Anima Indonesian Psychological Journal
2018, Vol. 33, No. 3, 178-189

https://doi.org/10.24123/aipj.v33i3.1694
0215-0158

Utilizing Fisher's-Z Transformation for Item Selection

Agung Santoso
Faculty of Psychology
University of Sanata Dharma

The previous work has shown that item selection method based on the use of corrected item-total correlation larger than .30 as the criterion provided the least errors of including items with low corrected item-total correlation in the population and excluding items with high corrected item-total correlation in the population. However, such method did not address the fact that corrected item-total correlation fluctuated across samples. Therefore, in smaller samples, the method provided larger errors. The current article proposed a new method for item selection that took into account the fluctuations of corrected item-total correlation across samples. The method was a significant test of correlation coefficient with the null hypothesis stating that the corrected item-total correlation was larger than or equal to .30. Four simulations were conducted to evaluate the proposed method and its modification. The results showed that the method was performed very well in reducing errors of including items with low corrected item-total correlation even in smaller sample sizes. However, the errors of excluding items with high corrected item-total correlation were large, particularly in small sample size. The large exclusion error was due to the lack of power to reject the null hypothesis when sample size was small. In larger samples, the proposed method and its modification and the method used criterion of corrected item total correlation larger than .30 performed equally well.

Keywords: corrected item-total correlation, item quality, item discrimination index, Fisher's-z transformation, inclusion error, exclusion error

Studi terdahulu menunjukkan bahwa metode seleksi item yang didasarkan pada penggunaan korelasi item-total terkoreksi sama dengan .30 sebagai kriteria $r_{it} \geq .3$ menghasilkan kesalahan terkecil dalam memasukkan item-item dengan korelasi item-total terkoreksi yang kecil di populasi dan menggugurkan item-item dengan korelasi item-total terkoreksi yang besar di populasi. Namun demikian, metode tersebut tidak dapat memecahkan permasalahan terkait dengan fakta bahwa korelasi item-total terkoreksi r_{it} berfluktuasi antar sampel. Oleh karena itu, dalam sampel yang lebih kecil, metode tersebut menghasilkan banyak kesalahan. Artikel ini mengajukan sebuah metode baru untuk melakukan seleksi item didasarkan pada korelasi item-total terkoreksi yang memperhitungkan fluktuasi korelasi item-total terkoreksi antar sampel. Metode yang ditawarkan adalah uji signifikansi koefisien korelasi dengan menggunakan hipotesis nul yang menyatakan bahwa korelasi item-total terkoreksi di populasi sebesar .30. Empat simulasi dilakukan untuk mengevaluasi metode yang diajukan dan modifikasinya. Hasil simulasi menunjukkan bahwa metode yang diajukan memberikan hasil yang sangat baik dalam mengurangi kesalahan memasukkan item dengan korelasi item-total terkoreksi yang kecil di populasi. Namun demikian, kesalahan mengugurkan item dengan korelasi item-total terkoreksi yang besar di populasi menjadi besar, khususnya dalam sampel dengan ukuran kecil. Hal ini terjadi karena kurangnya daya analisis untuk menolak hipotesis nul ketika ukuran sampel kecil. Dalam sampel dengan ukuran lebih besar, modifikasi dari metode yang diajukan dan metode dengan menggunakan kriteria besarnya korelasi item-total terkoreksi memberikan hasil yang setara.

