ICAL
International Conference on Assessment and Learning

TRANSFORMING ASSESSMENT AND LEARNING

13-15 October 2022 | Bali, Indonesia

2022

ACER

Transforming Assessment and Learning: Making the System Work. Proceeding of the 2022 International Conference on Assessment and Learning

ACER Indonesia
Jl. Hang Tuah IX No.9
Gunung, Kebayoran Baru
Jakarta 12120
Indonesia
Phone: +62 21 2751 6999
www.acer.org/id

ISBN 978-1-74286-697-0

Citation APA 7
Lesmini, A. (Ed.). (2022). Transforming Assessment and Learning: Making the System Work. Proceeding of the 2022 International Conference on Assessment and Learning. ACER Indonesia. https://doi.org/10.37517/978-1-74286-697-0

The opinions represented in these articles belong to the authors and this compilation does not warrant the accuracy of any conclusion expressed.

Table of Contents

CONTENTS

Table of Contents

CONTENTS

Table of Contents

CONTENTS

Validity evidence based on content: Controversies and quantification

Agung Santoso
Alice Whita Savira
Robertus Landung Eko Prihatmoko
University of Sanata Dharma

https://doi.org/10.37517/978-1-74286-697-0-01

Agung Santoso is currently an assistant professor in the Psychology Department at the University of Sanata Dharma. He earned his Ph.D. in Quantitative Psychology from the University of Notre Dame in 2019. His research interests are in quantitative methods - particularly in good practice in statistical analysis -, statistical modeling - particularly in measurement models and longitudinal study -, learning and academic analytics, and causal inference. Most of his works are in good practice in statistical analyses, including the ones presented at the recent conference.

Abstract

Validity is a crucial issue in test development because it represents the accuracy of test score interpretation in describing the measured attribute or construct. Among validity evidence that can be collected to support test score interpretation, the one based on test content incited controversies. The content-based validity evidence includes analyses of the relationship between test content and the construct the test purported to measure. Some authors opposed the use of such evidence in the validation process, arguing that the evidence can only provide hypotheses to be tested in the subsequent processes. Other authors argued that the content validation process provided meaningful information regarding test validity, particularly tests based on content such as achievement tests. Unfortunately, literature regarding such issues hitherto has not addressed the controversies adequately. Furthermore, several techniques for quantifying the content-based validity evidence have been proposed without any other article reviewing and evaluating them in comparison with the others. The current study was conducted to examine propositions made by prominent authors regarding validity based on test content and to compare proposed quantification techniques. Using Belfield Academic Search Engine, 1,841 articles with “content validity” in their titles, however only 28 of them met inclusion criteria and thus reviewed. The reviews showed that there were three positions regarding content-based validity evidence: (1) content validity is sufficient as sole evidence supporting test scores interpretation, (2) evidence can be based on the relationship between content and construct but support from other types of evidence is needed, and (3) evaluating representativeness of construct by test content is an important process, but it cannot satisfactorily provide validity evidence. As for the quantification techniques, there were four techniques proposed by different authors. Each focused on a different aspect of the relationship between construct and test content, such as whether or not test content is relevant (e.g., Aiken’s V and Polit’s CVI), essential (Lawshe’s CVR), or discriminant (Dixon and Johnston’s Discriminant Content Validity) to another construct. All but one of the techniques were variations of measures

of agreement between experts regarding how well an item represented a construct. Only one technique used statistics comparing means of judgments on the content relevance to different constructs. Most techniques have been evaluated and revised or corrected by other authors, while Dixon and Johnston's DCV has not because it was just published currently. The impact of the findings was discussed.

Introduction

Validity is a crucial psychometric quality of a test related to the accuracy of test score interpretation (AERA, APA, & NCME, 2014). The concept of validity has evolved from concepts emphasizing correlations of test scores with other variables, tripartite conceptualization, to the unitarian conceptualization of validity. In the unitarian conceptualization, the three types of validity in tripartite conceptualization did not stand by themselves as a discrete concept of validity but rather as classes of evidence to be collected in order to establish the validity of test score interpretation (AERA et al., 2014; Supratiknya, 2014; Santoso, 2010).

One type of evidence that invites a lot of controversy was validity evidence based on test content or known as content validity in tripartite conceptualization. The controversion of content-based validity evidence happened not only at the conceptual level but also at the practical level, particularly on quantification techniques.

At the conceptual level, the position of content-based validity evidence as the validity evidence for test scores interpretation was often questioned (for example Beckstead, 2009; Messick, 1993; Guion, 1977). The authors rejected the content-based validity evidence as the evidence for validity because the definition of validity was related to test score interpretation while content-based validity provided evidence of the representativeness of test domain by test content. The evidence for validity can only be provided by analyzing test scores obtained from field testing, that is by comparing the behavior of test scores from the field testing with the behavior expected or hypothesized from the underlying theory.

Several other authors defended the position of content-based validity as one of the validity evidence (Chalhoub-Deville, 2009; Lissitz, R. W. & Samuelsen, 2007a; 2007b). Content-based validity was considered important in the construction or development of tests with very specific domains such as competence or achievement tests. The validity of test scores interpretation could not be separated from the meaning and representativeness of test content. The class of evidence was also included in the recent *Standards* (AERA et al., 2014). However, there was not any argument made in the *Standards* for maintaining content-based evidence as a class of validity evidence.

Unfortunately, discourses on the controversies were rarely reviewed and published. For example, books on psychological assessment and testing (e.g., Gregory, 2015; Urbina, 2014) did not include such discussion. As a result, knowledge of the evolution of validity conceptualization, particularly of content-based validity, did not reach a bigger audience in psychology. For that reason, the current study was conducted to review and then present the position of content-based validity hitherto.

At the practical level, many authors had proposed techniques for quantifying judgments from experts on content-based validity (Aiken, 1980; Lawshe, 1975; Polit & Beck, 2006;

Johnston et al., 2014). The techniques were proposed because the effort of collecting evidence hitherto was considered superficial (Supratiknya, 2016) or subjective (Crocker & Algina, 2006).

However, to the current authors' knowledge, there was not any publication comparing and evaluating the techniques to find their strengths and limitations. Such knowledge could be used by test developers to choose appropriate techniques. For that reason, the current study also addressed issues related to the strength and limitations of quantification techniques.

Based on the presentation above, there were three questions addressed in the current study:

1. How was the position of content-based validity evidence in the concept of validity?
2. How were the comparison and evaluation of techniques for quantifying validity evidence based on test content?

Methods

Research Design

The current study was a literature review in the measurement area that included content-based validity as the topic of the resources not only methods to provide evidence of the validity of test score interpretation. The literature included topics on techniques for quantifying content-based validity.

Inclusion and Exclusion Criteria

The inclusion criteria used to select resources to be reviewed were:

1. Empirical studies or article reviews on content-based validity evidence and/or its quantification techniques
2. Content-based validity or its quantification was written as the topic of the resources
3. The year of publication was not limited because the author of the current study wanted to capture all arguments supporting or rejecting content-based validity as validity evidence.

The only exclusion criterion was that the content-validity evidence was treated not as the subject matter but only as the method to provide evidence of validity.

Data Collection

The articles reviewed in the current study were searched using Belfield Academic Search Engine (Base) that were then collected by buying them from the publishers. The authors also searched the articles from the bibliography included in reviewed articles. The keyword used in the search engine was *content validity* or *validity based on test content.* The results from the search were then selected following the aforementioned inclusion and exclusion criteria.

Analysis

The selected publications were then analyzed using the narrative thematic technique in the form of conclusions and discussion of findings or presentation in the literature (Higgins & Deeks, 2008; Pare & Kitsiou, 2017; Cronin, et al. 2008). Following the steps proposed by Cronin et al. (2008), the narrative thematic analysis was conducted using the PQRS method consisting of Preview, Question, Read and Summarize.

Results

Description of Resources Reviewed

The authors used the keyword "content validity" in the title column of Belfield Academic Search Engine to search for the articles to be reviewed. As many as 1841 journal articles were found, but only 24 of them met the inclusion criteria. We also add five more articles acquired from reading the materials and three chapters from three books that discussed content validity or validity in general.

Out of 32 resources, 15 of them were published in journals or books dedicated to discussing measurement topics, while 15 were published in substantive journals. It was interesting to find that out of 15 substantive resources, 7 of them were published in journals with nursing topics, while 8 were published in psychology and 2 in education. The results show that the topic of content validity seemed also to be an important issue in nursing research literature. The distribution of journals published articles on content validity topic can be seen in Table 1.

It is also worth noting that although the resources were published in different areas, the discussion on content-based validity evidence was similar. The similarity means that concerns regarding content-based validity evidence were the same across areas and may therefore across types of tests or scales.

Table 1. Distribution of journals that published articles on content validity

Title of Journals	Number of Articles
Measurement	
Measurement and Evaluation in Counseling and Development	2
International Journal of Selection and Assessment	2
Applied Psychological Measurement	2
Journal of Educational Measurement	1
Educational and Psychological Measurement	5
Book chapters on content validity	3
Psychology	
Industrial and Organizational Psychology	1
Personnel Psychology	3
British Journal of Health Psychology	2
Psychological Review	1
Psychological Reports	1
Educational	
Educational Researcher	2
Nursing	

Title of Journals	Number of Articles
International Journal of Nursing Studies	1
Nursing Research	1
Contemporary Nurse	1
Research in Nursing & Health	2
International Journal of Nursing Terminologies and Classifications	1
Western Journal of Nursing Research	1
Total	**32**

Controversies of Content Validity

There were three positions on the issue of whether content-based validity evidence could be considered as validity evidence. The first one argued that content-based validity evidence, or content validity as its proponent called it, was a different and independent type of validity. It provided information regarding validity of test score interpretation and was sufficient by itself. For example, Lizzits and Samuelsen (2007a) argued that content validity was essential to the evaluation of a test that was constructed based on a clear definition of the content domain such as the one in educational tests. The content validity, together with reliability, was then evaluated based on internal information from the test. Other authors that also had the same positions were, for example, Schmidt (2012) and Chalhoub-Deville (2009).

Furthermore, Lizzits and Samuelsen (2007a) stated that the content validity evaluated the adequacy of the process of constructing the test. Such a process could be applied to the creation of tests measuring a clear domain of interest as well as a latent construct. It is important to note that Lizzits and Samuelsen rejected the unitary concept of validity proposed by Messick (1989) and established their own concept.

