

## The effectiveness of item parceling to increase the model fit: A case study of PAPs

Anindita Dwi Hapsari\*; Wahyu Widhiarso

Universitas Gadjah Mada, Indonesia

\*Corresponding Author. E-mail: [anindita.dwi.h@mail.ugm.ac.id](mailto:anindita.dwi.h@mail.ugm.ac.id)

### ARTICLE INFO

#### Article History

**Submitted:**

11 April 2022

**Revised:**

20 March 2023

**Accepted:**

25 May 2023

#### Keywords

item parceling;  
confirmatory factor  
analysis; PAPs test

Scan Me:



### ABSTRACT

The impact of item parceling to improve model fit indexes in confirmatory factor analysis has been on debate amongst psychometricians. In this study, the effectiveness of item parceling was examined using *Tes Potensi Akademik Pascasarjana* (PAPs) or Postgraduate Academic Potential Test in Universitas Gadjah Mada. Item parceling approach, second-order approach, and item-based approach of confirmatory factor analysis (CFA) were used for examination. Data were collected from a sample of 1374 postgraduate candidates in 2017. The result found that model fit indices such as the chi-squared test, comparative fit index, Tucker-Lewis index, and standardized root mean square residual were improved in the item parceling approach when compared to item based approach. Interestingly, the root mean square error of approximation were deteriorating in the item parceling approach. The finding of this study suggested that model dimensionality and sample size should be carefully considered when using the item parceling approach.

This is an open access article under the [CC-BY-SA](https://creativecommons.org/licenses/by-sa/4.0/) license.



### To cite this article (in APA style):

Hapsari, A., & Widhiarso, W. (2023). The effectiveness of item parceling to increase the model fit: A case study of PAPs. *Jurnal Penelitian dan Evaluasi Pendidikan*, 27(1), 26-38.

doi:<https://doi.org/10.21831/pep.v27i1.49012>

## INTRODUCTION

The usage of structural equation modelling (SEM) is used in the social sciences field of study because of its capability to estimate relationships between unobserved constructs, known as latent variables, from observable variables (Hancock, 2003). SEM analytic techniques are rapidly used in social and behavioral science study for the causal modeling of multivariate and complex data sets which provided multiple measures of proposed constructs (Hair, 2019). By employing SEM analysis, the researchers are able to create a relationship model among multiple predictor and criterion variables, construct underlying unobservable latent variables, represent errors in measurement for observed variables, and statistically examine an earlier substantive or theoretical and measurement assumption of an empirical data (Chin, 1998).

One of the most popular techniques in SEM analysis family is confirmatory factor analysis (CFA). The main goal of CFA is to find out if the data fit a hypothesized measurement model. This hypothesized model can be developed based on theory or previous research. Suhr (2003) indicates that CFA allows the researchers to test the hypothesis regarding the existence of the relationship between observed variables and their underlying latent construct(s). Burns et al. (2001) implemented CFA to evaluate five different organization model of Attention-Deficit/Hyperactivity Disorder (ADHD) and oppositional defiant disorder (ODD) in DSV-IV. CFA was also used in Wongpakaran et al. (2017) study to confirming the factor structure of the re-



vised version of Multidimensional Scale of Perceived Social Support (MSPSS) for the population of elderly with depressive disorders. Another application CFA was used by [Birch et al. \(2001\)](#) to assess underlying factors of The Child Feeding Questionnaire (CFQ), a self-report for measuring parental beliefs, attitudes, and practices of children feeding.

However, the researchers rarely use CFA despite of its importance. Previous study found out that exploratory factor analysis (EFA) is more commonly used to measure construct validity method compared to CFA ([Bastos et al., 2010](#); [Ladhari, 2010](#)). Many researches only use EFA because of its effectiveness in identifying latent variable by exploring the relationship between observed variables. This can lead to a problem since EFA can open more statistical subjectivity in decision-making process ([Roberson et al., 2014](#)). Many psychological scales can produce well-defined EFA structure but unable to produce adequate results using CFA approach ([Marsh et al., 2009](#)). CFA can be a frightening spectre for researchers because of its convolution in analyzing data. Another obstacle that the researchers might face in using CFA approach is the complexity and the strict rules in determining fit model ([Reilly, 1995](#)). The example of this case is the two- and three- indicator rules where it demands that there is only one factor in each indicator loading and there is no measurement errors of the indicator correlation ([Bainter & Bollen, 2014](#)). These rules require a limited form for the number of observed variables, factor complexity one, and no correlation among error terms ([Kenny, 1979](#)). More recent study also found out that commonly used CFA model fit criteria are too restrictive when used for multifactor measurement scale ([Marsh, 2012](#); [Marsh et al., 2005](#)).

Several reasons why these two- and three- indicator rules are difficult to fulfil, especially for researchers in psychology, is because construct in psychology is more likely to be multidimensional rather than unidimensional ([Drolet & Morrison, 2001](#)). Construct in psychology usually complex in nature because human consists of many dimensions, including value, trait, or motivation. Psychological phenomena can be perceived as a complex model because they are determined by a potentially huge number of variables that are connected within each other. For example, a psychological construct such as depression carries a broad set of implications about more specific components of behavioral, emotional, cognitive, and physiological functioning ([McGrath, 2005](#)). [Burisch \(1984\)](#) experimented with a single-item measurement of depression that include unanchored 9-point scale ranging from 1 (often depressed, moody, self-conscious) to 9 (contented, self-assured, poised). [Burisch \(1984\)](#) found this single-item measurement are able to show a good criterion-related validity, sometimes even matching that of measurement with more items but vague construct validity. Constructs in psychology as one of the social science studies tend to be multidimensional because of its non-observable (latent) variables that can be measured using various kinds of indicators ([Widhiarso & Ravand, 2014](#)).