Kata kunci: korelasi item-total terkoreksi, kualitas item, indeks diskriminasi item, transformasi Fisher's-z, kesalahan inklusi, kesalahan eksklusif

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

Item quality is an important characteristic of a test that should be achieved in tests development. Items of low quality do not only reduce the reliability of test scores but also are detrimental to the test validity. An items quality that is often used to select items to be included in a test is item discrimination. Several methods have been proposed to select items based on items' discrimination index (Azwar, 2013; G. Domino & M. L. Domino, 2006; Hadi, 2005; Kline, 2005; Urbina, 2014). Santoso (2017) examined several methods of selecting items based on item discrimination index and found that the use of criterion based on coefficient of corrected item-total correlation provided the least errors of either including items that should not be in the test (i.e., inclusion error) or excluding items that should be in the test (i.e., exclusion error), particularly when sample size was large. However, the use of estimates of corrected item-total correlation ignores the fact that the estimates have a distribution across samples, standard error of which is affected by sample size. When the sample size is small, the standard error becomes large, making the fluctuation of corrected item-total correlation value across samples large. Consequently, the large fluctuation results in a large inclusion and exclusion errors. Such fact can be observed in Santoso's study showing that the use of criterion of corrected item-total correlation larger than .30 resulted in large inclusion and exclusion errors when sample size was small.

One way to amend such a limitation is by taking the distribution of the corrected item-total correlation into account in examining item quality by using statistical significance test. However, the performance of conventional use of statistical significance test, by testing a null hypothesis that the corrected item-total correlation in the population (r_{it}) was smaller than or equal to zero, was shown to be inferior compared to using selection method based on criterion of corrected item-total correlation larger than or equal to .30, particularly in terms of inclusion error (Santoso, 2017). The weakness of the method lays on the use of incorrect null hypothesis stating that the value of corrected item-total correlation in the population is zero. By using such hypothesis, one allows any items that have the value of corrected item-total correlation in the population larger than zero be included in a test so that, by enough statistical power (e.g., large enough sample size), even items with corrected item-total correlation in the population very close to zero in the population can be included in a test. For example, an item that has a value of corrected item-total correlation in the population equals to .10 in the population has a probability of .89 to be included in the test[1] when research samples are 1000. Such results introduce larger inclusion error as the sample size becomes larger.

The current study proposed the use of statistical significance test by using a null hypothesis stating that the corrected item-total correlation of the studied item is less than *P* in the population, as a new method of item selection. Here, *P* is the value of corrected item-total correlation that is considered good by researchers. By using such a method, only items that have corrected item-total correlation larger than *P* in the population are allowed to stay in the test. Therefore, although the analysis involves a very large sample size, thus a more powerful test, the method does not allow items with corrected item total correlation less than *P* be included in the test.

Two methods can be used to test the null hypothesis stating that the corrected item-total correlation in the population is less than *P*: (1) one sample t-test of correlation coefficient modified by Kraemer (1980); and (2) test of Fisher's-z transformation (Fisher, 1921), by assuming that the estimate of the corrected item-total correlation, follows normal distribution. The two methods approximate the test statistic for Pearson's product moment correlation when its value in the population is not zero, particularly when the analysis is conducted using small sample size. The current study compared the use of the proposed methods with the use of a criterion of corrected item-total correlation of the sample larger than .30 and conventional significance test of null hypothesis stating that the corrected item-total correlation in the population is equal or less than 0, to evaluate the proposed method effectiveness in selecting items. Recommendation based on which method provided lower inclusion and exclusion errors was therefore can be made.

Significance Test for Corrected Item-Total Correlation

Two statistical techniques that can be used to conduct significance test of corrected item-total correlation are the Kraemer's one sample t-test and Fisher's-z transformation test. Basically, one sample t-test of corrected item-total correlation is the same as t-test for correlation coefficient in general, because estimation of corrected item-total correlation in the popula-

[1] The estimate of the probability, of a population with certain value of corrected item-total correlation to have a significant test against null hypothesis of corrected item-total correlation equals to zero is calculated using R code provided in Appendix A.