Another proponent, Ebel (1956), stated more strongly that content validity was the basis for establishing other types of validity. For example, concurrent-criterion validity could be established by evaluating the correlation between the test under investigation and another test measuring the same construct that has already been proven to be valid. The availability of another valid test could only be justified using content validity because the search for a valid test without content validity would make the effort circular. Such a need for content validity was also mentioned by Lennon (1956) in a more moderate manner.

However, it is worth noting that both Ebel and Lizzits & Samuelsen (2007a) ignored the fact that the evaluation of the internal structure of a test and participants' response to items were two kinds of evidence that could also be provided without the need of other tests, thus lessen the need for content validity. Both kinds of evidence could also be obtained by evaluating internal information of the test under scrutiny (Embretson, 2007).
The second position can be seen as the direct opposite of the first position in that the authors rejected the representativeness of test content as validity evidence. The

rejection did not mean to diminish the importance of the process to evaluate the connection between test content and the construct intended to be measured. It was only that such a process was not considered enough to provide support for test score interpretation.

Messick (1989, 1993) was the prominent author who promoted the idea of the unitarian validity concept, in which all other types of validity were treated as types of evidence supporting the validity of test scores interpretation. Messick's opinion on content validity was actually not new. Several other authors such as Mosier (1947), Cureton (1951), Loevinger (1965), Guion (1977, 1978), and Tenopyr (1977) have already asserted that the process of evaluating the representativeness of construct by the test content could not be regarded as evidence of validity. Such a process only provided a hypothesis that needed empirical testing to support it. A more recent publication on job performance measures by Murphy (2009) showed that there were weak relationships between results from the content-based validation process and other validation processes. Other more current publications that also had the same position were Beckstead (2009a, 2009b).

The other problem of treating content representativeness as evidence for validity was that the process participants intended to conduct to arrive at a certain response may not be the same as the actual process the participants carried out (Messick, 1989). For example, the influence of social desirability and other response styles on the way test takers responded to the stimulus might not be identified in the evaluation of test content.

Although the authors in the second position rejected content representation as evidence of validity, they still considered the process of test content evaluation against the test domain was still extremely important. The importance of the process was due to the influence and limits of the nature and the content coverage imposed on the interpretation of test scores supported by other evidence.

The third group of authors posited a more moderate manner on content-based validity evidence. The authors' position was a middle ground between the first and second positions. While acknowledging the role of content representativeness in providing valid evidence, they pointed out the need of other types of evidence to establish the validity of test score interpretation. For example, Lennon (1956) stated that content-based evidence was needed when there was not any "dependable criterion variable" that could be readily accessed and when correlations with concurrent or future criteria were not meaningful as indicators of validity (p.297). However, Lennon also stated that evaluation of the representativeness of the content of the tests was not sufficient to establish validity because other information, that is the process the test takers employed to arrive at their response, was also needed.

Embretson (2007) also provided a moderate position regarding the content-based validity evidence. In the article, Embretson first showed the limitation of content validity in establishing the validity of test scores interpretation. For example, content-based validity evidence did not provide any evidence that test takers employed the skills, knowledge, or processes supposedly used in responding to test items. However, Embretson then included content-based validity evidence in eleven categories of evidence, particularly in Test Specification, Item Design Principles, and Domain Structure. Embretson argued that such evidence was crucial because the emphasis of

providing evidence of validity should be on the internal evidence not on nomological networks of constructs. Similar opinions were also held by other authors such as Sackett (2012) and also the new *Standards* (AERA, APA, & NCME, 2014).

We concluded that the first position was indefensible for several reasons. First, the process of content validation could only provide hypotheses about the relationship between test content and the construct underlying the test. Next, the content validation process could not provide sufficient evidence for the process presumed to be employed by test takers to arrive at certain responses. Furthermore, construct under-representativeness and construct irrelevant variance could only be identified by analyzing the test or item scores, not by evaluating the representativeness of the test content.

The third position was appealing because it provided a middle ground that accommodated both the first and second positions. However, the authors of the current article could not find any strong and fundamental argument for incorporating content representativeness as one type of evidence supporting the validity of the test scores interpretation. The only argument that we could find was in Lennon (1956). However, the articulated reason for conducting content validity was considered obsolete because there have been many methods developed to provide evidence for validity without the need for external criteria since. Therefore, the current authors chose to side with the second position in which the process of evaluating content representativeness was considered crucial but could not be considered as valid evidence.

However, we also found that the arguments posed by each position regarding content-based validity evidence could not be separated from the conceptualization of validity in general. For example, all authors in positions two and three believed that validity was attached not to the test but to the test score interpretation, therefore it was the score of the test that should be treated as the source of information regarding validity. While the prominent authors in position one believed otherwise. Therefore, the evaluation of the content validity position should be re-reviewed in a broader scope covering the conceptualization of validity in general. In turn, the conceptualization of validity itself could not be separated from how the measurement was conceptualized, therefore the future study should also include the philosophy underlying the definition of measurement.

Quantification of Content Validity

Although the current authors have chosen to side with the second position, issues regarding the quantification of content representativeness were still relevant. The quantification methods could be used in the process of establishing the relationship between test content and the test's underlying construct.

As far as the current authors could find, there were four methods for quantifying the content-representativeness that were presented in Table 2. The methods evaluated different aspects of content-representativeness. Lawshe's CVR, for example, evaluated whether an item was 'essential' to represent the construct measured, while Polit, Beck & Owen's CVI (2007) emphasized the item's relevance to the test. Aiken's V (1980) did not clearly identify what aspect of representativeness it measured. The author only mentioned that the judge needed to evaluate the content validity of the test. The newest

method proposed was the Discriminant Content Validity which evaluated how well an item represented the underlying construct and the extent to which test items represented a different construct.

It was important to mention here that all but one, the discriminant content validity, has already been evaluated by different authors and shown that there were some problems regarding their inference. Fortunately, those articles also provided the revision of the formula for either calculating the standard error of the statistics or drawing conclusions regarding the results. And based on the accuracy of the results from the revisions, the current authors strongly recommend their use.

Table 2. Comparison of quantification techniques for content-based validity

	Lawshe	Aiken	Polit, Beck & Owen	Dixon & Johnston
Source	Lawshe (1975)	Aiken (1980)	Polit, Beck & Owen (2007)	Johnston et al.. (2014)
What is evaluated	"Essential"	"Content Validity"	"Relevance"	"Discrimination"
How is the quantification conducted	Proportion of the essential judgment	Proportion of the ratings mean	inter-rater agreement on relevance	differences of relevance across constructs
Inference*	unclear	Multinomial or approximation of Normal Distribution	Inference of Fleiss Kappa	Mean difference tests such as t-test or Wilcoxon
Revision	Wilson, et al. (2012); Baghestani, et al. (2019)	Penfield et al. (2009)	Beckstead (2009a)	NA

Note: *Methods of inference found in the original article.

Conclusion

The first conclusion drawn from the discussion above was that the current author tended to agree with the second position regarding the content-based validity evidence, that content representativeness should not be considered as validity evidence. The current authors acknowledged the importance of the evaluation content representativeness based on the underlying construct therefore such a process should be still conducted in the explication processes or prior to collecting field testing data.

However, the current author realized that the discussion regarding content-based validity evidence could not be addressed comprehensively without addressing the validity conceptualization that in turn could not be separated from the conceptualization of measurement. Therefore, the current author encouraged such discussion in future

studies to provide a more comprehensive perspective on content-based validity evidence.

Regarding the quantification methods, we concluded that the use of the methods could still be continued as the process of establishing relationships between test content and the underlying construct should still be practiced. However, one should use the more recent formula or the revision of the original formula to provide a more accurate estimate and inference.

References

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing*. Washington, D.C: Amer Educational Research Assn.

Aiken, L. R. (1980). Content validity and reliability of single items or questionnaires. *Educational and Psychological Measurement*, 40 (4), 955–959. DOI:10.1177/001316448004000419

Baghestani, A. R., Ahmadi, F., Tanha, A., & Meshkat, M.(2019). Bayesian critical values for lawshe's content validity ratio. *Measurement and Evaluation in Counselling and Deveopment*, 52(1), 69 - 73. DOI:10.1080/07481756.2017.1308227

Beckstead, J. W. (2009a). Content validity is naught. *International Journal of Nursing Studies*, 46 (9), 1274–1283. DOI:10.1016/j.ijnurstu.2009.04.014

Beckstead, J. W. (2009b). On content validity: A response to Squires (2009). *International Journal of Nursing Studies*, 46 (9), 1286–1287. DOI:10.1016/j.ijnurstu.2009.04.014

Benson, J.(1981). A redefinition of content validity. *Educational and Psychological Measurement,* 41(3), 793 - 802

Chalhoub-Deville, M. (2009). Content validity considerations in language testing contexts. In R. W. Lissitz (Ed.), *The concept of validity: revisions, new directions, and applications* (pp. 241–263). Charlotte, NC: Information Age Publishing, Inc.

Crocker, L., Llarbre, M., & Miller, M. D., (1988). The generalizability of content validity ratings. *Journal of Educational Measurement*, 25(4), 287 - 299.

Crocker, L., & Algina, J. (2006). Introduction to Classical and Modern Test Theory (1st edition ed.). Mason, Ohio: Cengage Learning.

Cronin, P., Ryan, F., & Coughlan, M. (2008, January). Undertaking a literature review: a step-by-step approach. British Journal of Nursing, 17 (1), 38–43. Retrieved 2022-02-05, from http://www.magonlinelibrary.com/doi/10.12968/bjon.2008.17.1.28059 doi: 10.12968/bjon.2008.17.1.28059

Dixon, D., & Johnston, M. (2019). Content validity of measures of theoretical constructs in health psychology: Discriminant content validity is needed. *British Journal of Health Psychology,* 24, 477 - 484. DOI:10.1111/bjhp.12373

Ebel, R. L. (1956). Obtaining and reporting evidence on content validity. *Educational and Psychological Measurement*, 16(3),269 - 282. DOI:10.1177/001316445601600301

Embretson, S. E. (2007). Construct Validity: A Universal Validity System or Just Another Test Evaluation Procedure? Educational Researcher, 36(8), 449-455.

