1957). In this method participants are divided into two groups, namely, high score and low score participants. High score group consists of the highest 25% of the participants in the try out, while low score group consists of the lowest 25%. When the *t*-test of an item shows significant difference, the item is considered as having a high discrimination power. However, this method has two weaknesses. Firstly, it dichotomizes total score into

Correspondence concerning this article should be addressed to Agung Santoso, Universitas Sanata Dharma Paingan, Maguwoharjo, Depok, Sleman, Yogyakarta 55282. E-mail: agungsan_psy@yahoo.com

two discrete high-low groups (Maxwell & Delaney, 1993). Secondly, it halves the sample size. These weaknesses contribute to the decrease of statistical power of this analysis, resulting in the low sensitivity in identifying items with moderate discrimination power.

The second method is the significance of the corrected item-total correlation (r_{it}; Hadi, 2005). If the r_{it} is significant, the item is considered as having a high discrimination power. The weakness of this method lies in its ignorance of the correlation coefficient of the r_{it} when including or excluding items. An item with very low r_{it} can be significant when the sample size is large enough. For instance, $r_{it} = .14$ can be significant when the $n \geq 20$. Such tendency to keep items with low correlation coefficient is caused by the significant test that uses $r_{it} = 0$ as its null hypothesis. A significant result only means there is enough evidence that the correlation in the population is larger than zero.

The r_{it} coefficient is the third method frequently used for item selection (Azwar, 2013; G. Domino & Domino, 2006; Kline, 2005). The minimum $r_{it} = .30$ or .25 is generally used to keep an item in the instrument (Azwar). An item is discarded when its r_{it} is smaller than the minimum criteria. This method has weakness, that is, an ignorance of the variation of r_{it} between samples, particularly when the sample size is small. Small sample size makes the estimation of r_{it} between samples high, or in other words, the precision is low. Low precision increases the possibility to generate moderate r_{it} from the population with $r_{it} = 0$. This means the findings from the sample are less accurate in representing the population, while the expectation is that the sample may represent the population.

Considering weaknesses of these three methods, it is important to study which method is the best for item selection and what conditions may generate most optimal item selection process. Item selection is considered best when it generates the least errors in excluding and including items. Optimal condition in this study is limited to sample size, because it is the only sample measurement that can be controlled by the researcher.

Method

A simulation study was employed to examine these three item selection methods. It focused on the number of errors in each method, either in excluding items that should be included, or including items that should be excluded. The method that generates least errors is considered as the best.

Several conditions were determined by the researcher in creating simulation data using the R Program (see Appendix A). The first condition is the number of items. The researcher decided to use only one variation, that is, 40 items with high corrected item-total correlation (r_{it})[1] in the population (Group 1) and 10 items with low r_{it} in the population (Group 2). This number of items was chosen based on the number of items commonly found in psychological research in Indonesia according to the researcher's experience.

The r_{it} in the population, which is the second condition of the simulation, was calculated by first determining the correlation between items (r_{ii}). This was easier compared to determining r_{it} before r_{ii}. Two variations were applied in this condition.

(1) Correlation between items in Group 1 (40 items) and correlation between items in Group 2 (10 items) were determined at .30. Correlation between items in Group 1 and 2 was determined 0.0 (see Figure 1). This implies that low discrimination items in the instrument actually measured another construct which was not meant to be measured. In this condition, r_{it} for all 40 items was .513, while r_{it} for the rest 10 items was .116 in the population. (The formula to calculate r_{it} in this simulation is described in Appendix B, while R Program codes used is in Appendix A).

(2) Correlation between items in Group 1 was determined .30 while the r_{ii} in Group 2 was 0.0. Correlation between items in Group 1 and 2 was determined 0.0 (see Figure 2). This means that there was only one dominant factor in the instrument. In this condition, the r_{it} for the 40 items was .527, while r_{it} for the rest 10 items was 0.0 in the population.

Each sample had a sample size from 50, which represented small sample, to 500, which represented large sample size, with an interval of 50. These sample sizes were the third condition determined by the researcher. These variations were chosen to represent sample size in contemporary psychological research in Indonesia.