tion is based on Pearson's product moment correlation coefficient. The *t* statistic of an obtained correlation value, when the sample drawn from a bivariate normal distribution with correlation value in the population equals to zero, is as the following:

$$t(r \mid \rho = 0, v) = \frac{(r\sqrt{v})}{\sqrt{(1-r^2)}} \qquad (1)$$

where $v = n - 2$. The value resulted in Equation (1) follows a *t* distribution with *v* degrees of freedom (Fisher, 1915; Kraemer, 1980). When the value of the correlation in the population equals to zero, the exact distribution is a complex function and can only be approximated by using:

$$z = \sqrt{n}\,(r - \rho) / (1 - \rho^2) \qquad (2)$$

when sample size is very large. A closer approximation that works well when the correlation coefficient in the population is non-zero with smaller sample size is as the following:

$$t(r \mid \rho, v) = (r - \rho)\,\sqrt{v} / \sqrt{(1 - r^2)(1 - \rho^2)} \qquad (3)$$

that follows *t* distribution with *v* degrees of freedom (Kraemer, 1980) .

The value obtained from (3) is evaluated by using *t* distribution with $= n - 2$. If the value of *t* obtained from the sample is larger than the critical value of $t_{df,\alpha/2}$, and one may conclude that, then the null hypothesis that the correlation coefficient in the population is less than or equal to is rejected, leading to inclusion of item in the scale. Otherwise, the item is excluded from the scale.

Another approximation was obtained by using the normalizing and variance-stabilizing transformation (Fisher, 1915, 1921). First, the value of correlation coefficient obtained from sample (*r*) and the value of correlation coefficient criterion ρ is transformed to Fisher's-z by the following formula:

$$z(r) = \frac{1}{2} \; ln \; \left(\frac{1 + r}{1 - r} \right) = arctanh(r) \qquad (4)$$

Then, we calculate the statistic of difference between the two *z* values as the following:

$$z(r - \rho) = \sqrt{n - 3}\,(z(r) - z(\rho)) \qquad (5)$$

The value obtained from (5) is evaluated based on standard normal distribution. If the obtained $z(r - \rho)$ is larger than $z_{\alpha/2}$ and $r > \rho$ then the null hypothesis is rejected.

Although developed using different approximations, the statistical tests resulting from the two techniques demonstrates very high agreement. The author conducted a simulation to illustrate this point, which R-codes are provided in the Appendix B. The simulation showed that the differences of obtained *p*-values ranged from - 0.00047 to 0.00047, with mean of - 3.02 * 10^{-6}. The results suggested a negligible difference of *p*-values obtained that may result in similar conclusion about the significance test from the two techniques. Therefore, the author used only Fisher's-z transformation test in conducting the current study.

Method

Conditions for data generating procedures in current study followed Santoso (2017). There was only one condition for the number of items in the test, which was fifty. The fifty items consisted of forty items set to have high values of corrected item-total correlation in the population (Group 1) representing *good items* and ten items set to have low values of corrected item-total correlation in the population (Group 2) representing *bad items*. The author used two procedures of generating items data to have high and low corrected item-total correlation by setting the correlation between items (ρ_{ii}) first and then calculated corrected item-total correlation in the population resulting from the structure of the correlation between items. Here, the correlation between items of the Group 2 is the first independent variable manipulated in current study.

In the first procedures, the correlations between items in Group 1 were set to be 0.3, while the correlations between items in Group 2 were set to be 0.0. The correlations between Group 1 items with Group 2 items were set to be 0.0. This condition reflected a situation in which the test measured only one latent factor with some random disturbance from items that poorly measures the latent factor, or a condition of pure reliability problem condition. In such condition the corrected item-total correlation in the population for the forty items of Group 1 were 0.527, while the corrected item-total correlation in the population of the ten items of Group 2 were 0.0[2].

In the second procedures, the correlations between items in Group 1 and Group 2 were set to be 0.3, while the correlations between Group 1 items and Group 2

[2] The derivation of the formula to obtain corrected item-total correlation in the population and its application in R can be seen in the Appendix B of Santoso (2017).

items were set to be 0.0. Such condition reflected a situation in which the forty good items measured one latent factor while the other ten measured another latent factor, while the correlation of the two latent factors were zero in the population. Such conditions reflects a validity problem in which one test measured more than one latent factor but treated as if it measures only one factor. In such condition, the corrected item-total correlation in the population for Group 1 items were .513, while the correlation for Group 2 items were .116.