The obstacle with the chi-squared test index as one indicator of model fit is that researchers may fail to reject an inappropriate model in small and large sample sizes ([Gatignon, 2016](#)). Sample size is crucial in factor analysis because it relates to the stability of the parameter estimates. Common rules of thumb for determining adequate sample size for a particular application of CFA in the size of the sample are more than 200 ([Myers et al., 2011](#)). The error of estimation can be reduced by increasing the number of observation which can increase statistical accuracy. In this way, using a large sample size for a model parameter is crucial ([Jackson, 2001](#)). Another reason to use a large sample size in CFA approach is the model result will be replicated and cross-validated in future case. [Bandalos \(1997\)](#) mentioned that structural equation models usually made for its predictive purposes, for replication in future cases, or as a representation of some underlying construct. Since larger sample size can represent a population better compared to smaller sample size, it can contribute to models' predictive purposes.

Nonconvergence is also likely to happen in CFA. Nonconvergence happened when the iterative maximum likelihood estimation procedure did not converge within the maximum iterations ([Boomsma & Hoogland, 2001](#)). Nonconvergents and improper solutions occurred more frequently for smaller sample sizes and for models with fewer indicators of each factor

(Anderson & Gerbing, 1984). Nonconvergence could occur in factor analysis because of three main factors, including sample size, size of factor loadings, and number of indicators per factor (Boomsma, 1985). Other than that, several factors can also contribute to the statistical performance of the test such as the distribution of the observed variables, the sample size and the structural misspecification of the test (Kolenikov & Bollen, 2012).

### Item Parcels

In the default application of SEM, individual items are used as indicators of the underlying construct that is purportedly measured by a given scale. Item-based approach is still widely used in the social field of study research. Another alternative approach is to aggregating items into several parcels as the indicators of the target construct or referred as parceling. Parceling is a measurement strategy that is commonly used in multivariate approaches to psychometrics, particularly for use with latent-variable analysis techniques that can be categorized as an aggregate-level indicator consisted of the sum or average of two or more items, responses, or behaviors (Little et al., 2002). However, parceling is not the same as creating subscales. Subscale usually based on certain theory that might have an explanation of its relation, whereas parceling did not have any relation and done in non-theoretical approach (Bandalos, 2008). The difference between a model that use item based approach and parceling approach in SEM analysis can be seen in Figure 1.

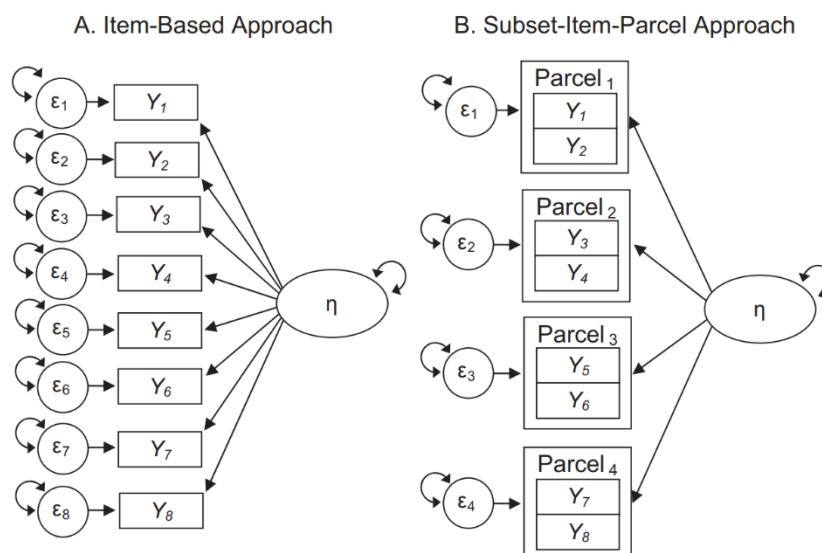


Figure 1. Comparison of Model Using Item-Parceling Approach and Item-Based Approach (Matsunaga, 2008)

Parcels can be utilized to reduce the number of indicators of a scale that consisted of many items and surveys when performing CFA (Bandalos & Finney, 2001). Parceling is usually done in studies with dichotomously or coarsely categorized measurement indicators to meet certain assumption of normality and continuity of statistical fit test (Bandalos & Finney, 2001). Marsh and Hau (2004) proposed several advantages of item parceling such as more definitive rotational results, less violation of normality assumptions, increased reliability, fewer parameters to be estimated resulted on more stable parameters. Parceling was found to be more beneficial in increasing scale communality and the common-to-unique factor for each indicator thus, it can reduce the random error score (Matsunaga, 2008). Forming parcels by combining items reduces coarsely distributed data that are unrelated to the construct and increases the ratio of the true score to the total score, and make the reliability higher (Matsunaga, 2008). Compared

to item-based approach, Bandalos (2002) study also demonstrates that chi-square value, CFI, and RMSEA indexes parcel-based solutions are better than item-based approach. From the modelling perspective, SEM solutions based on parcelled data provide stabler estimation and fit the data better than their item-based counterparts (Holbert & Stephenson, 2002).

The usage of item-parceling approach evoke several opinions from researcher. Marsh et al. (1998) stated that, although parcel-based models in general show good performance, item-based solutions would be preferable when all relevant factors including the parameter estimates are taken into consideration. Another important critique of parceling concerns multidimensionality and model misspecification. When the given scale is multidimensional, the use of parcels can obscure rather than clarify the factor structure of the data (Bandalos, 2002).