The researcher used 1000 samples for each sample size in each r_{it} condition. For each sample, the researcher calculated r_{it} for each item, and then selected items using the three methods; (a) r_{it} significance; (b) *t*-test significance; and (c) r_{it} coefficient. Four r_{it} were determined as the minimum inclusion

[1] To make it easier to read, in this article the symbol r_{it} represents corrected item-total correlation.

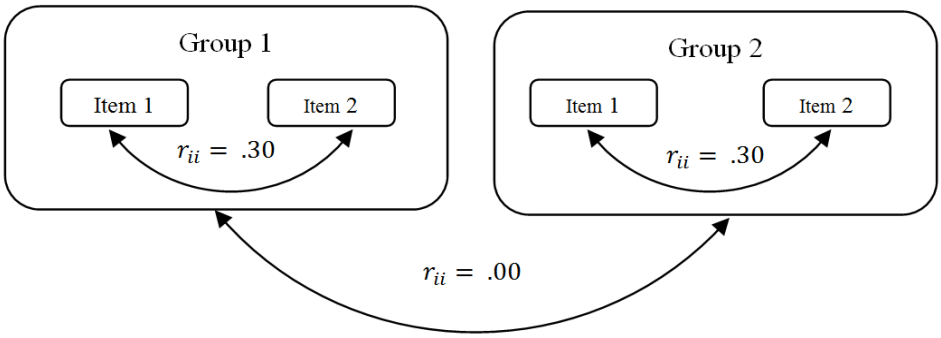

Figure 1. Illustration of the correlation between items in each group and correlation between items in Group 1 and 2 in the first condition.

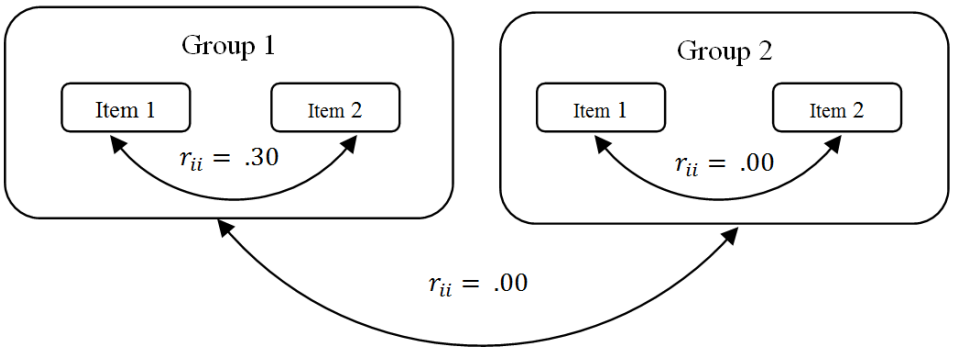

Figure 2. Illustration of the correlation between items in each group and correlation between items in Group 1 and 2 in the second condition.

criteria: (a) $r_{it} \geq .20$; (b) $r_{it} \geq .25$; (c) $r_{it} \geq .30$; and (d) $r_{it} \geq .40$. The number of items excluded from Group 1 (the error of excluding good items) and included from Group 2 (the error of including bad items) was documented. The 1000 samples were tabulated to summarize the number of errors in each method.

The tabulation is presented in tables and graphs. For succinctness, only four sample sizes were displayed in tables as they represented other sample sizes. These four sample sizes were 50, 100, 250, and 500. Results from other sample sizes are presented in graphs.

Based on this simulation the researcher evaluated which method is the best. The best method is the one that makes least errors, both in excluding items that should be included, or including items that should be excluded in all sample sizes.

Results

The result of the simulation is displayed in Table 1 and 2, and Figure 3 to 6. When correlation between poor items was determined at .30, or poor items were considered measuring another construct, item selection based on significance test, that is r and t significance test, tended to include items regardless of whether or not those items had high r_{it} in the population (see Table 1). This tendency became stronger when the sample sizes increased. This is because the decrease of the standard error of the r and t when the sample size increased. It means that r and t significance test tend to include items that should be excluded. Such tendency was stronger when using r significance test as compared to t significance test, because the power of r significance test in stronger than t-test. One reason for this power difference was the use of 25% highest and lowest in the t-test.