The second independent variable was sample sizes that were chosen to be 50, 100, 250 and 500 representing small to large sample sizes, resamples 1000 times each. For each sample, the author calculated the value of the corrected item-total correlation and tested the null hypothesis stating that the corrected item-total correlation in the population was less than .30, .20, and 0 by using Fisher's-z transformation test and conventional NHT of ρ_{it} = 0f . The R codes and implementation of the codes is given in Appendix D. The author also used criteria of the corrected item-total correlation in the sample larger than .25 and .30 as item selection procedures to evaluate the performance of the proposed method. Then, the author calculated the number of items from Group 1 that was excluded (exclusion error) and the number of items from Group 2 that was included (inclusion error) based on information from the aforementioned methods. The two errors were the outcome variables of the current simulation. The results from one thousand samples were then tabulated to summarize the number of inclusion errors and exclusion errors made across one thousand samples for each method. The results of tabulation were then presented in tables. The author compared the results from the simulation to evaluate which methods provided the least inclusion and exclusion errors.

Results

The results of the simulation are presented in Tables 1 and 2. It can be seen that the proposed method provided substantially smaller error of including the Group 2 items in the test compared to the other methods. It means that the proposed method tended to exclude items with small item-total correlation in the population either in the condition when the Group 2 items had no correlation to each other or when the Group 2 items had moderate correlation to each other. Compared to the other methods, the use of significance test to test null hypothesis of ρ_{it} = 0, performed the worst. Increasing the sample size did not reduce even increased the errors of including Group 2 items in the test when the correlation between Group 2 items were not zero.

However, the proposed method provided a very large exclusion error, that was excluding good items that should be retained in the test. The large inclusion error was particularly happened when sample size is small. It means that the large inclusion error was caused by small analysis power to reject the null hypothesis stating that the corrected item-total correlation was less than or equal to .30. In the larger sample size condition, the inclusion errors of the proposed method decreased substantially, while the exclusion error was still small. For example, when n = 250 and correlation between Group 2 items were zero, the use of corrected item-total correlation in the population equals .30 provided .114 proportion of samples that have one to five items in Group 1 being excluded while the proportion of samples including items in Group 2 was zero.

It is also notable that the overall performance of the proposed method was inferior to the use of criteria of corrected item-total correlation in the sample larger than .30 and .20. The inferior performance of the proposed method was due largely to inclusion errors that were related to low power in smaller sample sizes. To improve the proposed method, the author proposed a way to determine P so that the probability of attaining sample corrected item-total correlation larger than or equal to P in a population with corrected item-total correlation equals to .30 was equal to .90. The R code to obtain the adjusted R is presented in Appendix C.

The author conducted another simulation to evaluate the effect of different P determined in the way mentioned in the previous paragraph. The results of the simulation are presented in Tables 3 and 4.

The results show that the use of the adjusted P reduced the exclusion errors obtained in item selection based on Fisher's-z transformation test, while maintaining a good level of inclusion errors particularly in sample sizes larger than .50 When the correlation between items in Group 2 was set to .30 and sample size of 250, the Fisher's-z transformation test even outperformed the use of criterion of corrected item-total correlation in the sample larger than or equal to .30. When sample size reached one thousand the three methods provided no errors. The author concluded that the use of the adjusted criteria performed better than the use of significance test with null