To prove this argument, a side-by-side comparison of item-based approach and item-parceling approach in the purpose of obtaining adequate absolute model score index should be applied with the same test and data. This study wants to find out the efficiency of parceling method compared to item-based method and second-order method to determine the absolute model score index with the same test and sample size.

## RESEARCH METHOD

### Participants

Participants in this study are the candidates for postgraduate program of the year 2017 (N=1374) in UGM. The data score in this PAPs test in a secondary data from Psychodiagnostic Test Development Unit (UPAP) UGM as the official administrator of PAPs test.

### Instruments

The purposed test used for this analysis is Postgraduate Academic Potential Test or *Test Potensi Akademik Pascasarjana* (PAPs) test. PAPs test was developed by UPAP to predict candidates' potential in pursuing a higher level of education (Widhiarso et al., 2015). The previous version of PAPs test possessed high reliability with the score of 0.92 (Belinda, 2015). The component score of PAPs test is obtained by summing every correct item on each subtest. Each correct item will be given 1 score and the wrong item will be scored 0. There is no score reduction for the wrong item. The most updated version of PAPs test is PAPs test series E. PAPs test consists of three subtests, namely verbal subtest, quantitative subtest, and figural subtest. Each subtest consists of four to five components and 50 items. In total, this test consists of 150 items. The specification of each subtest is elaborated in Table 1.

Table 1. Specification of PAPs Test Series E

Subtest	Component	Number of Items
Verbal	Synonym	15
	Antonym	15
	Word Analogy	15
	Analytic	5
Quantitative	Number Series	10
	Arithmetic	10
	Algebraic Concept	10
	Quantitative Computation	10
	Geometric Reasoning	10
Figural	Diagram	15
	Series	15
	Classification	10
	Picture Analogy	10

### Data Analysis Procedure

Before conducting SEM analysis, the technique of handling missing data should be properly considered because equations and formulas in SEM analysis requires complete data. In PAPs test, all wrong answer and missing data from each participants are replaced with zero.

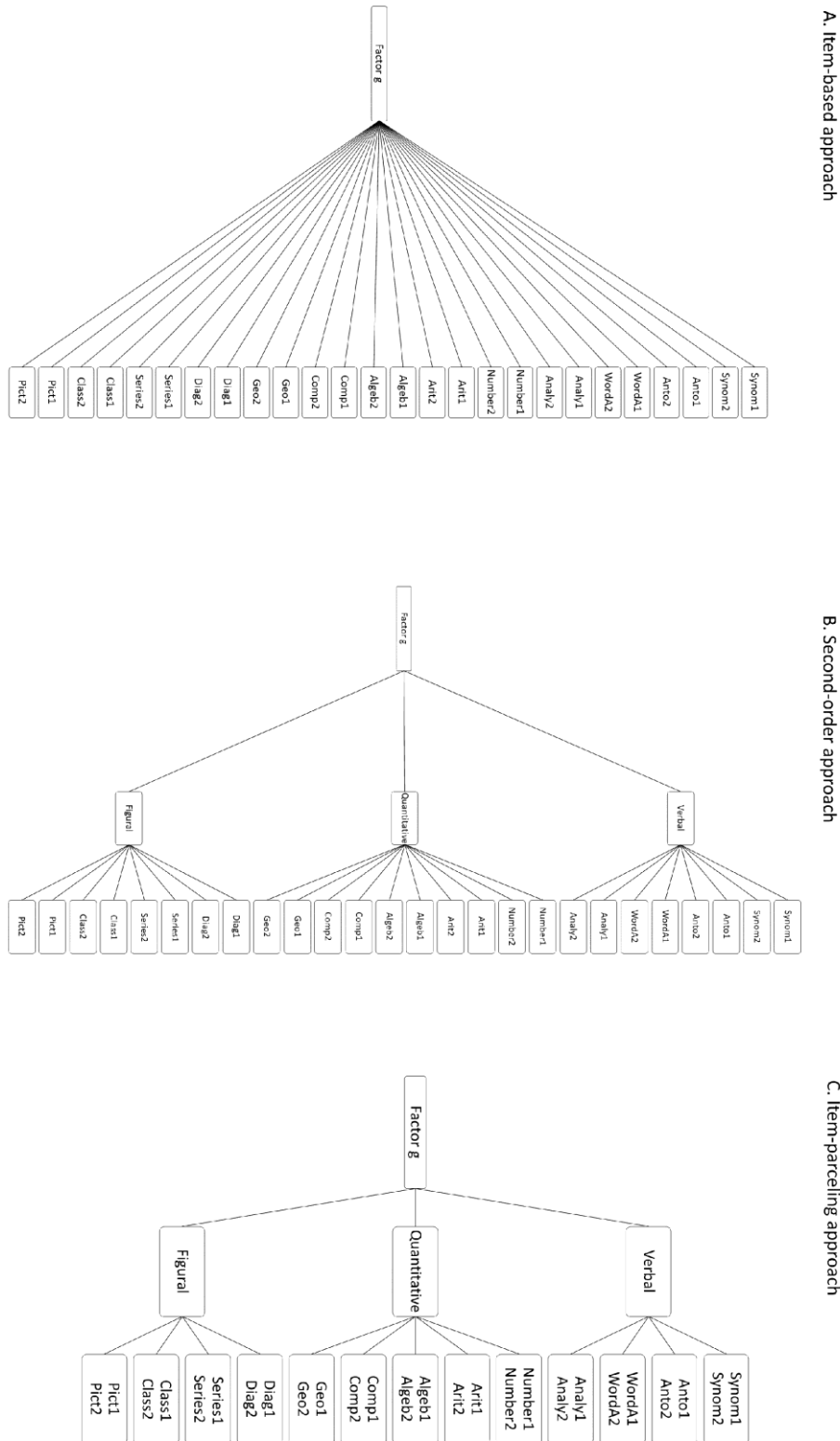


Figure 2. Comparison of Item-Based Approach, Second-Order Approach, and Item-Parceling Approach for Verbal Subtest

There are three types of approaches in this study; item-based approach, second-order approach, and item-based approach. Second-order approach was used in this study because PAPs test is an assessment that measures latent construct of broad general intelligence. Second-order approach were made based on the latent construct measured in PAPs test: verbal, quantitative, and figural. Figure 2 shows the differences between each approach when applied in verbal sub-test. Then, model-fit indexes in SEM analysis on item parceling analysis and individual item analysis will be compared.