Fitzpatrick, A. R.(1983). The meaning of content validity. *Applied Psychological Measurement*, 7(1),3 - 13. DOI:10.1177/014662168300700102

Gregory, R. J. (2015). Psychological testing: History, principles, and applications (7th ed.). Essex: Pearson Education Limited.

Guion, R. M. (1977). Content Validity: The Source of My Discontent. Applied Psychological Measurement, 1 (1), 1 – 10.

Guion, R. M. (1978). "Content validity" in moderation. *Personnel Psychology*, 31(2), 205 - 213.

Higgins, J. P., & Deeks, J. J. (2008). Selecting studies and collecting data. In J. P. Higgins & S. Green (Eds.), Cochrane handbook for systematic reviews of interventions (pp. 151 – 185). Chichester, West Sussex, UK: John Wiley & Sons, Inc.

Johnston, M., Dixon, D., Hart, J., Glidewell, L., Schr¨oder, C., & Pollard, B. (2014, May). Discriminant content validity: A quantitative methodology for assessing the content of theory-based measures, with illustrative applications. *British Journal of Health Psychology*, 19 (2), 240–257. DOI:10.1111/bjhp.12095

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28 (4), 563–575. DOI: 10.1111/j.1744-6570.1975.tb01393.x

Lennon, R. T. (1956). Assumptions underlying the use of content validity. *Educational and Psychological Measurement*, 16(3), 294 - 304. DOI:10.1177/001316445601600303

Lissitz, R. W., & Samuelsen, K. (2007a). Further Clarification Regarding Validity and Education. *Educational Researcher*, 36 (8), 482–484. DOI:10.3102/0013189X07311612

Lissitz, R. W., & Samuelsen, K. (2007b). A Suggested Change in Terminology and Emphasis regarding Validity and Education. *Educational Researcher*, 36 (8), 437–448.

Loevinger, J.(1957). Objective tests as instruments of psychological theory. *Psychological Reports*, 3, 635 - 694. DOI: 10.2466/pr0.1957.3.3.635

Loevinger, J. (1965). Person and population as psychometric concepts. *Psychological Review,* 72, 143-155. DOI: 10.1037/h0021704

Lynn, M. R. (1986, November). Determination and quantification of content validity. *Nursing Research*, 35 (6), 382 – 386. DOI:10.1097/00006199-198611000-00017

Mastaglia, B., Toye, C., & Kristjanson, L. J. (2003). Ensuring content validity in instrument development: Challenges and innovative approaches. *Contemporary Nurse*, 14(3), 281 - 291. DOI:10.5172/conu.14.3.281

Messick, S. (1989). Validity. In L. R L (Ed.), *Educational Measurement* (3rd ed., pp. 13 – 103). New York, NY: American Council on Education dan Macmillan.

Messick, S. (1993). Foundations of Validity: Meaning and Consequences in Psychological Assessment. *ETS Research Report Series*, 1993 (2), i–18. doi: 10.1002/j.2333-8504.1993.tb01562.x

Michell, J. (2009). Invalidity in Validity. In R. W. Lissitz (Ed.), *The concept of validity: Revisions, New Directions, and Applications* (pp. 111–133). Charlotte, NC: Information Age Publishing, Inc.

Murphy, K. R. (2009). Content validation is useful for many things, but validity isn't one of them. *Industrial and Organizational Psychology*, 2(4), 453 - 464. DOI: 10.1111/j.1754-9434.2009.01173.x

Par´e, G., & Kitsiou, S. (2017). Methods for Literature. In F. Lau & C. Kuziemsky (Eds.), *Reviews Handbook of eHealth Evaluation: An Evidence-based Approach*. University of Victoria.

Polit, D. F., & Beck, C. T. (2006, October). The content validity index: Are you sure you know what's being reported? Critique and recommendations. *Research in Nursing & Health*, 29 (5), 489–497. DOI: 10.1002/nur.20147

Polit, D.F., Beck, C. T., & Owen, S. V. (2007). Is the CVI an acceptable indicator of content validity? Appraisal and recommendations. *Research in Nursing & Health*, 30(4), 459 - 467. DOI:10.1002/nur.20199

Sackett, P. R. (2012). Cognitive tests, constructs, and content validity: A commentary on Schmidt (2012). *International Journal of Selection and Assessment,* 20(1), 24 - 27. DOI: 10.1111/j.1468-2389.2012.00576.x

Santoso, A. (2010). Validitas: Dulu dan Sekarang. In A. Supratiknya & T. Susana (Eds.), *50 Tahun Himpunan Psikologi Indonesia: Redefinisi Psikologi Indonesia dalam Keberagaman*, (pp. 333 – 346). Himpunan Psikologi Indonesia (HIMPSI).

Schmidt, F. L.(2012). Cognitive tests used in selection can have content validity as well as criterion validity: A broader research review and implications for practice. *International Journal of Selection and Assessment*, 20(1), 1 - 13. DOI:10.1111/j.1468-2389.2012.00573.x

Supratiknya, A. (2014). *Pengukuran Psikologis*. Penerbit Universitas Sanata Dharma.

Supratiknya, A. (2016). *Kuantifikasi validitas isi dalam asesmen psikologis*. Penerbit Universitas Sanata Dharma.

Tenopyr, M. L (1977). Content-construct confusion. *Personnel Psychology,* 30(1), 47 - 54. DOI: 10.1111/j.1744-6570.1977.tb02320.x

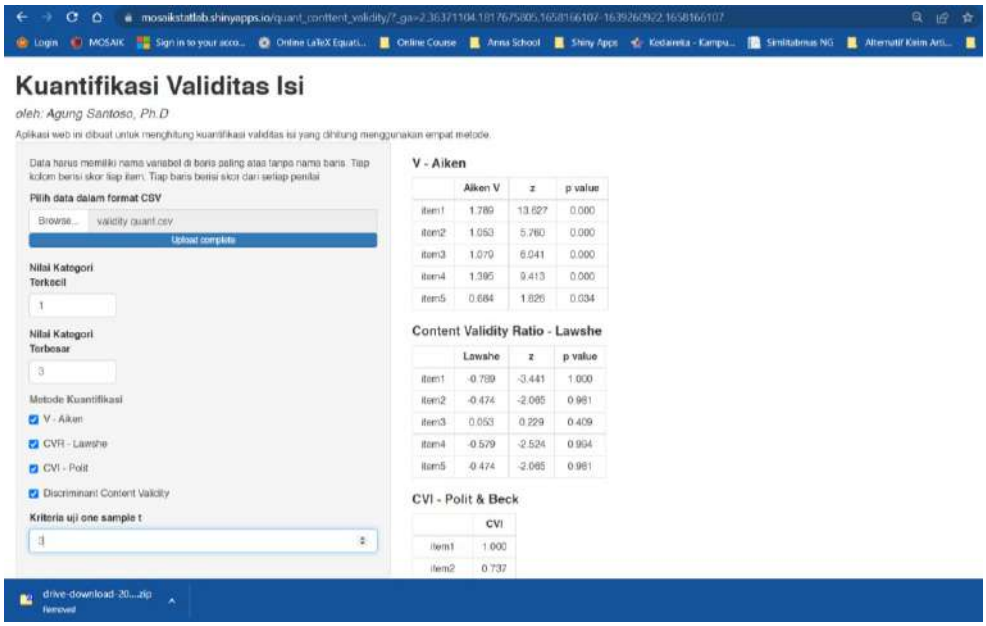
Urbina, S. (2014). *Essentials of psychological testing*. John Wiley & Sons, Incorporated.

Wilson, F. R., Pan, W., & Schumsky, D. A. (2012, July). Recalculation of the critical values for Lawshe's content validity ratio. *Measurement and Evaluation in Counseling and Development*, 45 (3), 197–210. doi: 10.1177/0748175612440286

Wynd, C. A., Schmidt, B., Schaefer, M. A. (2003). Two quantitative approaches or estimating content validity.

Appendix

Web Application for Quantifying Content-Based Validity Evidence
https://mosaikstatlab.shinyapps.io/quant_conttent_validity/?_ga=2.36371104.1817675805.1658166107-1639260922.1658166107



Keywords: content validity, validity evidence, validity based on content, quantification of content validity,

Relationship between students' economic, cultural, and social status, school climate and student achievement in Indonesia

Alejandra Osses-Vargas
Ursula Schwantner
Raymond J. Adams
Nicole Wernert

Australian Council *for* Educational Research, Australia

https://doi.org/10.37517/978-1-74286-697-0-02

Alejandra Osses-Vargas is a sociologist with postgraduate studies in education. Student learning assessment is her main area of interest, with a focus on instrument development and factors explaining outcomes inequalities. Alejandra's career focus is on providing empirical evidence that contributes to inform education policy and support organisations in areas related to assessment and evaluation. Alejandra has worked at the University of Chile and provided consultancy services to other tertiary education institutions in Chile. Currently, she is a Research Fellow at the Australian Council for Educational Research (ACER).

Abstract

PISA 2018 data shows that, in Indonesia, the relationship between students' economic, cultural, and social status and their scores on reading achievement is weak. However, there is great variation between schools in average reading achievement with over 40% of the explainable variation in reading achievement being between-schools. In addition, in Indonesia, there is low social diversity across schools (students within schools are likely to be of similar economic, cultural, and social background). These conditions raise the question of whether school factors play a role – specifically whether school climate has a compensating, mediating or moderating effect on the relationship between students' and schools' economic, cultural, and social status and achievement. Using regression analysis, it was found that, in Indonesian schools, the composition of student body in terms of economic, cultural, and social status plays a significant role in explaining differences in reading achievement between schools. School climate was found to compensate and mediate the relationship between school economic, cultural, and social status and school reading achievement, particularly those dimensions related to classroom climate, student respect for diversity, school discrimination climate, student sense of belonging and teacher support and directed instruction practices. Only student sense of belonging was found to play moderating role in the relationship between school economic, cultural, and social status and school reading achievement, a finding that requires further investigation.