Table 1
Result of the Simulation With ρ_{ii} Between Group 2 Items was .00

Number of Errors	Exclusion Errors					Inclusion Errors				
	r_{it} significance	$r = .25$	$r = .30$	$\rho_{it} = .20$	$\rho_{it} = .30$	r_{it} significance	$r = .25$	$r = .30$	$\rho_{it} = .20$	$\rho_{it} = .30$
					n = 50					
0	782	725	500	62	3	579	655	841	982	996
1 – 5	217	271	474	475	79	421	345	159	18	4
5 – 10	1	4	25	291	181	0	0	0	0	0
> 10	0	0	1	172	737	0	0	0	0	0
					n = 100					
0	998	980	908	642	66	614	949	993	1000	1000
1 – 5	2	20	92	353	549	386	51	7	0	0
5 – 10	0	0	0	5	267	0	0	0	0	0
> 10	0	0	0	0	118	0	0	0	0	0
					n = 250					
0	1000	1000	1000	1000	885	584	1000	1000	1000	1000
1 – 5	0	0	0	0	114	416	0	0	0	0
5 – 10	0	0	0	0	1	0	0	0	0	0
> 10	0	0	0	0	0	0	0	0	0	0
					n = 1000					
0	1000	1000	1000	1000	1000	617	1000	1000	1000	1000
1 – 5	0	0	0	0	0	383	0	0	0	0
5 – 10	0	0	0	0	0	0	0	0	0	0
> 10	0	0	0	0	0	0	0	0	0	0

Table 2
Result of the Simulation With ρ_{ii} Between Group 2 Items was .30

Number of Errors	Exclusion Errors					Inclusion Errors				
	r_{it} significance	$r = .25$	$r = .30$	$\rho_{it} = .20$	$\rho_{it} = .30$	ρ_{it} significance	$r = .25$	$r = .30$	$\rho_{it} = .20$	$\rho_{it} = .30$
					n = 50					
0	704	647	427	43	0	252	305	515	894	990
1 – 5	295	346	531	373	39	693	662	477	106	10
5 – 10	1	7	39	341	155	55	33	8	0	0
> 10	0	0	3	243	806	0	0	0	0	0
					n = 100					
0	1000	951	836	514	39	138	525	811	943	995
1 – 5	0	49	164	471	387	668	463	188	57	5
5 – 10	0	0	0	15	342	194	12	1	0	0
> 10	0	0	0	0	232	0	0	0	0	0
					n = 250					
0	1000	1000	999	999	779	12	878	981	979	1000
1 – 5	0	0	1	1	219	456	122	19	21	0
5 – 10	0	0	0	0	2	532	0	0	0	0
> 10	0	0	0	0	0	0	0	0	0	0
					n = 1000					
0	1000	1000	1000	1000	1000	0	1000	1000	1000	1000
1 – 5	0	0	0	0	0	1	0	0	0	0
5 – 10	0	0	0	0	0	999	0	0	0	0
> 10	0	0	0	0	0	0	0	0	0	0

hypothesis using *P* = .30 and slightly better than criteria of corrected item-total correlation equals to in the sample is larger than or equal to .30 particularly when sample size is moderate. However, when item selection was con-ducted in a very large sample size, the three methods provided accurate inclusion and exclusion of items.

Discussion

The study was conducted to propose a new method to conduct item selection based on item discrimination properties. The proposed method and its modification were compared to the other methods currently used in item selection.

The results of the simulation showed that the proposed method was best in addressing the problems of inclusion of items that has low corrected item-total correlation in the population, or inclusion of bad items. However, the proposed method did not perform well in dealing with exclusion errors, particularly when the power of the analysis was small due to smaller sample size. The overall performance of the proposed method was also inferior compared to the use of criterion of corrected item-total correlation in the sample larger than or equal to .30.