Item parcels were calculated by taking the mean of each item in each component. Each component consisted of five to 15 items. There were 13 parcels produced for each test in which each parcel consisted of five until 15 items. The smallest parcel consisted of five items and the largest parcel consisted of 15 items. Data analysis in this study was conducted using Mplus version 7 software.

There are several recommended indices than shows the model fit score: chi-squared test, root mean square error of approximation (RMSEA), comparative fit index (CFI), Tucker-Lewis Index (TLI), and standardized root mean square residual (SRMR) (McDonald & Ho, 2002; Hooper et al., 2008; Kline, 2016). A model can be categorized to have a good model fit when the CFI and TLI score are above 0.90 and RMSEA and SRMR score below 0.08 (Hu & Bentler, 1999).

## FINDINGS AND DISCUSSION

### Findings

Based on Table 2, the descriptive statistics result analyzed from 1.374 sample of PAPs test series E1 shows that the mean of each component ranges from 1.67 to 6.06. Overall, this test has fairly symmetrical data distribution (Synonym, Antonym, Word Analogy, Number Series, Quantitative Comparison, Serial, Classification, and Picture Analogy) to moderately skewed data distribution (Analytical, Arithmetic, Algebra Concept, Geometric Reasoning, and Diagram).

Table 2. Descriptive Statistics of PAPs Test Series E1

Component	Mean	SD	Min.	Max.	Skew.	Kurt.
Synonym	5.05	2.878	0	13	0.380	-0.822
Antonym	4.97	4.648	0	14	0.275	-1.575
Word Analogy	6.06	2.466	0	13	0.292	-0.546
Analytic	1.67	1.411	0	5	0.705	-0.520
Number Series	3.88	3.172	0	10	0.403	-1.373
Arithmetic	3.69	2.201	0	10	0.734	0.081
Algebraic Concept	3.36	2.283	0	10	0.927	0.397
Quantitative Comp.	3.79	1.850	0	10	0.351	0.044
Geometric Reasoning	2.39	1.770	0	9	0.764	0.268
Diagram	3.85	2.289	0	15	0.782	0.853
Series	5.49	3.971	0	15	0.424	-1.194
Classification	3.67	2.230	0	9	0.343	-1.018
Picture Analogy	3.51	1.730	0	9	0.306	0.035

Table 3 shows the descriptive statistics result analyzed from 1.374 sample of PAPs test series E2. Similar with the the E1 series, this test has symmetrical data distribution (Synonym, Analogy, Word Analogy, Analytical, Number Battery, Quantitative Comparison, Serialization, Classification, and Picture Analogy) to moderately skewed data distribution (Arithmetic, Algebra Concept, and Diagram) to highly skewed data distribution (Geometric Reasoning).



Table 3. Descriptive Statistics of PAPs Test Series E2

Component	Mean	SD	Min.	Max.	Skew.	Kurt.
Synonym	4.83	2.935	0	13	0.311	-1.054
Antonym	4.96	4.374	0	14	0.265	-1.515
Word Analogy	6.22	2.209	0	13	0.341	-0.156
Analytic	2.53	1.781	0	5	0.112	-1.432
Number Series	4.57	3.423	0	10	0.374	-1.476
Arithmetic	3.59	2.534	0	10	0.596	-0.757
Algebraic Concept	3.14	2.012	0	10	0.843	0.415
Quantitative Comp.	3.79	2.149	0	10	0.478	-0.403
Geometric Reasoning	2.35	1.826	0	10	1.075	1.345
Diagram	4.53	2.136	0	13	0.746	1.100
Series	5.08	3.801	0	14	0.452	-1.153
Classification	3.93	2.584	0	10	0.293	-1.089
Picture Analogy	4.60	2.563	0	10	0.309	-0.997

### Model Fit

Table 4 shows the result factor analysis of the PAPs test series E1 score. As previously mentioned, the indices used to measure model fit were Chi-square index, CFI, TLI, RMSEA, and SRMR. According to [Hu and Bentler \(1999\)](#), a data can be said to be in accordance with the model if TLI and CFI values  $> 0.90$  and RMSEA and SRMR values  $< 0.08$ .