Introduction

The relationship between students' economic, cultural, and social status and learning outcomes is a common and well-documented subject in educational research (see, for example, Berkowitz et al., 2017; Broer et al., 2019; Hattie, 2009; O'Connell, 2019). In Indonesia, this relationship is weak and students' economic, cultural, and social status only explains a small part of the variation in reading achievement between students (OECD, 2019b). Based on this finding, Indonesia is classified as an equitable education system. PISA results also show that Indonesian schools significantly differ in their average reading achievement. In fact, Indonesia is among the countries where over 40% of the explainable variance in student achievement is between schools (OECD, n.d.).

The additional fact that Indonesia is among the countries that exhibit the lowest levels of social diversity across schools (OECD, 2019c) sparks the interest for investigating whether school factors, such as school climate, have a compensating, mediating or moderating effect on the relationship between students' and schools' economic, cultural, and social status and achievement. That is, it may be that school climate has an additive positive contribution to student achievement, compensating for negative school's economic, cultural, and social status. In contrast, it may be more the case that school's economic, cultural, and social status determines school climate (for example, in schools serving students from high economic, cultural, and social status bullying rarely occurs), thus school climate mediates the relationship between economic, cultural, and social status and student achievement. Finally, school climate may moderate the relationship between economic, cultural, and social status and student achievement, such that the strength of the relationship between economic, cultural, and social status and student achievement varies as school climate varies (Berkowitz et al., 2017).

Understanding the contribution of economic, cultural, and social status and school climate in student achievement are central topics of interest in the newly established Indonesian Minimum Competency Assessment (Pusat Asesmen Pendidikan, 2022). Therefore, this study focuses on understanding the role of students' perception of school climate in the relationship between economic, cultural, and social status and student achievement in Indonesia. This can be valuable evidence to consider in the design of policy actions aimed at improving equity of educational outcomes between schools in Indonesia.

School Climate

School climate has been a construct of interest since the early 1960s (Wang & Degol, 2016). School climate research has arisen from two different research domains: that of organisational psychology (focussing on teachers; for example, Halpin & Croft, 1963) and school and classroom effectiveness research (focussing on students; see (Reynolds et al., 2014). The interest in school climate (especially that in the school effectiveness domain) exists due to a belief that the climate of a school has influence, directly or indirectly, on the outcomes (both achievement and affective-behavioural) of students.

Many studies have found that elements of school climate are related to student achievement, starting with the seminal work of Brookover and colleagues (Brookover et al., 1978) who found that student and staff perceptions of social norms and expectations in the school were related to student achievement. Other elements of school climate that

have been found to be related to student achievement are academic optimism (Hoy et al., 2006); school safety and experiences of bullying (Gronna & Chin-Chance, 1999; Juvonen et al., 2011); teacher support (Lee, 2012; Wang & Holcombe, 2010); and a sense of school belonging or identification (Wang & Holcombe, 2010), for example.

However, to date, there is no standard definition of school climate, with many different conceptualisations and ways in which it is defined and operationalised. Reviews of school climate research have concluded that school climate is multi-dimensional, covering domains such as perceptions of safety and discipline; academic expectations; social relationships; school facilities; and school connectedness (Berkowitz et al., 2017; Cohen et al., 2009; Freiberg, 1999; Wang & Degol, 2016; Zullig et al., 2010). In addition, school climate research should be based on multiple informants - students, teachers or other members of the school community - to provide the perceptions of school climate (Wang & Degol, 2016). However, school climate research continues to vary from study to study in both the way school climate is operationalised and who the informants are. Therefore, it is important that the definition and chosen informants are clear for each study of school climate.

For the current study, the definition of school climate that has been used is that used in PISA 2018. In PISA 2018, school climate is conceptualised as a multi-dimensional construct representing different aspects of the school experience (OECD, 2019c, pp. 37–38) grouped into three broad spheres:

1. *Student Disruptive Behaviour*, encompassing physical and emotional security of school members, disciplinary climate, and frequency of student disruptive behaviour.
2. *Teaching and Learning*, encompassing classroom practices and teacher behaviours that shape the experience of learning and promote socio-emotional development of children.
3. *School Community,* encompassing the nature of the relationships that students, teachers, the school principal, parents, and the local community establish within the school setting.

Method

This study investigated the role of students' perceptions of school climate in the relationship between economic, cultural, and social status and student reading achievement in Indonesia using PISA 2018 data. PISA 2018 was administered in 79 countries to country-representative samples of 15-year-old students enrolled in school, regardless of the grade they were studying. The Indonesian sample comprised 12,008 students distributed across 397 schools.

PISA is a triennial survey that assesses the extent to which students near the end of their compulsory education have acquired knowledge and skills deemed essential for full participation in social and economic life (OECD, 2019a). PISA assesses students' reading, mathematics and science literacy and administers contextual questionnaires to students and school principals to gather information about the environment in which students live and learn.

Indicators used in the study

Reading achievement

The current study used students' scores on reading literacy as the outcome variable. Reading literacy is defined "students' capacity to understand, use, evaluate, reflect on and engage with texts in order to achieve one's goals, develop one's knowledge and potential, and participate in society." (OECD, 2019a, p. 27). For each student, PISA database provides 10 plausible values of reading achievement which are combined for analyses following specifications provided by the OECD (2009). School average reading achievement was calculated from students' scores, also following OECD (2009) specifications.

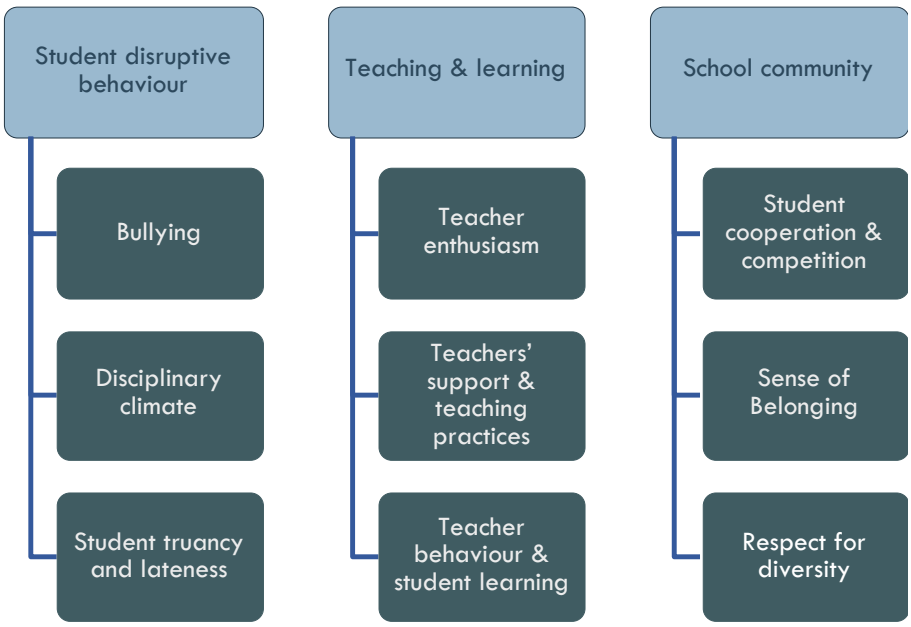
Economic, cultural, and social status

Students' economic, cultural and social status is defined in PISA as a "measure of students' access to family resources (financial capital, social capital, cultural capital and human capital) and the social position of the student's family/household" (OECD, 2019b, p. 52). Operationally, the PISA index of economic, social and cultural status (ESCS) is derived from information relating to students' family: parents' education and occupation, and household possessions. The ESCS score estimated by the OECD for each student, as provided in the PISA database, was used in the current study. School average ESCS (Schl_ESCS) was computed using student-level data.

Students' perception of School climate

The current study used PISA's definition of school climate – a multidimensional construct representing different aspects of the school experience. PISA identifies four spheres on which school climate can be operationalised: safety, teaching and learning, school community, and institutional environment (OECD, 2019c, pp. 37–38). The first three were used in this study and operationalised using indicators in Figure 1. Definition and specific items are detailed in the Appendix (Table 3).

Figure 1. Operationalisation of students' perception of School climate



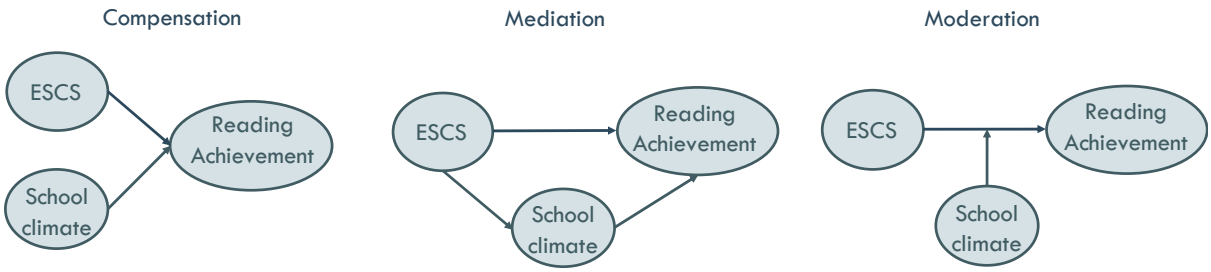
Analysis

Before analysing the role of school climate in the relationship between economic, cultural, and social status and student achievement, the theoretical model of school climate proposed by PISA was tested for Indonesia using confirmatory factor analysis. An alternative model was also tested to explore whether a different factor structure could better account for the relationship between the analysed school climate indicators in Indonesia.

Three models were tested to analyse the role of school climate in the relationship between economic, cultural, and social status and student achievement: a compensation, a mediation, and a moderation model (see Figure 2). Models were specified at the student and school levels using regression analyses that recognise the nested structure of educational data (Raudenbush & Bryk, 2002) – using replicate weights for computing standard errors (OECD, 2009). Results were compared between models.

The compensation model was used to explore whether a positive school climate contributes value to reading achievement on top of that contributed by ESCS. The mediation model investigated whether ESCS determines the level of school climate, which then influences achievement. The moderation model explored whether the relationship between ESCS and reading achievement is different for different levels of school climate.