The modification of the proposed method, by using the adjusted *P* criteria, provided better results than the proposed method. The exclusion errors were reduced while the inclusion errors were maintained to be small. The modification of the proposed method

Table 3

Result of the Simulation with ρ_{ii} Between Group 2 Items was 0.0 Using Adjusted P for Each Sample Size

Number of	Exclusion Errors			Inclusion Errors		
Errors	*r = .30*	ρ_{it} = .30	ρ_{it} = *adj.P*	*r = .30*	ρ_{it} = .30	ρ_{it} = *adj.P*
n = 50						
0	489	3	257	866	999	945
1 – 5	491	81	641	134	1	55
5 – 10	18	179	87	0	0	0
> 10	2	737	15	0	0	0
n = 100						
0	901	80	749	985	1000	997
1 – 5	99	519	249	15	0	3
5 – 10	0	263	2	0	0	0
> 10	0	138	0	0	0	0
n = 250						
0	999	886	997	1000	1000	1000
1 – 5	1	114	3	0	0	0
5 – 10	0	0	0	0	0	0
> 10	0	0	0	0	0	0
n = 1000						
0	1000	1000	1000	1000	1000	1000
1 – 5	0	0	0	0	0	0
5 – 10	0	0	0	0	0	0
> 10	0	0	0	0	0	0

Note. adj.P = adjusted criteria by finding the value of P so that the probability of samples drawn from ρ_{it} = .30 have values larger than P is equal to .90

Table 4

Result of the Simulation With ρ_{ii} Between Group 2 Items was 0.3 Using Adjusted P for Each Sample Size

Number of	Exclusion Errors			Inclusion Errors		
Errors	*r = .30*	ρ_{it} = .30	ρ_{it} = *adj.P*	*r = .30*	ρ_{it} = .30	ρ_{it} = *adj.P*
n = 50						
0	399	1	197	522	975	676
1 – 5	557	31	624	463	25	320
5 – 10	41	137	153	15	0	4
> 10	3	831	26	0	0	0
n = 100						
0	842	40	690	803	999	911
1 – 5	157	411	301	196	1	88
5 – 10	1	317	9	1	0	1
> 10	0	232	0	0	0	0
n = 250						
0	999	781	996	989	1000	998
1 – 5	1	217	4	11	0	2
5 – 10	0	1	0	0	0	0
> 10	0	1	0	0	0	0
n = 1000						
0	1000	1000	1000	1000	1000	1000
1 – 5	0	0	0	0	0	0
5 – 10	0	0	0	0	0	0
> 10	0	0	0	0	0	0

Note. adj.P = adjusted criteria by finding the value of P so that the probability of samples drawn from ρ_{it} = .30 have values larger than P is equal to .90

was also slightly superior to the use of corrected item-total correlation larger than or equal to .30 criteria, in a condition in which the correlation among bad items (i.e., Group 2 items) were not zero. The performance of the modified method was improved because the adjusted *P* as the criterion took into account the dependency of the variability of the estimates of corrected item-total correlation on the sample size. The smaller the sample size, the larger the variability of the estimates became, therefore the value of the criterion was adjusted to increase power of the analysis. The adjustment was also dependent on the sample size so that the adjustment was large when the sample size was small resulting in higher power. The dependency of the adjustment followed the dependency of the variability so that it was not too small that it may increase the inclusion of bad items.

Based on the overall performance of the proposed method and its modification in the current study, it seems to be reasonable to abandon the proposed method and turn to the use of criterion of corrected item-total correlation larger than or equal to .30 instead. However, the readers should be reminded that although the overall performance of the proposed method and its modification were inferior to the use of the criterion , the performance of the proposed method and its modification were substantially superior in reducing inclusion of bad items, particularly in smaller sample sizes. Because the inclusion of bad items, particularly those that measure unintended constructs, greatly impairs the reliability as well as the validity of a test, the proposed method and, particularly, its modification should also be utilized accompanying the use of the criterion.

The results of the current study supported the suggestions frequently made in the textbooks of test constructions to use the criterion of corrected item-total correlation larger than or equal to .30. However, the results also showed potential alternatives of the criterion that might be more beneficial in dealing with inclusion errors, particularly when the sample size was small. Though still having a problem with statistical power to reduce exclusion errors, the proposed method and its modification performed substantially better than the criterion of corrected item-total correlation larger than or equal to .30.