Table 4. Model Fit Indices for PAPs Test Series E1

Approach	Chi square	CFI	TLI	RMSEA	SRMR
Verbal					
Item-based	2032.966	0.960	0.958	0.023	0.026
Second-order	1978.162	0.962	0.961	0.022	0.026
Item-parceling	21.290	0.994	0.983	0.084	0.015
Quantitative					
Item-based	3758.785	0.842	0.835	0.040	0.045
Second-order	3239.391	0.873	0.867	0.036	0.046
Item-parceling	63.923	0.973	0.945	0.093	0.031
Figural					
Item-based	3118.753	0.892	0.887	0.035	0.038
Second-order	2880.619	0.905	0.900	0.033	0.038
Item-parceling	0.305	0.999	1.003	0.000	0.003

Factor analysis score of PAPs test E1 ( $N = 1.374$ ) in item-based approach showed that CFI and TLI score for verbal subtest are 0.960, and 0.958. RMSEA and SRMR scores were 0.023 and 0.026. From these indices, it could be proved that item-based approach had satisfactory model fit scores for verbal subtest. For quantitative and figural subtest, the CFI and TLI scores of this subtest using item-based approach did not provide the scores required to achieve a good model fit ( $< 0.90$ ). When using second-order approach, CFI and TLI scores of all subtests were higher than item-based approach. RMSEA score on all subtest was also decreases, whereas SRMR score remained the same for verbal and figural subtest and higher for quantitative subtest. Item-parceling approach provided higher CFI and TLI score for all subtest compared to item-based and second-order approach. However, it provided higher RMSEA score for verbal and figural subtest. On the other hand, item-parceling approach provided lower SRMR score for all subtests and RMSEA score for figural subtest.

Table 5 shows factor analysis score result for PAPs test E2 ( $N = 1374$ ). In item-based approach, CFI and TLI scores for quantitative and figural subtest did not exceed the minimum requirement to achieve adequate model fit, whereas the RMSEA and SRMR scores already fulfilled the requirement. Second-order approach in this test were analyzed using AMOS version 22 for figural subtest because the iteration was not found when the analysis was done using

Mplus version 7. It provided higher CFI and TLI scores for all subtest but not adequate for figural subtest. Second-order approach did not give any effect on the RMSEA and SRMR scores of verbal subtest but it gave lower score for quantitative and figural subtests. Item-parceling approach provides adequate CFI, TLI, and SRMR scores to achieve model fit. But it only provided good RMSEA score for figural subtest.

Table 5. Model Fit Indices for PAPs Test Series E2

Approach	Chi square	CFI	TLI	RMSEA	SRMR
Verbal					
Item-based	1909.763	0.969	0.968	0.021	0.025
Second-order	1887.244	0.970	0.968	0.021	0.025
Item-parceling	15.399	0.996	0.989	0.070	0.011
Quantitative					
Item-based	3439.783	0.868	0.863	0.037	0.041
Second-order	2834.764	0.903	0.899	0.032	0.038
Item-parceling	67.575	0.977	0.954	0.095	0.028
Figural					
Item-based	3460.273	0.883	0.879	0.038	0.040
Second-order*	3224.866	0.895	0.890	0.036	-
Item-parceling	6.112	0.998	0.995	0.039	0.008

Note. \*) Analyzed using AMOS 22

Based on the aforementioned descriptions, it can be seen that item parceling produce better output in CFI, TLI, and SRMR score as the score improved compared to second-order approach and item-based approach for both version of the test. However, RMSEA score for both version of the test deteriorates.

## Discussion

Based on the analysis of three model approaches, it can be seen that overall parceling improves several model fit indexes such as chi square value, CFI, TLI, and SRMR in both PAPs test series E1 and E2 compared to item-based approach and second-order approach. In addition, item-parceling approach tends to show a better score on model fit indexes compared to item-based model in both versions of the test for all subtests. The finding of this study corresponds with the previous study which stated that the usage of item parceling can provide a better model fit score (Nasser & Wisenbaker, 2003). De Bruin (2004) stated that parceling possesses unlimited advantages such as higher reliability, richer scale points, better approximation of an interval-scale level, and normality. These advantages can make parceling technique satisfies the assumptions of factor analysis better when it is compared to individual items approach. Balanced item-to-construct parceling was found to improve model fit by reducing the p/f ratios and it also shows a better model fit compared to item-level model (Wilkinson, 2007).

Interestingly, the finding of this study discovers that the RMSEA index for parceling produces bigger number for several components in both versions of the test. RMSEA index for verbal and quantitative in PAPs test series E1 and verbal, quantitative, and figural for series E2 present bigger scores for parceling approach compared to item based and second order approach. This finding is not in line with the previous one which stated that SEM analysis employing parcel based approach resulted in bigger score of CFI and lower score of RMSEA and SRMR (Orcan, 2013). It is generally known that RMSEA is an indicator that shows the difference between observed variance matrix per degree of and the hypothesized covariance matrix which represent the model (Chen, 2007). RMSEA is closely related to degree of freedom, thus it tends to be more sensitive to the complexity of the model and the number of estimated parameters in the model (Byrne, 2012). RMSEA works by seeking a way to balance out the model as measured using noncentrality parameter operationalized as the models degree of freedom by con-



sidering models parsimony. By adding more parameters, the degrees of freedom of the model decreases. If the degree of freedom of a model decreases, RMSEA corrects models by dividing the noncentrality parameter by its degrees of freedom and will create a smaller score (Bandalos, 2018). Even with a correctly specified models, RMSEA can be lower than the cut-offs. It is frequently happened if it has small degree of freedom (Kenny et al., 2014). In the parcelling based approach, the parameter estimated in models might be reduced and produced low degree of freedom. Due to the low number of parameters, there are some variances that cannot be defined because it cannot fit to any parameter. This causes a small number of explained variance. The amount of indicators and parameters in creating a parcel must be properly noted to anticipate this issue. Thus, the parcels can be classified into two: those with one component as represented by an indicator and the ones containing only odd or even items.