Figure 2. Models used in analysis

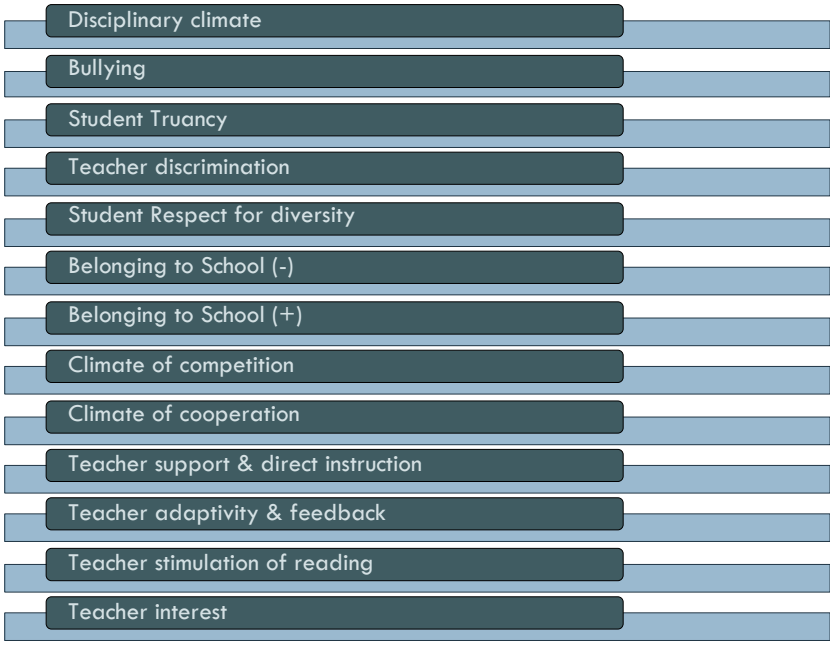


Results

The model of school climate for Indonesia

The three-spheres model of school climate theorised by PISA explained 31% of the variance in the school climate latent variable for Indonesia data. An exploratory factor analysis model was specified to investigate how indicators of school climate relate in Indonesia. The resulting model, which explained 65% of variance in the school climate latent variable, suggested 13 dimensions (see Figure 3).

Figure 3. School climate model for Indonesia



*** Belonging to school (-) refers to negative worded items and belonging to school (+) refers to positive worded items. See detail of items in the Appendix*

It was found that when indicators of school climate are combined in the three theoretical spheres/dimensions suggested by PISA, the effect of individual indicators on reading achievement is masked. For instance, the sphere of *Student disruptive behaviour* – comprised by disciplinary climate, bullying and student truancy in the PISA theoretical model (see Figure 1), had a non-significant correlation to reading achievement. However, in the 13-dimension model, *Disciplinary climate*, *Bullying* and *Student Truancy* showed significant and varying correlations to achievement (.23, -.04 and -.04, respectively).

In the PISA model, the effect of individual indicators within a dimension are cancelled. The 13-dimension model provides more detail about the effect of specific school climate dimensions on school achievement, which is of more interest from a policy perspective. Therefore, the 13-dimension model was retained for analyses.

The initial model

School averages of ESCS and school climate dimensions were computed and three models at the student level were specified as in equations 1 to 3.

$$Reading\ achievement = Intercept + Std.\ ESCS \quad \text{equation 1}$$

$$Reading\ achievement = Intercept + Std.\ ESCS\ + Schl.ESCS \quad \text{equation 2}$$

$$Reading\ achievement = Intercept + Std.\ ESCS\ + Schl.ESCS + Schl.climate \quad \text{equation 3}$$

Results of fitting each of these models are displayed in Table 1. Model 1 (corresponding to equation 1) shows that students' ESCS plays a minor role in explaining differences in students' reading achievement – only 8% of the variation in reading achievement is explained by this variable. Results indicate that students in schools with higher average ESCS achieve higher reading scores and that the effect of school average ESCS cancels the effect of student ESCS on reading achievement (Model 2). The inclusion of school ESCS accounts for an additional 10% of the variation in student reading achievement.

Table 1. Initial student-level models

STUDENT LEVEL MODELS

	Model 1	Model 2	Model 3
(Intercept)	401.08 ***	442.10 ***	373.13 ***
	(4.72)	(6.16)	(9.41)
ESCS	19.14 ***	1.62	1.62
	(2.20)	(1.06)	(1.06)
Schl_ESCS		43.70 ***	15.16 ***
		(3.85)	(3.07)
schl_disclima			11.23 *
			(5.63)
schl_bullying			-2.19
			(12.07)
schl_truancy			-11.63
			(8.88)
schl_Tdiscrim			-38.07 ***
			(5.69)
schl_Srespect			35.58 ***
			(6.08)
schl_Sbelneg			37.30 ***
			(8.86)
schl_Sbelpos			-44.34 ***
			(11.78)
schl_Scompete			6.32
			(8.03)
schl_Scooper			-7.39
			(7.41)
schl_TSupDirin			-22.55 **
			(7.95)
schl_TAdapFeed			-8.29
			(10.33)
schl_TstimRead			3.06
			(10.78)
schl_Tinterest			-11.70
			(7.32)
R2	0.08	0.18	0.39
R2.adj	0.08	0.18	0.39
Num. obs.	12008	12008	12008
Deviance			
Dispersion			

*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1

All variables with the "schl_" suffix correspond to school averages of students related variables. See Appendix for abbreviations and the items corresponding to each indicator of school climate.

The inclusion of school indicators related to students' perception of school climate accounts for an additional 21% of the variation in student reading achievement (Model 3). School-level indicators of students' perception of disciplinary climate in language lessons, students' respect for diversity, and students' sense of belonging expressed through negatively worded items show a positive and statistically significant relationship to student reading achievement. School-level indicators of students' perception of teacher discrimination, students' sense of belonging expressed through positively worded items, and students' perception of teacher support and directed instruction show a negative and statistically significant relationship to student reading achievement.

Given that Model 2 shows that school average ESCS cancels the effect of student ESCS on reading achievement and, after confirming that the pattern of significant indicators and effect size shown by Model 3 was similar for school level analysis, compensation, mediation and moderation analysis were performed using all variables aggregated at the school level (see Model 5 in **Error! Reference source not found.**).

Compensation analysis

Results indicate that a positive school climate, as manifested by high levels of disciplinary climate in language lessons, students' respect for diversity, and sense of belonging expressed thdirected instructionded items, and low levels of teacher discrimination practices, sense of belonging expressed through positively worded items, and teacher support and directed-instruction, has a compensating positive contribution

to school average reading scores, beyond the contribution of school ESCS. This finding is indicated by comparing Models 4 and 5 in **Error! Reference source not found.**.

The compensation effect is illustrated by the fact that school average ESCS explains 39% of the variation in school average reading achievement (Model 4) and the inclusion of school climate indicators accounts for an additional 47% of the variation in school achievement.

Mediation analysis

The mediating role of school climate in the relationship between school average ESCS and school average reading achievement is also investigated by comparing Models Models 4 and 5 in **Error! Reference source not found.**. Results indicate that school c limate has a partial mediating effect on school average reading achievement. This finding is observed in that the inclusion of school climate reduces the effect of school average ESCS on school average reading achievement. However, given that the effect of school average ESCS is still significant in Model 5, a partial mediation effect is observed.

Table 2. School-level models

SCHOOL LEVEL MODELS			
	Model 4	Model 5	Model 6
(Intercept)	441.37 ***	374.37 ***	383.24 ***
	(7.74)	(8.27)	(13.03)
Schl_ESCS	44.99 ***	17.19 ***	23.17 ***
	(4.33)	(2.87)	(6.67)
schl_disclima		11.42 *	16.06
		(4.99)	(13.09)
schl_bullying		-2.83	-13.64
		(10.91)	(23.67)
schl_truancy		-11.26	10.32
		(8.40)	(15.69)
schl_Tdiscrim		-37.94 ***	-36.88 **
		(5.19)	(12.51)
schl_Srespect		35.65 ***	53.06 ***
		(5.59)	(12.96)
schl_Sbelneg		36.00 ***	42.66 *
		(8.12)	(21.14)
schl_Sbelpos		-43.81 ***	-95.98 ***
		(10.70)	(23.91)
schl_Scompete		5.79	-4.97
		(7.95)	(15.70)
schl_Scooper		-7.07	-1.94
		(8.34)	(19.87)
schl_TSupDirir		-22.65 **	-42.17 *
		(7.61)	(19.37)
schl_TAdapFeec		-8.81	-20.14
		(9.26)	(27.21)
schl_TstimReac		4.03	13.46
		(8.74)	(17.33)
schl_Tinteres1		-11.44	9.89
		(8.39)	(19.24)
schl_ESCS.discl			5.25
			(7.29)
schl_ESCS.bully			-5.48
			(11.99)
schl_ESCS.truan			14.15 .
			(8.57)
schl_ESCS.Tdisc			-2.26
			(6.40)
schl_ESCS.Sresp			12.65 .
			(7.08)
schl_ESCS.Belne			3.02
			(10.49)
schl_ESCS.Belpo			-33.50 **
			(12.37)
schl_ESCS.Scomp			-4.03
			(8.91)
schl_ESCS.Scoop			1.68
			(10.43)
schl_ESCS.TsupD			-17.83 .
			(9.79)
schl_ESCS.TAdFe			-7.95
			(13.84)
schl_ESCS.StimR			7.24
			(9.58)
schl_ESCS.Tintr			9.84
			(9.69)
R2	0.39	0.86	0.88
R2.adj	0.39	0.86	0.87
Num. obs.	397	397	397
*** p < 0.001; ** p < 0.01; * p < 0.05; . p < 0.1			