The results also confirmed the previous study by Santoso (2017) that the sample size required to provide tolerable amount of errors of including bad items and excluding good items was at least 250 with moderate amount of items. In smaller sample sizes, all methods tended to provide larger errors. The current study also confirmed the disadvantages of using statistical test of testing the null hypothesis of corrected item-total correlation in the population equals to zero in selecting items. The method provided large amount of errors and could not be ameliorated by increasing sample sizes.

Limitations

In the current study, the author only considered two conditions of correlation between bad items, including zero correlation and correlation of .30 between bad items, while constraining the correlation between bad and good items to be zero. There might be other conditions of correlation structure between items that can be included in future studies. The current author assumed the items scores as continuous, while in substantive research the items scores can be discrete. Future studies may include conditions in which the items scores were discrete, either dichotomous or polytomous. The proposed method was based on the assumption of normality of r_{it}, that might be the cause of the lower power analysis. Future studies may investigate the use of methods that may relax the normality assumption such as bootstrap method in conducting the statistical inference.

Conclusions and Recommendation

The significance test for testing the null hypothesis of $\rho_{it} = 0$ should not be used as the method to select items in the future. The item selection procedure needs a minimum sample size of 250 based on current simulation study. However, it should be noted that in current study, the author did not include condition of sample size between 100 and 250. Therefore sample sizes between the two values might provide good enough results. The item selection should be based on the criteria of sample corrected item-total correlation larger than or equal to .30 accompanied by the modified proposed method to also take into account the possibility of including bad items, particularly in smaller sample size.

References

Azwar, S. (2013). *Reliabilitas dan validitas* (4th ed.). Yogyakarta: 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

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika, 10*(4), 507-521. https://doi.org/10.2307/2331838

Fisher, R. A. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron, 1,* 3-32.

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.* 2455 Teller Road, Thousand Oaks California 91320 United States: SAGE Publications, Inc. Retrieved from http://methods.sagepub.com/book/psychological-testing

Kraemer, H. C. (1980). Robustness of the distribution theory of the product moment correlation coefficient. *Journal of Educational Statistics, 5*(2), 115-128. https://doi.org/10.2307/1164676

Santoso, A. (2017). Comparing t test, r_{it} significance test, and r_{it} criteria for item selection method: A simulation study. *Anima Indonesian Psychological Journal, 32*(2), 99-108. https://doi.org/10.24123/aipj.v32i2.588

Urbina, S. (2014). *Essentials of behavioral science ser. : Essentials of psychological testing* (2). Somerset, US: John Wiley & Sons, Incorporated. Retrieved from http://site.ebrary.com/lib/alltitles/docDetail.action?docID=10891094

(Appendix follows)

Appendix A

R Code for Estimating The Probability of A Population with A Certain r_{it} to Have A Significant Test Against Null Hypothesis of $r_{it} = 0$

Code to find the critical value of t under null hypothesis, when *n* = 1000

```
t0<-qt(.975,(1000-2))
```

Code to find the value of r corresponded with t0

```
t0.to.r<-function(t,n){
           sqrt(t^2/(t^2+(n-2)))}
r<-t0.to.r(t0,1000)
```

Code to approximate the probability of having significant *r* when *r* in the population is .1 (*r0* = .1)

```
r.test<-function(r,r0,n){
           res<-NULL
           t<-(r-r0)*sqrt((n-2)/(1-r^2))
           pval<-pt(t,df=n-2,lower.tail=FALSE)
           res$t<-t
           res$pval<-pval
           return(res)
}
r.test(r,0.1,1000)$pval
```

Or we can also use the code below after obtaining t0:

Code to find non centrality parameter for t distribution when *r* = 0.1 and *n* = #1000

```
ncp<-r.test(0.1,0,1000)$t
```

Code to approximate the probability of having significant *r* when *r* in the population is .1

```
pt(t0,df=1000-2,ncp=ncp,lower.tail=FALSE)
```