In the same condition, item selection method that was based on correlation coefficient had a tendency to exclude items regardless of whether or not has high r_{it} in the population (see Table 1). This tendency was weaker as sample size increased. It was because the r_{it} estimation getting more stable when the sample size increased, so that the r_{it} on the sample was getting closer to the population. When the sample size reached 500, errors in excluding and including

Table 1
Result of the Simulation with r_{ii} Between Poor Items was .30

Number of Errors	Exclusion Errors						Inclusion Errors					
	r_{it} significance	*t*-test	*r* = .20	*r* = .25	*r* = .30	*r* = .40	r_{it} significance	*t*-test	*r* = .20	*r* = .25	*r* = .30	*r* = .40
						n = 50						
0	700	0	809	660	412	45	287	713	159	335	540	852
1	203	0	146	206	258	89	209	179	198	233	235	113
2	50	1	25	77	135	109	183	70	166	164	109	26
3	25	1	14	23	72	110	118	20	151	114	63	5
4	10	2	5	13	44	101	91	10	110	77	34	3
5	3	1	0	7	28	99	55	2	86	41	11	1
6	3	5	1	6	16	66	32	3	63	24	6	0
7	3	11	0	4	15	68	15	3	38	5	1	0
8	2	12	0	1	2	61	5	0	16	5	1	0
9	1	8	0	1	4	46	4	0	11	2	0	0
≥ 10	0	954	0	0	10	165	0	0	0	0	0	0
						n = 100						
0	997	20	994	966	838	200	117	600	254	529	778	991
1	3	32	6	30	125	211	163	224	222	238	154	9
2	0	40	0	3	27	172	150	97	166	119	48	0
3	0	54	0	0	8	101	148	41	138	59	18	0
4	0	67	0	1	0	86	123	19	81	31	1	0
5	0	64	0	0	1	62	106	10	67	14	0	0
6	0	60	0	0	1	51	71	5	36	9	1	0
7	0	53	0	0	0	41	62	2	21	1	0	0
8	0	61	0	0	0	32	38	2	10	0	0	0
9	0	59	0	0	0	7	14	0	5	0	0	0
≥ 10	0	429	0	0	0	28	0	0	0	0	0	0
						n = 250						
0	1000	837	1000	1000	998	711	24	332	532	882	992	1000
1	0	126	0	0	2	186	48	258	229	96	8	0
2	0	29	0	0	0	55	70	175	113	16	0	0
3	0	6	0	0	0	27	77	99	68	6	0	0
4	0	1	0	0	0	12	113	61	28	0	0	0
5	0	0	0	0	0	3	123	37	13	0	0	0
6	0	0	0	0	0	4	113	23	14	0	0	0
7	0	1	0	0	0	1	139	12	3	0	0	0
8	0	0	0	0	0	1	115	2	0	0	0	0
9	0	0	0	0	0	0	112	0	0	0	0	0
≥ 10	0	0	0	0	0	0	0	0	0	0	0	0
						n = 500						
0	1000	999	1000	1000	1000	974	0	130	790	991	1000	1000
1	0	1	0	0	0	25	1	142	169	8	0	0
2	0	0	0	0	0	1	8	146	24	1	0	0
3	0	0	0	0	0	0	25	137	13	0	0	0
4	0	0	0	0	0	0	18	132	2	0	0	0
5	0	0	0	0	0	0	52	121	0	0	0	0
6	0	0	0	0	0	0	54	77	0	0	0	0
7	0	0	0	0	0	0	124	53	1	0	0	0
8	0	0	0	0	0	0	173	33	1	0	0	0
9	0	63	0	0	0	10	0	0	0	0	0	0
≥ 10	0	322	0	0	0	6	0	0	0	0	0	0

items become very insignificant. It means that this method tended to generate items similar to population when poor items actually measured another construct. The comparison between four criteria of the *r* showed that $r_{it} = .40$ tended to cause more errors in excluding good items. This tendency can be seen when the sample size reached 500; 2.6% of the sample excluded at least one good item. In contrast, the criterion $r_{it} = .20$ and $r_{it} = .25$ tended to cause more errors in including poor items. When sample size reached 500, 20.8% of the sample included at least one poor item using $r_{it} = .20$, and