Moderation model

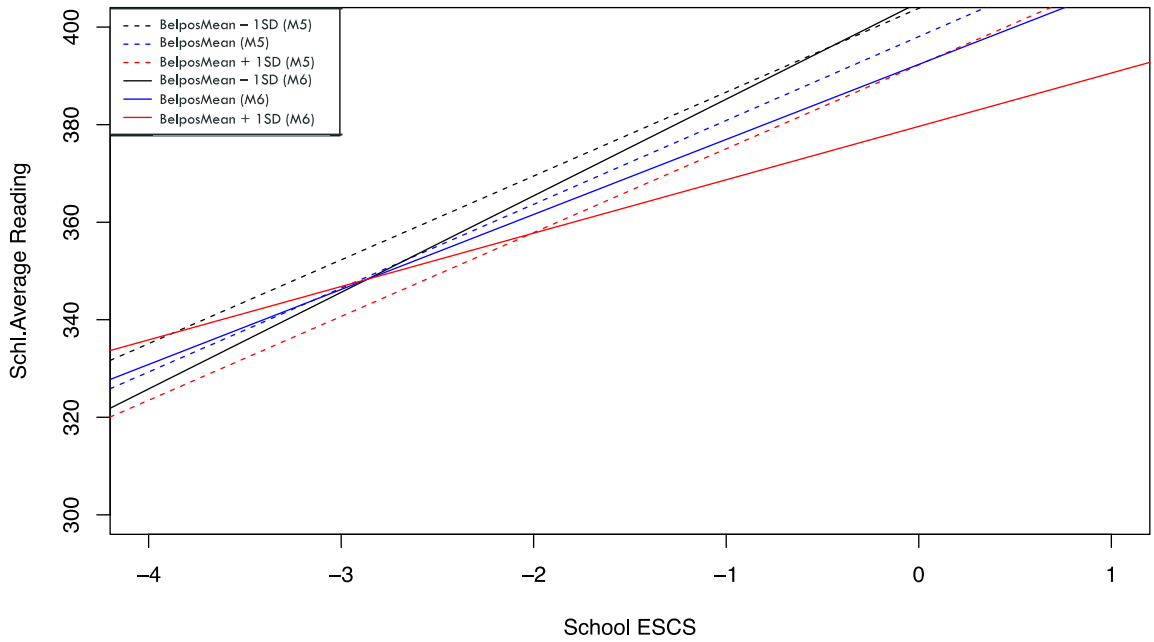
Investigating whether school climate moderates the role of school average ESCS on school average reading achievement entails introducing interaction effects between each indicator of school climate and school average ESCS (Model 6 in **Error! Reference s ource not found.**). Statistically significant interactions will be an indication of moderation effects.

Results indicate that the only significant interaction between school average ESCS and school climate indicators is verified in relation to students' sense of belonging expressed through positively worded items.

Figure 4 displays the relationship between school average ESCS and school average reading achievement for three different levels of students' sense of belonging expressed through positively worded items: at its mean (blue line), one standard deviation below

its mean value (black line) and one standard deviation above its mean value (red line). All other school climate indicators are fixed at their average value. In the figure, dotted lines represent the direct effect of students' sense of belonging expressed through positively worded items in Model 5; continuous lines represent the moderation effect (Model 6). The figure shows that in schools with more school belonging expressed through positively worded items, the relationship between school ESCS and school average reading achievement becomes negative from positive.

Figure 4. Moderation effect of students' sense of belonging expressed through positively worded items in the relationship between school average ESCS and school average reading achievement



Discussion and Conclusion

Associations between economic, cultural, and social status indicators have been extensively explored in different disciplines such as sociology, psychology, and health for a variety of outcomes, and throughout various stages of life and contexts. The relationship between economic, cultural and social status indicators and learning outcomes is also a common and well-documented subject in educational research (see, for example, (Broer et al., 2019; Conger et al., 2010; Hattie, 2009; O'Connell, 2019; Yang-Hansen, 2008).

Although understanding the relationship between economic, cultural, and social status and student achievement is crucial for monitoring equity in education, the former is a factor that cannot be modified. Therefore, understanding the role that other factors that can be modifiable by policy actions play in this relationship is key for acting upon inequity. School climate is one of those modifiable factors.

In Indonesia, students' economic, cultural, and social status plays only a minor role in explaining differences in achievement between students. However, Indonesian schools show low diversity in terms of economic, cultural, and social status, which means students tend to attend schools where their schoolmates tend to have similar economic, cultural, and social status. In this context, schools' average economic, cultural, and

social status does play a significant role in explaining differences in achievement between schools – that is, students in schools with higher average economic, cultural, and social status achieve higher scores in PISA.

In this study it has been shown that school climate compensates and partially mediates the relationship between economic, cultural, and social status and student reading achievement. Compensation is verified in that a positive school climate contribute value to school achievement on top of that contributed by school average economic, cultural, and social status.

Partial mediation is verified in that the effect of school climate indicators on school achievement is statistically significant. Considering the contribution of these school climate indicators reduces the effect of school average economic, cultural, and social status on school achievement but it does not eliminate its effect. Therefore, the importance of unpacking what positive school climate is and how it can be modified.

In Indonesia, a positive school climate, according to students' perceptions, is manifested by several indicators. In the first place, a positive school climate is manifested by high levels of positive disciplinary climate in language lessons, high students' respect for diversity and high students' sense of belonging to school expressed through positively worded items. A positive school climate is also manifested in low levels of teacher discrimination as perceived by students, low levels of students' sense of belonging to school expressed through negatively worded items and low levels of teacher support to students and directed-instruction practices. Policy actions aimed at improving school climate in these dimensions may have an impact on improving school average reading achievement in schools that enrol students coming from low economic, cultural, and social status contexts.

Two indicators of school climate require further attention in future research in Indonesia: students' sense of belonging to school and teacher support and directed instruction practices.

In PISA, students' sense of belonging to school is assessed using six items, three of which are worded positively and three of which are worded negatively (see Table 3 in Appendix). In this study it was found that in Indonesia a positive school climate is given by high levels of students' sense of belonging to school expressed through positively worded items and low levels of sense of belonging to school expressed through negatively worded items. This means that, in Indonesia, a sense of belonging expressed by not feeling like an outsider, lonely or awkward at school does not contribute to a positive school climate. It may be the case that, in Indonesia, this type of items are indicators of a construct different than "sense of belonging to school". The importance of further investigating the "students' sense of belonging to school" items also relate to the finding that only a sense of belonging expressed through positively worded items acts as a moderator factor in the relationship between economic, cultural and social status indicators and school average reading achievement.

The second indicator of school climate that requires further attention is teacher support and directed instruction. In Indonesia this indicator has a negative association with student and school average ESCS and a negative relationship to the school climate latent variable. This expresses that teacher support and directed-instruction practices in schools serving students from low economic, cultural, and social status are not conducive to a positive school climate promoting high achievement. In Indonesia,

receiving more support from teachers may be related to the fact that students are not performing well and, therefore, need more attention from teachers.

References

Berkowitz, R., Moore, H., Astor, R. A., & Benbenishty, R. (2017). A Research Synthesis of the Associations Between Socioeconomic Background, Inequality, School Climate, and Academic Achievement. *Review of Educational Research*, *87*(2), 425–469. https://doi.org/10.3102/0034654316669821

Broer, M., Bai, Y., & Fonseca, F. (2019). *Socioeconomic Inequality and Educational Outcomes: Evidence from Twenty Years of TIMSS* (Vol. 5). Springer International Publishing. https://doi.org/10.1007/978-3-030-11991-1

Brookover, W. B., Schweitzer, J. H., Schneider, J. M., Beady, C. H., Flood, P. K., & Wisenbaker, J. M. (1978). Elementary School Social Climate and School Achievement. *American Educational Research Journal*, *15*(2), 301–318. https://doi.org/10.2307/1162468

Cohen, J., Mccabe, E. M., Michelli, N. M., & Pickeral, T. (2009). School Climate: Research, Policy, Practice, and Teacher Education. *Teachers College Record*, *111*(1), 180–213. https://doi.org/10.1177/016146810911100108

Conger, R. D., Conger, K. J., & Martin, M. J. (2010). Socioeconomic Status, Family Processes,and Individual Development. *Journal of Marriage and Family*, *72*(3), 685–704. https://doi.org/10.1111/j.1741-3737.2010.00725.x

Freiberg, H. J. (Ed.). (1999). *School Climate: Measuring, Improving and Sustaining Healthy Learning Environments*. Routledge. https://doi.org/10.4324/9780203983980

Gronna, S. S., & Chin-Chance, S. A. (1999). *Effects of School Safety and School Characteristics on Grade 8 Achievement: A Multilevel Analysis*. https://eric.ed.gov/?id=ED430292

Halpin, A. W., & Croft, D. B. (1963). *The Organizational Climate of Schools*. Midwest Administration Center, University of Chicago.

Hattie, John. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. Routledge.

Hoy, W. K., Tarter, C. J., & Hoy, A. W. (2006). Academic Optimism of Schools: A Force for Student Achievement. *American Educational Research Journal*, *43*(3), 425–446. https://doi.org/10.3102/00028312043003425

Juvonen, J., Wang, Y., & Espinoza, G. (2011). Bullying Experiences and Compromised Academic Performance Across Middle School Grades. *The Journal of Early Adolescence*, *31*(1), 152–173. https://doi.org/10.1177/0272431610379415

Lee, J.-S. (2012). The effects of the teacher–student relationship and academic press on student engagement and academic performance. *International Journal of Educational Research*, *53*, 330–340. https://doi.org/10.1016/j.ijer.2012.04.006

O'Connell, M. (2019). Is the impact of SES on educational performance overestimated? Evidence from the PISA survey. *Intelligence*, *75*, 41–47. https://doi.org/10.1016/j.intell.2019.04.005

OECD. (n.d.). *PISA 2018 Technical Report—PISA*. Retrieved November 7, 2022, from https://www.oecd.org/pisa/data/pisa2018technicalreport/

OECD. (2009). *PISA Data Analysis Manual: SAS Second Edition*. https://www.oecd.org/pisa/pisaproducts/pisadataanalysismanualspssandsassecondedition.htm

OECD. (2019a). *PISA 2018 Results (Volume I): What Students Know and Can Do*. PISA, OECD Publishing. https://www.oecd.org/pisa/publications/

OECD. (2019b). *PISA 2018 Results (Volume II): Where All Students Can Succeed*. OECD. https://doi.org/10.1787/b5fd1b8f-en

OECD. (2019c). *PISA 2018 Results (Volume III): What School Life Means for Students' Lives*. OECD. https://doi.org/10.1787/acd78851-en

Pusat Asesmen Pendidikan. (2022). *Asesmen Nasional*. https://pusmendik.kemdikbud.go.id/an/page/survei_lingkungan_belajar

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical Linear Models: Applications and Data Analysis Methods*. SAGE.