Another reason why item parcelling can improve model fit indices in PAPs is because of the test unidimensionality. Unidimensionality of measurement is defined by its internal and external consistency (Anderson, 1984). A unidimensional measurement shows good internal consistency between each component. Each component within a unidimensional structure must have a close relationship in construct validity. PAPs test consists of three components: quantitative, verbal, and figural. A construct validity study of the PAPs test has been conducted before and the result of the study proves that the PAPs test unidimensionality model is supported by the fulfilment of fit indices against established criteria (Widhiarso, 2019). Previous studies have found out that parcelling method is more suitable when it is applied in unidimensional measurement compared to multidimensional measurement. The majority of the research and literature related to parcelling support the evidence that the dimensional nature of a measured construct might have an impact towards the validity and accuracy of parcelling techniques. Item parcelling in unidimensional structure can increase the model-data fit and provides less biased solution for coarsely categorized item with non-normal data distribution (Bandalos, 2002). Many researchers suggest that parcelling should be done only for unidimensional construct (Bandalos & Finney, 2001). It is also stated that item parcelling can effectively improve model fit indices when it is applied to a unidimensional scale (Little et al., 2002). Thus, it is safe to say that unidimensional construct of intelligence measured in PAPs test supports the usage of item-parcelling approach.

The parcelling method in this study is a homogenous parcelling in which each item is aggregated based on their component and shares a similar characteristic with each other. When aggregating items based on their construct, the variance of each parcel is also represented by its component (Cole et al., 2015). The factor loadings of homogenous and heterogenous parcels can impact the model fit indices score. Homogenous parcels tend to show better performance in model fit indices compared to heterogeneous parcels because the factor loadings in homogenous parcels are smaller than heterogenous parcels. When the factor loading is smaller, the chi-square, RMSEA, and SRMR scores get smaller while the CFI and TLI scores get bigger (Cole et al., 2015). Lower chi-square, RMSEA, and SRMR scores and higher CFI and TLI scores become the desired goals when analysing data using CFA. It can help fulfilling the existing criteria for a good model fit. Smaller factor loading can improve the model fit score better because it can reduce the weight of observed covariances (Cole & Preacher, 2014). The rule of thumb of this concept is that the model tends to fit better when the observed covariances is smaller or almost zero. When the observed variance is smaller, it reduces the discrepancies between the observed and implied covariances (Cole et al., 2015). Item parcels tend to substantially increase the power to detect miss-specification within the structural model without affecting the parameter accuracy (Rhemtulla, 2016).

The result of this study indicates different impact of item parcelling between each components. A construct validity study of PAPs by Widhiarso (2019) found that quantitative subtest shows higher factor loading compared to verbal subtest and figural subtest. Amongst all of the components of the test, verbal subtest have the lowest factor loading. This might be caused by the distinctive features of quantitative subtest. Quantitative subtest contains several attributes,

such as thinking systemically as well as using precise strategies and speed in data processing and accuracy. These attributions tend to show small variances between each subject and does not show a huge difference, especially with the lack of cultural influence and verbal interference in the subtest. Those things make quantitative subtest generates an almost uniform way of thinking between each subject. This uniform way of reasoning makes quantitative subtest produces low variation among its data (low data variation). As previously mentioned, smaller variances can blur out the discrepancies between observed and implied covariances (Cole et al., 2015). Thus, it can make parceling approach worked better with homogenous data. The limitation of this study is that this study does not provide information about how sample size can influence the effectiveness of item parceling. This study did not have the alteration of the sample size of PAPs test included in the analysis process. Therefore, the present study cannot dispense comprehend information about item parceling efficacy on different sample sizes.

## CONCLUSION

Overall, it is safe to say that parceling approach is a useful technique for a measurement with unidimensional construct. Item parceling can alleviate some obstacles such as unclear internal consistency, coarsely distributed data, and correlation errors between items. However, when a measurement possesses an unclear construct, parceling might not be suitable to conduct, depends on how a parcel is constructed. Therefore, item parceling is more suitable to be used in CFA rather than EFA. Parceling is primarily a way to obtain the estimation of latent construct and the focus is on accurate estimates of the direct and indirect effects between latent factors. Parceling is commonly used for a long scale that usually consisted of more than 50 of 100 items. In this length of scale, medium-sized sample (example  $N = 200$ ) might not work for parceling as it can create estimation issue. This might be one of the limitations of this study. Since this study does not analyze the impact of parceling methods when applied to different sample size, future study should consider the effect of sample size on the usage of parceling techniques in PAPs test.

## REFERENCES

- Anderson, J. C., & Gerbing, D. W. (1984). The effect of sampling error on convergence, improper solutions, and goodness-of-fit indices for maximum likelihood confirmatory factor analysis. *Psychometrika*, *49*(2), 155–173. <https://doi.org/10.1007/bf02294170>
- Bainter, S. A., & Bollen, K. A. (2014). Interpretational confounding or confounded interpretations of causal indicators? *Measurement: Interdisciplinary Research and Perspectives*, *12*(4), 125–140. <https://doi.org/10.1080/15366367.2014.968503>
- Bandalos, D. L., & Finney, S. J. (2001). Item parceling issues in structural equation modeling. In G. A. Marcoulides & R. E. Schumacker (Eds.), *New developments and techniques in structural equation modeling* (pp. 269–296). Lawrence Erlbaum Associates.
- Bandalos, D. L. (1997). Assessing sources of error in structural equation models: The effects of sample size, reliability, and model misspecification. *Structural Equation Modeling: A Multidisciplinary Journal*, *4*(3), 177–192. <https://doi.org/10.1080/10705519709540070>
- Bandalos, D. L. (2002). The effects of item parceling on goodness-of-fit and parameter estimate bias in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *9*(1), 78–102. [https://doi.org/10.1207/s15328007sem0901\\_5](https://doi.org/10.1207/s15328007sem0901_5)
- Bandalos, D. L. (2008). Is parceling really necessary? A comparison of results from item parceling and categorical variable methodology. *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(2), 211–240. <https://doi.org/10.1080/10705510801922340>