Table 2

Result of the Simulation with r_{ii} Between Poor Items was .00

Number of Errors	Exclusion Errors						Inclusion Errors					
	r_{it} significance	*t*-test	*r* = .20	*r* = .25	*r* = .30	*r* = .40	r_{it} significance	*t*-test	*r* = .20	*r* = .25	*r* = .30	*r* = .40
					n = 50							
0	774	0	866	707	492	91	607	831	435	677	846	979
1	159	1	107	197	239	111	314	152	368	272	139	21
2	50	3	23	62	136	123	74	16	162	50	15	0
3	10	0	1	19	52	117	4	1	31	1	0	0
4	4	3	3	8	31	106	1	0	3	0	0	0
5	3	4	0	4	22	76	0	0	1	0	0	0
6	0	4	0	2	12	92	0	0	0	0	0	0
7	0	8	0	1	9	50	0	0	0	0	0	0
8	0	13	0	0	2	63	0	0	0	0	0	0
9	0	14	0	0	3	35	0	0	0	0	0	0
≥ 10	0	936	0	0	0	101	0	0	0	0	0	0
					n = 100							
0	1000	23	1000	984	895	276	592	821	800	953	987	1000
1	0	54	0	16	88	243	325	162	182	45	13	0
2	0	53	0	0	15	155	72	16	17	2	0	0
3	0	73	0	0	2	117	11	1	1	0	0	0
4	0	74	0	0	0	65	0	0	0	0	0	0
5	0	85	0	0	0	57	0	0	0	0	0	0
6	0	77	0	0	0	34	0	0	0	0	0	0
7	0	67	0	0	0	19	0	0	0	0	0	0
8	0	54	0	0	0	14	0	0	0	0	0	0
9	0	63	0	0	0	10	0	0	0	0	0	0
≥ 10	0	322	0	0	0	6	0	0	0	0	0	0
					n = 250							
0	1000	879	1000	1000	1000	830	603	727	993	1000	1000	1000
1	0	102	0	0	0	140	310	232	7	0	0	0
2	0	11	0	0	0	19	74	37	0	0	0	0
3	0	5	0	0	0	9	11	4	0	0	0	0
4	0	2	0	0	0	2	2	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
6	0	1	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
≥ 10	0	0	0	0	0	0	0	0	0	0	0	0
					n = 500							
0	1000	1000	1000	1000	1000	996	579	673	1000	1000	1000	1000
1	0	0	0	0	0	3	323	269	0	0	0	0
2	0	0	0	0	0	1	84	49	0	0	0	0
3	0	0	0	0	0	0	14	9	0	0	0	0
4	0	0	0	0	0	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	0	0	0	0	0	0	0
9	0	0	0	0	0	0	0	0	0	0	0	0
≥ 10	0	0	0	0	0	0	0	0	0	0	0	0

1% sample using $r_{it} = .25$. Criterion $r_{it} = .30$ had the smallest percentage of errors from all sample sizes compared to other methods (see Table 3).

Figure 3 and 4 provided more detailed information on the percentage of accuracy in including good items and excluding poor items when poor items measured another construct. Figure 3 showed that criterion $r_{it} = .20$ generated the highest accuracy in including good items compare to other criteria in all sample sizes. However, this criterion had a relatively low percentage of accuracy in excluding poor items in all sample sizes (see Figure 4). Of all item selection criteria, the most accurate criterion in including good items and excluding poor items is the criterion

$r_{it} \geq .30$.

In the condition where r_{ii} between poor items was 0.0 in the population – which means poor items did not measure any construct or are only measurement mistakes – item selection based on significance tests tended to include items that should be excluded. In all sample sizes, the percentage of errors in including at least one poor item using r_{it} significance is approximately 40% and using *t*-test approximately 20-30% (see Table 2). Such errors when using signi-