Reynolds, D., Sammons, P., De Fraine, B., Van Damme, J., Townsend, T., Teddlie, C., & Stringfield, S. (2014). Educational effectiveness research (EER): A state-of-the-art review. *School Effectiveness and School Improvement*, *25*(2), 197–230. https://doi.org/10.1080/09243453.2014.885450

Wang, M.-T., & Degol, J. L. (2016). School Climate: A Review of the Construct, Measurement, and Impact on Student Outcomes. *Educational Psychology Review*, *28*(2), 315–352. https://doi.org/10.1007/s10648-015-9319-1

Wang, M.-T., & Holcombe, R. (2010). Adolescents' Perceptions of School Environment, Engagement, and Academic Achievement in Middle School. *American Educational Research Journal*, *47*(3), 633–662. https://doi.org/10.3102/0002831209361209

Yang-Hansen, K. (2008). Ten-year trend in SES effects on reading achievement at school and individual levels: A cross-country comparison. *Educational Research and Evaluation*, *14*(6), 521–537. https://doi.org/10.1080/13803610802576759

Zullig, K. J., Koopman, T. M., Patton, J. M., & Ubbes, V. A. (2010). School Climate: Historical Review, Instrument Development, and School Assessment. *Journal of Psychoeducational Assessment*, *28*(2), 139–152. https://doi.org/10.1177/0734282909344205

Appendix

Table 3. Definition and operationalisation of school climate

Sphere	Indicator	Definition	Items	PISA itemcode	Response categories
Student behaviour	Disciplinary climate (schl_disclima)	Positive values on these items mean that the student enjoys a better disciplinary climate in language-of-instruction lessons	Students don't listen to what the teacher says.	ST097Q01TA	4=never or hardly ever 3=some lessons 2=most lessons 1=every lesson
			There is noise and disorder.	ST097Q02TA	
			The teacher waits long for students to quiet down.	ST097Q03TA	
			Students cannot work well.	ST097Q04TA	
			Students don't start working for a long time after the lesson begins.	ST097Q05TA	
	Bullying (schl_bullying)	Positive values on these items indicate that the student is more exposed to bullying at school	Other students left me out of things on purpose	ST038Q03NA	1=Never or almost never 2=A few times a year 3=A few times a month 4=Once a week or more
			Other students made fun of me	ST038Q04NA	
			I was threatened by other students	ST038Q05NA	
	Student truancy and lateness (schl_truancy)	Positive values on these items indicate more truancy	I <skipped> a whole school day	ST062Q01TA	1=Never 2=One or two times 3=Three or four times 4=Five or more times
			I <skipped> some classes	ST062Q02TA	

Sphere	Indicator	Definition	Items	PISA itemcode	Response categories
School community	Respect for diversity – teacher discrimination (schl_Tdiscrim)	Positive values on these items indicate teachers discriminate students by cultural groups	Teachers in your school: They have misconceptions about the history of some cultural groups	ST223Q02HA	1= To none or almost none of them 2= To some of them 3 =To most of them 4= To all or almost all of them
			Teachers in your school: They say negative things about people of some cultural groups	ST223Q04HA	
			Teachers in your school: They blame people of some cultural groups for problems faced by <country of test>	ST223Q05HA	
			Teachers in your school: They have lower academic expectations for students of some cultural groups	ST223Q08HA	
	Respect for diversity – students (schl_Srespect)	Positive values on these items indicate more respect for diversity	How well does the following describe you: I respect people from other cultures as equal human beings	ST217Q01HA	1= Very much like me 2= Mostly like me 3 =Somewhat like me 4 =Not much like me 5= Not at all like me
			How well does the following describe you: I treat all people with respect regardless of their cultural background	ST217Q02HA	
			How well does the following describe you: I give space to people from other cultures to express themselves	ST217Q03HA	
			How well does the following describe you: I respect the values of people from different cultures	ST217Q04HA	
	Sense of belonging at school	Positive values on these items mean that the student has a stronger	Thinking about your school: I make friends easily at school (schl_Sbelpos)	ST034Q02TA	1 =Strongly agree 2= Agree 3= Disagree

Sphere	Indicator	Definition	Items	PISA itemcode	Response categories
		sense of belonging at school	Thinking about your school: I feel like I belong at school (schl_Sbelpos)	ST034Q03TA	4 =Strongly disagree
			Thinking about your school: Other students seem to like me (schl_Sbelpos)	ST034Q05TA	
			Thinking about your school: I feel like an outsider (or left out of things) at school (schl_Sbelneg)	ST034Q01TA	
			Thinking about your school: I feel awkward and out of place in my school (schl_Sbelneg)	ST034Q04TA	
			Thinking about your school: I feel lonely at school (schl_Sbelneg)	ST034Q06TA	
	Perceived competition (schl_Scompete)	Positive values on these items mean that students perceive that other students at the school compete with each other	Think about your school, how true: Students seem to value competition	ST205Q01HA	1= Not at all true 2= Slightly true 3 =Very true 4 =Extremely true
			Think about your school, how true: It seems that students are competing with each other	ST205Q02HA	
			Think about your school, how true: Students seem to share the feeling that competing with each other is important	ST205Q03HA	
	Perceived cooperation (schl_Scooper)	Positive values on these items mean that students perceive that other students at the school co-operate with each other	Think about your school, how true: Students seem to value cooperation	ST206Q01HA	
			Think about your school, how true: It seems that students are cooperating with each other	ST206Q02HA	

Sphere	Indicator	Definition	Items	PISA itemcode	Response categories
			Think about your school, how true: Students seem to share the feeling that cooperating with each other is important	ST206Q03HA	

Sphere	Indicator	Definition	Items	PISA itemcode	Response categories
Teaching and learning	Teacher support (schl_TSupDirins)	Positive values on these items mean that students perceive their teacher support learning	How often during <test language lessons>: The teacher shows an interest in every student's learning	ST100Q01TA	1 =Every lesson 2 =Most lessons 3 =Some lessons 4 =Never or hardly ever
			How often during <test language lessons>: The teacher gives extra help when students need it	ST100Q02TA	
			How often during <test language lessons>: The teacher helps students with their learning	ST100Q03TA	
			How often during <test language lessons>: The teacher continues teaching until the students understands	ST100Q04TA	
	Teaching practice – directed instruction (schl_TSupDirins)	Positive values on these items mean that students perceive their teachers to use teacher-directed practices frequently	How often during <test language lessons>: The teacher sets clear goals for our learning	ST102Q01TA	
			How often during <test language lessons>: The teacher asks questions to check whether we have understood what was taught	ST102Q02TA	
			How often during <test language lessons>: [...] the teacher presents a short summary of the previous lesson	ST102Q03TA	

Sphere	Indicator	Definition	Items	PISA itemcode	Response categories
			How often during <test language lessons>: The teacher tells us what we have to learn	ST102Q04TA	
	Teaching practice – stimulation of reading (schl_TStimRead)	Positive values on these items mean that the students perceive their teacher to provide stimulation for reading	In your <test language lessons>, how often: The teacher encourages students to express their opinion about a text	ST152Q05IA	1= Never or hardly ever 2= In some lessons 3 =In most lessons 4 =In all lessons
			In your <test language lessons>, how often: The teacher helps students relate the stories they read to their lives	ST152Q06IA	
			In your <test language lessons>, how often: The teacher shows students how the information in texts builds on [...]	ST152Q07IA	
			In your <test language lessons>, how often: The teacher poses questions that motivate students to participate actively	ST152Q08IA	
	Teaching practice – adaptive instruction (schl_AdapFeed)	Positive values on these items mean that students perceived their language-of-instruction teachers to be more adaptive	How often in <test language lessons>: The teacher adapts the lesson to my class needs and knowledge	ST212Q01HA	1 =Never or almost never 2= Some lessons 3 =Many lessons 4= Every lesson or almost every lesson
			How often in <test language lessons>: The teacher provides individual help when a student has difficulties [...]	ST212Q02HA	
			How often in <test language lessons>: The teacher changes the structure of the lesson on a topic that most [...]	ST212Q03HA	
	Teacher support – feedback (schl_AdapFeed)	Positive values on these items mean that students perceive their teachers to provide frequent feedback	How often during <test language lessons>: The teacher gives me feedback on my strengths in this subject	ST104Q02NA	
			How often during <test language lessons>: The teacher tells me in which areas I can still improve	ST104Q03NA	

Sphere	Indicator	Definition	Items	PISA itemcode	Response categories
			How often during <test language lessons>: The teacher tells me how I can improve my performance	ST104Q04NA	
	Teacher behaviour – interest (schl_Tinterest)	Positive values on these items mean that students perceived their language-of-instruction teachers to be enthusiastic about teaching	Thinking of past two <test language lessons>: It was clear to me that the teacher liked teaching us	ST213Q01HA	1= Strongly disagree 2 =Disagree 3= Agree 4 =Strongly agree
			Thinking of past two <test language lessons>: The enthusiasm of the teacher inspired me	ST213Q02HA	
			Thinking of past two <test language lessons>: It was clear that the teacher likes to deal with the topic of the lesson	ST213Q03HA	

The psychometric properties of Dyadic Coping Inventory (DCI)

Amorita Christella A
Fitri Andriani
Airlangga University

https://doi.org/10.37517/978-1-74286-697-0-03

Abstract

A survey conducted by The Indonesian Central Bureau of Statistics (Badan Pusat Statistik/BPS Indonesia) last year has found that the divorce rate in Indonesia had increased up to 53,50% compared to the data in 2020, and more than half of the divorce cases happened because of "unsolved conflicts between couples". Out of various studies on marital or romantic relationships done internationally, it suggests that dyadic coping could be the protective factor in the relationship between stress and the couples' well-being. According to a meta-analysis study held in 2015 along with systematic literature reviews published in 2019 and 2022 regarding the construct, the Systemic Transactional Model (STM) of dyadic coping proposed by Guy Bodenmann, as well as the instrument used to measure it, the Dyadic Coping Inventory/DCI (also developed from the same model), are considered the most frequently used model and instrument in assessing dyadic coping together with other related construct (e.g., relationship satisfaction). The Dyadic Coping Inventory has been translated into 25 languages and has been adapted in 14 countries. Unfortunately, this instrument had only been translated from English to Indonesian and had not been tested, specifically for its validity and reliability on the Indonesian population. The purpose of this study is to examine the psychometric property of the Indonesian version of Dyadic Coping Inventory. The study is conducted by following the International Test Commission (ITC) guidelines for translating and adapting tests (2017). A total of 104 participants were involved in this research (during pilot study). Results showed that the instrument has an overall high internal consistency (α = .93). Evidence for validity based on content and response processes were also acquired through experts' judgement and cognitive interviews. Further studies should include a larger sample size to test the Indonesian version of DCI and perform a factorial analysis (validity based on internal structure).