- Bandalos, D. L. (2018). *Measurement theory and applications for the social sciences*. Guilford Press.
- Bastos, J. L., Celeste, R. K., Faerstein, E., & Barros, A. J. (2010). Racial discrimination and health: A systematic review of scales with a focus on their psychometric properties. *Social Science & Medicine*, 70(7), 1091–1099. <https://doi.org/10.1016/j.socscimed.2009.12.020>
- Belinda, B. (2015). *Karakteristik psikometri tes PAPs seri A1*. Undergraduate thesis, Universitas Gadjah Mada, Yogyakarta.
- Birch, L., Fisher, J., Grimm-Thomas, K., Markey, C., Sawyer, R., & Johnson, S. (2001). Confirmatory factor analysis of the child feeding questionnaire: A measure of parental attitudes, beliefs and practices about child feeding and obesity proneness. *Appetite*, 36(3), 201–210. <https://doi.org/10.1006/appe.2001.0398>
- Boomsma, A., & Hoogland, J. J. (2001). The robustness of LISREL modeling revisited. In R. Cudeck, S. du Toit, & D. Sörbom (Eds.), *Structural equation modeling, present and future ER* (Preliminary version, pp. 139–168). Scientific Software International.
- Boomsma, A. (1985). Nonconvergence, improper solutions, and starting values in Lisrel maximum likelihood estimation. *Psychometrika*, 50(2), 229–242. <https://doi.org/10.1007/bf02294248>
- Burisch, M. (1984). Approaches to personality inventory construction: A comparison of merits. *American Psychologist*, 39(3), 214–227. <https://doi.org/10.1037//0003-066x.39.3.214>
- Burns, G. L., Boe, B., Walsh, J. A., Sommers-Flanagan, R., & Teegarden, L. A. (2001). A confirmatory factor analysis on the DSM-IV ADHD and ODD symptoms: What is the best model for the organization of these symptoms? *Journal of Abnormal Child Psychology*, 29, 339–349. <https://doi.org/10.1023/A:1010314030025>
- Byrne, B. M. (2012). *Structural equation modeling with Mplus: Basic concepts, applications, and programming*. Routledge Academic.
- Chen, F. F. (2007). Sensitivity of goodness of fit indexes to lack of measurement invariance. *Structural Equation Modeling: A Multidisciplinary Journal*, 14(3), 464–504. <https://doi.org/10.1080/10705510701301834>
- Chin, W. W. (1998). Commentary: Issues and opinion on structural equation modeling. *MIS Quarterly*, 22(1), vii–xvi. <http://www.jstor.org/stable/249674>
- Cole, D. A., Perkins, C. E., & Zelkowitz, R. L. (2015). Impact of homogeneous and heterogeneous parceling strategies when latent variables represent multidimensional constructs. *Psychological Methods*, 21(2), 164–174. <https://doi.org/10.1037/met0000047>
- Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, 19(2), 300–315. <https://doi.org/10.1037/a0033805>
- De Bruin, G. P. (2004). Problems with the factor analysis of items: Solutions based on Item Response Theory and item parcelling. *SA Journal of Industrial Psychology*, 30(4). <https://doi.org/10.4102/sajip.v30i4.172>
- Drolet, A. L., & Morrison, D. G. (2001). Do we really need multiple-item measures in service research? *Journal of Service Research*, 3(3), 196–204. <https://doi.org/10.1177/109467050133001>
- Gatignon, H. (2016). *Statistical analysis of management data*. Springer.
- Hair, J. F. (2019). *Multivariate data analysis*. Cengage.

- Hancock, G. R. (2003). Fortune cookies, measurement error, and experimental design. *Journal of Modern Applied Statistical Methods*, 2(2), 293–305. <https://doi.org/10.22237/jmasm/1067644980>
- Holbert, R. L., & Stephenson, M. T. (2002). Structural equation modeling in the communication sciences, 1995–2000. *Human Communication Research*, 28(4), 531–551. <https://doi.org/10.1111/j.1468-2958.2002.tb00822.x>
- Hooper, D., Mullen, M. R., & Coughlan, J. (2008). Structural equation modeling: Guidelines for determining model fit. *Electronic Journal on Business Research Methods*, 6(1), 53–60. <https://doi.org/10.21427/D7CF7R>
- Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6(1), 1–55. <https://doi.org/10.1080/10705519909540118>
- Jackson, D. L. (2001). Sample size and number of parameter estimates in maximum likelihood confirmatory factor analysis: A Monte Carlo investigation. *Structural Equation Modeling: A Multidisciplinary Journal*, 8(2), 205–223. [https://doi.org/10.1207/S15328007SEM0802\\_3](https://doi.org/10.1207/S15328007SEM0802_3)
- Kenny, D. A. (1979). *Correlation and causality*. Wiley.
- Kenny, D. A., Kaniskan, B., & McCoach, D. B. (2014). The performance of RMSEA in models with small degrees of freedom. *Sociological Methods & Research*, 44(3), 486–507. <https://doi.org/10.1177/0049124114543236>
- Kline, R. B. (2016). *Principles and practice of structural equation modeling*. Guilford Press.
- Kolenikov, S., & Bollen, K. A. (2012). Testing negative error variances. *Sociological Methods & Research*, 41(1), 124–167. <https://doi.org/10.1177/0049124112442138>
- Ladhari, R. (2010). Developing e-service quality scales: A literature review. *Journal of Retailing and Consumer Services*, 17(6), 464–477. <https://doi.org/10.1016/j.jretconser.2010.06.003>
- Little, T. D., Cunningham, W. A., Shahar, G., & Widaman, K. F. (2002). To parcel or not to parcel: Exploring the question, weighing the merits. *Structural Equation Modeling: A Multidisciplinary Journal*, 9(2), 151–173. [https://doi.org/10.1207/s15328007sem0902\\_1](https://doi.org/10.1207/s15328007sem0902_1)
- Marsh, H. W., Hau, K., & Grayson, D. (2005). Goodness of fit in structural equation models. In *Contemporary psychometrics: A festschrift for Roderick P. McDonald* (pp. 275–340). Lawrence Erlbaum Associates.
- Marsh, H. W., Hau, K.-T., Balla, J. R., & Grayson, D. (1998). Is more ever too much? The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research*, 33(2), 181–220. [https://doi.org/10.1207/s15327906mbr3302\\_1](https://doi.org/10.1207/s15327906mbr3302_1)
- Marsh, H. W., & Hau, K.-T. (2004). Explaining paradoxical relations between academic self-concepts and achievements: Cross-cultural generalizability of the internal/external frame of reference predictions across 26 countries. *Journal of Educational Psychology*, 96(1), 56–67. <https://doi.org/10.1037/0022-0663.96.1.56>
- Marsh, H. W., Lüdtke, O., Robitzsch, A., Trautwein, U., Asparouhov, T., Muthén, B., & Nagengast, B. (2009). Doubly-latent models of school contextual effects: Integrating multilevel and structural equation approaches to control measurement and sampling error. *Multivariate Behavioral Research*, 44(6), 764–802. <https://doi.org/10.1080/00273170903333665>
- Marsh, H. W. (2012). Application of confirmatory factor analysis and structural equation modeling in sport and exercise psychology. In G. Tenenbaum, R. C. Eklund