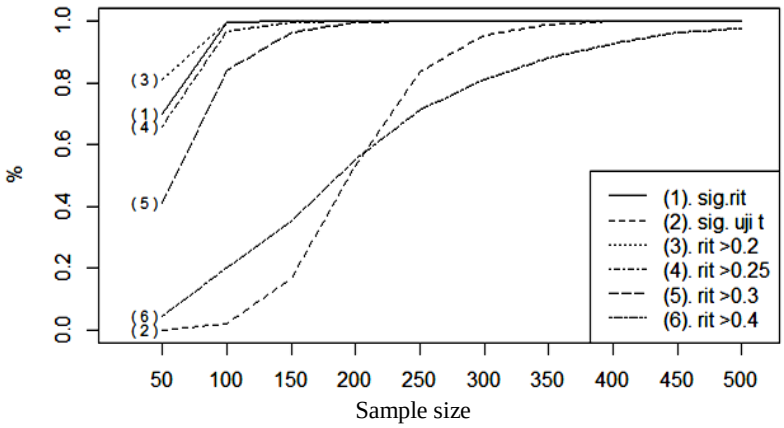

Figure 3. Percentage of the accuracy in including good items when r_{ii} between poor items was .30.

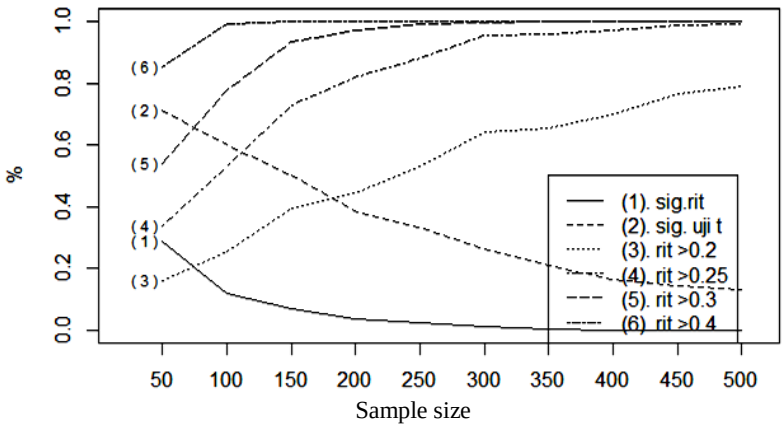

Figure 4. Percentage of the accuracy in excluding poor items when r_{ii} between poor items was .30.

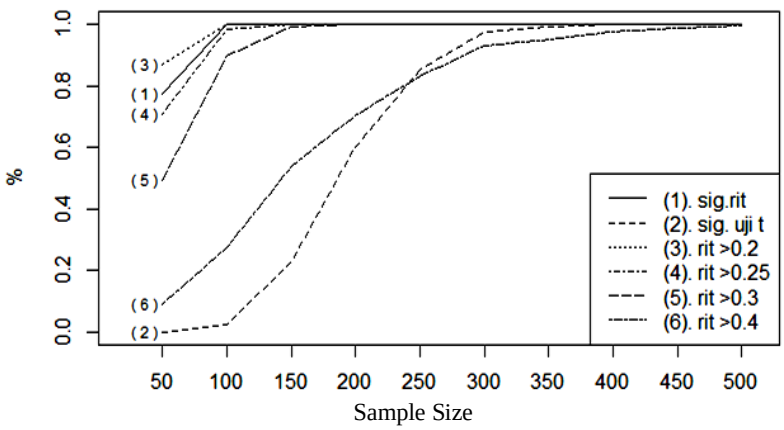

Figure 5. Percentage of accuracy in excluding poor items when r_{ii} between poor items was 0.0.

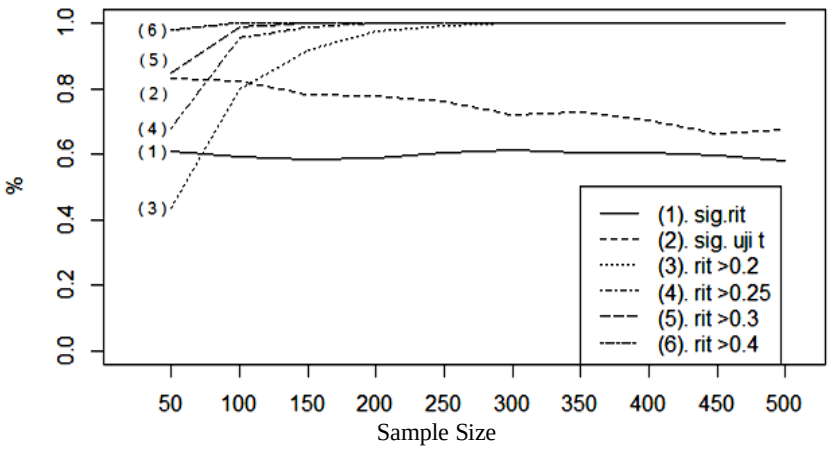

Figure 6. Percentage of accuracy in excluding poor items when r_{ii} between poor items was 0.3.