(Appendix continues)

Appendix B

R Code for Illustrating The High Agreement between Kraemer's one sample t-test and Fisher's-z Transformation Test

Code of function to conduct Fisher's-z transformation test:

```
z.rit.test=function(r,r.crit,N){
z.crit=0.5*log((1+r.crit)/(1-r.crit))
zz=0.5*log((1+r)/(1-r))
SE=1/(sqrt(N-3))
z=(zz-z.crit)/SE
pnorm(z,lower.tail=F)
}
```

Code of function to conduct Kraemer's one sample t-test:

```
kraemer.rit.test=function(r,r.crit,N){
se=sqrt((1-r^2)*(1-r.crit^2))
t=(r-r.crit)*sqrt(N-2)/se
pt(t,lower.tail = F,df=(N-2))
}
```

Code to obtain r_{it} values ranging from 0 to .95 with .01 intervals:

```
rtrial=seq(0,.95,.01)
```

Conducting the Fisher's-z transformation test and Kraemer's one sample t-test of r_{it} values in rtrial, to obtain the p-values, and save the results in z.result and t.result, respectively.

```
z.result=z.rit.test(rtrial,.3,100)
t.result=kraemer.rit.test(rtrial,.3,100)
```

Take the difference of p-values obtained from Fisher's-z transformation test and Kraemer's one sample t-test and find the minimum, maximum and mean of the difference.

```
dif.p=z.result-t.result
min(dif.p)
max(dif.p)
mean(dif.p)
```

(Appendix continues)

Appendix C

R Code for Obtaining The Criterion P in Testing The Null Hypothesis ρ_{it} = P

The code to obtain the non-centrality parameter of the *t* distribution given ρ_{it} = .3 and save it in t0:

```
t0=r.test(0.3,0,n-2)
```

Note that the command r.test has been defined in Appendix A. After obtaining the non-centrality parameter of the t distribution, we calculated the value of t that has the cumulative probability of .1 in a non-central t distribution with non-centrality parameter was equal tot0.

```
t=qt(.1,ncp=t0$t,df=n-2)
```

We calculated the value of r_{it} that corresponded to the value of t obtained and used it as P in testing the null hypothesis of ρ_{it} = P.

```
rcrit=sqrt((t^2/(t^2+n-2)))
```

(Appendix continues)

Appendix D

R Codes for Conducting Fisher's-Z Transformation Test for Item Selection

The R codes for conducting Fisher's-Z transformation test for item selection is given below:

```
F.rit=function(x,r.crit=0.3){
N=dim(x)[1]
z.rit.test=function(r,r.crit,N){
z.crit=0.5*log((1+r.crit)/(1-r.crit))
z=0.5*log((1+r)/(1-r))
SE=1/(sqrt(N-3))
z=(z-z.crit)/SE
pnorm(z,lower.tail=F)
}
rit2=function(x){
rit.d=NULL
nit=dim(x)[2]
rsum=rowSums(x)
for(i in 1:nit){
sumd=rsum-x[,i]
rit.d=c(rit.d,cor(x[,i],sumd))
}
rit.d
}
r=rit2(x)
F.sig=t.sig=NULL
for(i in 1:length(r)){
F.sig=c(F.sig,z.rit.test(r[i],N=N,r.crit=r.crit))
}
list('n'=N,'Fisher.rit'=cbind(r,F.sig))
}
```

To run the analysis, one needs to run the previous command first, load the data and save it in an R object, and then write the command as follows:

```
F.rit(dat,r.crit=0.3)
```

The `dat` is an R object consists of the data of item scores for all participants that have been loaded. The data should be arranged so that columns represent items, and rows represent participants. The line `r.crit=0.3` tells the command to test the null hypothesis of $\rho_{it} \leq .3$. One can change the value to another preferred value or use the adjusted criterion. The procedure to obtain the adjusted is described in Appendix C.