(Eds.), *Handbook of sport psychology* (pp. 774–798). Wiley.  
<https://doi.org/10.1002/9781118270011.ch35>

- Matsunaga, M. (2008). Item parceling in structural equation modeling: A primer. *Communication Methods and Measures*, 2(4), 260–293. <https://doi.org/10.1080/19312450802458935>
- McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods*, 7(1), 64–82. <https://doi.org/10.1037/1082-989x.7.1.64>
- McGrath, R. E. (2005). Conceptual complexity and construct validity. *Journal of Personality Assessment*, 85(2), 112–124. [https://doi.org/10.1207/s15327752jpa8502\\_02](https://doi.org/10.1207/s15327752jpa8502_02)
- Myers, N. D., Ahn, S., & Jin, Y. (2011). Sample size and power estimates for a confirmatory factor analytic model in exercise and sport. *Research Quarterly for Exercise and Sport*, 82(3), 412–423. <https://doi.org/10.1080/02701367.2011.10599773>
- Nasser, F., & Wisenbaker, J. (2003). A Monte Carlo study investigating the impact of item parceling on measures of fit in confirmatory factor analysis. *Educational and Psychological Measurement*, 63(5), 729–757. <https://doi.org/10.1177/0013164403258228>
- Orcan, F. (2013). *Use of item parceling in structural equation modeling with missing data*. Doctoral dissertation, Florida State University.
- Reilly, T. (1995). A necessary and sufficient condition for identification of confirmatory factor analysis models of factor complexity one. *Sociological Methods & Research*, 23(4), 421–441. <https://doi.org/10.1177/0049124195023004002>
- Rhemtulla, M. (2016). Population performance of SEM parceling strategies under measurement and structural model misspecification. *Psychological Methods*, 21(3), 348–368. <https://doi.org/10.1037/met0000072>
- Roberson, L. L., Aneni, E. C., Maziak, W., Agatston, A., Feldman, T., Rouseff, M., ... Nassir, K. (2014). Beyond BMI: The "Metabolically healthy obese" phenotype & its association with clinical/subclinical cardiovascular disease and all-cause mortality - A systematic review. *BMC Public Health*, 14(1), 14. <https://doi.org/10.1186/1471-2458-14-14>
- Suhr, D. D. (2003). Exploratory or confirmatory factor analysis?. *Statistics and Data Analysis*, 200-31. <https://support.sas.com/resources/papers/proceedings/proceedings/sugi31/200-31.pdf>
- Widhiarso, W. (2019). Pembuktian validitas terkait struktur tes potensi akademik pascasarjana (PAPS) Universitas Gadjah Mada. *Jurnal Psikologi*, 46(2), 145-162. <https://doi.org/10.22146/jpsi.38223>
- Widhiarso, W., & Ravand, H. (2014). Estimating reliability coefficient for multidimensional measures - A pedagogical illustration. *Review of Psychology*, 21(2), 111–121. <https://core.ac.uk/download/pdf/33286765.pdf>
- Widhiarso, W., Azwar, S., Suhapti, R., & Haryanta. (2015). Analisis dan penyempurnaan aitem-aitem tes PAPS seri A1. *Seri Technical Report UPAP*, 2(2), 1–7.
- Wilkinson, T. J. (2007). Assessment of clinical performance: Gathering evidence. *Internal Medicine Journal*, 37(9), 631–636. <https://doi.org/10.1111/j.1445-5994.2007.01483.x>
- Wongpakaran, T., Wongpakaran, N., Sirirak, T., Arunpongpaial, S., & Zimet, G. (2017). Confirmatory factor analysis of the revised version of the Thai multidimensional scale of perceived social support among the elderly with depression. *Aging & Mental Health*, 22(9), 1149–1154. <https://doi.org/10.1080/13607863.2017.1339778>