ficance tests tended to decrease when poor items did not measure any construct and increase when poor items measured another construct.

The percentage of errors in including at least one poor item or excluding at least one good item when using r_{it} tended to decrease with the increase of sample size. For instance, the percentage of errors in excluding at least one good item when *n = 50* was approximately 29.3% using the criterion $r_{it} \geq .25$. Using the same criterion, the percentage decreased to 1.6% when *n* = 100, and 0% when *n* = 250.

Of all four criteria using r_{it} method, criterion $r_{it} \geq .30$ had the lowest percentage of errors (see Table 3), while criterion $r_{it} \geq .25$ was the second lowest. When *n* = 250, both criteria did not generate any error from the 1000 sample replications (see Table 2). The other two criteria had relatively high percentage of errors but still lower than significance-based methods.

Figure 5 and 6 describe the percentage of accuracy in including good items and excluding poor items when the poor items did not measure any construct. It can be clearly seen that significance-based item selection methods tended to have lower accuracy in excluding items as compared to the other method, particularly in the larger sample sizes. Item selection method using $r_{it} \geq .25$ and $r_{it} \geq .30$ tended to generate less errors especially when $n \geq 200$.

Based on the results of the simulation above, it is concluded that the criteria $r_{it} \geq .30$ generates items that were the closest to the condition of the population. The other two significance-based methods were not satisfying because they did not accurately represent the population. Additionally, the simulation also offered an estimation about the sample size for item selection try out. When poor items were assumed to measure another construct, selection criteria $r_{it} \geq .30$ needed at least 200 samples to generate reasonably small errors. Other criteria needed larger sample size (i.e., more than 200) to generate comparably small errors. When poor items were assumed as purely measurement mistakes, selection criterion $r_{it} \geq .30$ needed at least 150 samples to produce small errors; and other criteria needed larger sample size.

Table 3
Percentage of the Total Errors from All Sample Sizes

Criteria	$r_{ii} = .00$	$r_{it} = .30$	Total	Rank*
r_{it} significance	.240	.460	.350	5
t-test significance	.315	.480	.398	6
$r_{it} \geq .20$	.076	.216	.146	3
$r_{it} \geq .25$	.044	.102	.073	2
$r_{it} \geq .30$	.041	.078	.060	1
$r_{it} \geq .40$	.145	.179	.162	4

Note. *The ranking starts from the smallest to largest: 1 = lowest errors, 6 = highest errors.

Discussion

The current study has examined item selection methods commonly used in psychological research in Indonesia. A good selection method is the one that generates least errors, both errors in excluding good items and including poor items.

The results showed that significance-based methods, either *r* or *t*-test, tended to generate more errors in including poor items. These errors increased as the

sample size increased, particularly when poor items were assumed to measure another construct. This means these two methods tended to include poor items in the instrument.

The increase of errors as the sample size increased can be explained as follows: Larger sample size increased the analytical power which means more sensitivity to detect small correlations in the population. Errors in including poor items increased when sample size was increased in the first simulation, because although r_{it} was determined .116 (small effect size), the increased sample size increased analytical power so that more sensitivity to identify those small correlations. This explanation was confirmed in the second simulation, where errors in including poor items were relatively stable when the sample size increased. In the second simulation, the r_{it} was determined .00 in the population, so that the sample size did not increase the possibility to reject the null hypothesis.

The r_{it} method tended to generate errors in excluding good items. However, these errors decreased when the sample size was increased. This decrease was because the estimation of r_{it} became more stable with the increase of sample size. Of all four r_{it} criteria proposed, in general the criterion $r_{it} \geq .30$ generated least errors.

Another finding of the study was, when poor items measured another construct, errors in including poor items was increased. This means that these poor items did not just distort the measurement validity, but also distort the item selection process because these items were tended to be included, particularly in small sample size. It means that item-total correlation tended to give false information about the item quality when the instrument included a different construct from the purpose of the measurement.

There are three implications of the limitations of item-total correlation in this regard: (1) item-total correlation did not provide information about item validity, because when some items measured another construct, those items were harder to exclude; (2) the explication process of the construct intended to be measured and item writing process were crucial processes in the instrument construction because the researcher must make sure the items only measure constructs relevant with the aim of the measurement; and (3) item selection process needs to have larger sample size.

In relation to the third implication above, the study found that criterion $r_{it} \geq .30$ needed at least 200 respondents for the item selection accuracy to approach 100%. If the researcher is confident with the unidimensionality of the items, a sample size of 150 can be considered adequate.

Limitations

This study showed that criterion $r_{it} \geq .30$ generated the best result compared to other methods. However, this criterion was still unsatisfying when the sample size was small (n = 50 to 150). The researcher hypothesized that this is because of the use of correlations from the samples were rendered representative of the population without considering variations between samples. Therefore, criteria other than $r_{it} \geq .30$ may be needed to consider variations between samples so that the result can be more satisfying.

Conclusion

Based on the explanation above the researcher concluded that significance-based selection methods were less suggested because they tended to generate more errors, particularly in including poor items.

Selection method using r_{it} criteria was better because it generated less errors, particularly when sample size was larger. Of all four criteria proposed, the criterion $r_{it} \geq .30$ was the best.

Poor items that measured another construct tended to make item-total correlation provided false information about the item quality. This error can be managed by using larger sample size. Based on the conditions of this study, it is found that 200 samples were adequate. In different conditions, such as when correlation between poor items that measured another construct was larger than .30, the researcher hypothesized that the sample size needed becomes larger.

Recommendations

The results of the current study showed the superiority of the criterion $r_{it} \geq .30$ as compared to other criteria and methods. Therefore, it is recommended that this criterion is used in analyzing item quality. Significance-based methods are not recommended because those methods tended to generate more errors, particularly in including poor items. Such errors increased when the sample size increased.

Larger sample sizes are suggested in item selection because this study found that criterion $r_{it} \geq .30$ generated more accurate items closer to the condition of the population when then sample sizes were large.

The simulation also showed that when unidimensionality assumption was rejected, item-total correlation from smaller sample sizes tended to provide

false information about the item quality. Therefore, beside larger sample size, the processes of construct explication and item writing are important to make sure the items only measures one construct. It is also suggested that item-total correlation is no longer considered as an indicator of item validity because of the finding of this study.

References

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River, N.J: Pearson.

Azwar, S. (2013). *Reliabilitas dan validitas* (4th ed.). Pustaka Pelajar.

Domino, G., & Domino, M. L. (2006). *Psychological testing : An Introduction (2)*. Cambridge, GB: Cambridge University Press. Retrieved from http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10130394

Edwards, A. L. (1957). *Techniques of attitude scale construction.* New York: Appleton-Century-Crofts. Retrieved from https//catalog.hathitrust.org/Record/000617435

Hadi, S. (2005). Aplikasi ilmu statistika di fakultas psikologi. *Anima Indonesian Psychological Journal, 20*(3), 203–229.

Kline, T. (2005). *Psychological testing: A practical approach to design and evaluation.* Thousand Oaks, CA: SAGE Publications, Inc. Retrieved from http://methods.sagepub.com/book/psychological-testing

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin, 113*(1), 181–190. http://dx.doi.org/10.1037/0033-2909.113.1.181

(Appendices follow)

Appendix A
R Code for Simulation

The content of Appendix A is kept by the authors.
Interested readers may contact the authors to obtain it
(agungsan_psy@yahoo.com)

Appendix B
Calculating Corrected Item-Total Correlation using Variance-Covariance Matrix

The content of Appendix B is also kept by the authors.
Interested readers may contact the authors to obtain it
(agungsan_psy@yahoo.com)