Introduction

The unsolved conflicts between couples in Indonesia

A survey conducted by The Indonesian Central Bureau of Statistics (Badan Pusat Statistik/BPS Indonesia) last year has found that the divorce rate in Indonesia had increased up to 53,50% compared to the data in 2020 (±447.743 cases), and more than half of the divorce cases (±279.205 cases) happened because of "unsolved conflicts

between couples" (Annur, 2022; Jayani, 2021). According to ICD 10, how parents deal with stress can be a risk factor for their children's development. Bronfenbrenner's theory of ecological systems also explains how various social environments (e.g., the micro-system) could affects a child's development, which he called 'the transactional influence' (Lerner & Damon, 2006; Rosa & Tudge, 2013; Tudge & Rosa, 2020). The act of modelling is one of the earliest methods for children to learn about their surroundings (Crain, 2014). Parents are often said to be one of the sources of that modelling, As the child grows up, the result of the modelling might develop into something more permanent like the child's habits, personality traits, or social competence (Murray & Farrington, 2005; Scaglioni et al., 2018). There's also the classic nature-versus-nurture debate. Physiologically speaking, humans have that innate fight-or-flight stress response, but as the epigenetics perspective emerges, which stress response we dominantly use in daily life or life-threatening situation could be influenced by our surroundings (Boyce, 2019). All of the previous findings take us back to the parents.

Dyadic coping as the protective factor

There are a lot of models of dyadic coping, but the one being used in this study is the systemic transactional model (STM) developed by Guy Bodenmann (1995), as well as the instrument used to measure it, the Dyadic Coping Inventory (DCI) (Bodenmann, 2008a). Based on a meta-analysis study held in 2015 along with systematic literature reviews published in 2019 and 2022 regarding the construct, the model and instrument developed by Bodenmann are considered the most frequently used model and instrument in assessing dyadic coping together with other related construct (e.g., relationship satisfaction) (Bodenmann et al., 2019; Falconier et al., 2015; Nepomuceno et al., 2022).

According to the Systemic Transactional Model/STM (Bodenmann, 1995; Kayser & Bodenmann, 2005), the process of dyadic coping (DC) starts with 'stress communication', It happens when one partner sends stress signals verbally (e.g., complaining about work) or non-verbally (through facial expressions or behavior). The other partner then assesses what the partner is going through and decides whether to give positive or negative dyadic coping.

The positive and negative dyadic coping can each be further specified (Bodenmann et al., 2019; Kayser & Bodenmann, 2005). Positive dyadic coping includes supportive dyadic coping (SDC), which occurs when one shows empathy and help his/her partner by giving practical solution; delegated dyadic coping (DDC) occurs when one takes over his/her partner regular activities in order to help; partners could also engage in common dyadic coping (CDC), this form of dyadic coping occurs when both partners try to deal with stress together.

Negative dyadic coping (NDC) is characterized by behaviors such as forcing oneself to give support to the partner, withdrawing from the partner in times of need, and/or insulting the partner verbally/non-verbally. While literature about the negative dyadic coping is scarce, some of studies found that this form dyadic coping tends to be shown through verbal aggression, anger, having insomnia, and social dysfunction in men and women (Bodenmann et al., 2010, 2011).

As of the present, the model and instrument measuring dyadic coping proposed by Bodenmann is the only one that measures couples' coping with daily stressors. While the other models and instruments available, were often developed from studies involving couples' coping with chronic illnesses, this condition would create different type of stress

that might need different kind of coping. In Indonesia itself, the dyadic coping had often been examined in studies involving couples coping with chronic illnesses or mental disorders (Nurnaningsih et al., 2018; Purba & Risnawaty, 2021). Cutrona et al., (2018) provided a review on the difference between dyadic coping and other similar construct (e.g., social support).

Psychometric properties of the Dyadic Coping Inventory (DCI)

The Dyadic Coping Inventory is a self-administered questionnaire with 37 items. It is a 5-point Likert scale. Participants can rate their response to each question by choosing 1 (very rarely/never) to 5 (very often/always) options. It takes about 10-15 minutes to complete the scale. For scoring, the instruction given by the manual (Bodenmann, 2008b) is to sum up items 1-35 after reverse coding the items from negative dyadic coping dimension. The last two items are not included in the scoring as both are only evaluation items.

The original DCI has high internal consistency for its subscales with α ranging from .71 till .92 (Bodenmann, 2008a). Most of the adapted versions of DCI also have overall high internal consistency with α above .80 (Nepomuceno et al., 2022).

As for validity, it was first tested using Principal Component Analysis (PCA) on data collected from German, Italian, and French samples. It had over 50% of the variance explained, respectively (Bodenmann, 2008b). Most of the studies adapting DCI in different countries used Confirmatory Factor Analysis or CFA and it resulted in a 5 + 5 + 2 structure (Nepomuceno et al., 2022). From these results, it can be seen that this instrument can measures coping from one's own perspective and the partner's perspective and so it showed the 5 + 5 factors. Almost all of the adaptation studies also divided the positive dyadic coping into problem-focused and emotion-focused coping hence the last 2 factors. In some of the meta-analysis studies, it turned out that dyadic coping correlated strongly and positively to marital quality and relationship satisfaction (Bodenmann, 2008a; Falconier et al., 2015). These could be the evidence for convergent validity of dyadic coping.

The purpose of this study

Based on the benefits of having couples do dyadic coping, its effect on the relationships' well-being, and possibly children's development, we try to examine the Indonesian version of the Dyadic Coping Inventory, so that a valid and reliable instrument can be used in further studies.

Method

We conducted the study by following the International Test Commisions (2017) guidelines for translating and adapting tests. We received both permission and the Indonesian translation of DCI from the original author. The Indonesian translation DCI was translated by another researcher in a non-psychometric study a few years ago (Dermawan et al., 2015). We continued working on the translated version by doing the next steps, which were backward-translation until pilot study.

Participant

Several participants were recruited to take part in cognitive interviews (validity based on response processes) and ±100 participants were involved in the pilot study. They had to fulfill two criteria to participate in the study; (1) an Indonesian citizen, and (2) have been married for at least 2 years. All of the participants were recruited through social media, and they gave signed informed consent before participating.

We also collected sociodemographic data during the pilot study. A total of 104 participants with 61.5% men (n = 64) and 38.5% women (n = 40) were involved in the study. 13.5% (n = 14) were aged 20-30, 25% (n = 26) were aged 31-40, 49.5% (n = 51) were aged 41-50, and 12% (n = 13) were aged 51-60. 76% (n = 65) were of Javanese ethnicity. As of educational background, 8.7% (n = 9) graduated high school, 9.6% (n = 10) got a diploma, 49% (n = 51) got a bachelor degree, and 32.7% (n = 34) got their master's degree. For marriage duration, about 69.2% (n = 52) have been married for more than 10 years, 28.8% (n = 30) have been married for 2-10 years, while 1.9% (n = 2) have been married for more than 30 years. Lastly, 74.1% (n = 77) of the participants have more than 2 kids.

Measures

Sociodemographic variables

We asked for the participants sociodemographic data before they started filling the questionnaire. Some of the data collected were of their gender, age, ethnicity, educational background, marriage duration, and number of kids.

The Dyadic Coping Inventory (DCI)

The DCI developed by (Bodenmann, 2008a) consisted of 37 items. It's a self-administered questionnaire with 5-point Likert scale. The dimensions measured in the instrument are based on items reflecting stress communication (SC = 8 items), forms of dyadic coping (SDC = 10 items; DDC = 4 items, CDC = 5 items; and NDC = 8 items) and evaluation of dyadic coping (EDC = 2 items). The scoring for DCI is done by summing up 35 items after reverse coding 8 items from NDC and excluding 2 items from EDC.

Data Analysis

The backward-translation was done by one translator with a bachelor degree on English Literature. We compared the original version and backward-translation version using analysis suggested by Sperber (2004) to see linguistic aspects such as similarity and comparability from both versions.

Evidence for validity based on content were acquired using Polit and Beck (2007) theory on I-CVI and S-CVI. We also looked for evidence for validity based on response processes by conducting cognitive interview, and applied thematic analysis (Willis, 2014) to the results.

Finally, we determined the reliability of the DCI (Indonesian version) by looking at the instrument's internal consistency (Cronbach's Alpha) using Jamovi version 2.3.

Findings

The backward-translation process

Three raters were involved in the process of comparing the original version and the backward-translation version. All of them have education backgrounds in psychology and is fluent in both English and Indonesian languages. Through this step, we acquired results on the similarity and comparability aspects of the instrument. All items had mean scores below 2,50 on both aspects, which considered acceptable (meaning that all of the

items were of similar meaning despite having different sentence structures than the original version).

Evidence for validity based on content

We acquired evidence for validity based on content through expert judgement. Several raters were involved in this stage, they were selected based on their expertise on fields such as relationships, stress, and coping. Some of them are psychologists, postgraduate psychology students, and a lecturer. After calculating the I-CVI value & S-CVI value (Table 1) based on Polit's theory, and after reviewing the experts' feedback, we came up with the draft of the instrument.

Table 1. S-CVI value of the DCI – Indonesian version

Parameter			Conclusion
Relevancy	Importance	Clarity	
0,99	0,98	0,97	The instrument is valid and can be used as an assessment tool, with a minor revision according to I-CVI results

Evidence for validity based on response processes

The cognitive interviews were done before we conducted the pilot study. Several participants were involved and most of them understood the instructions and what each item of the instrument was meant to measure. We inferred this conclusion based on the result of thematic analysis (Willis, 2014). We found two themes related to dyadic coping, one was communication pattern (verbal or non-verbal) and the other one was coping strategies (how they cope can be linked back to the definitions of each form of dyadic coping).

Based on their comments about the response descriptions, we changed it from "very rarely to very often" into "never to always". Some of the participants said the former was rather ambiguous in the Indonesian language and that they were more familiar with the "never to always" type of Likert scale.

The participants also gave feedback on some of the items of the instrument. We took that into consideration and finally came up with the version to be used in pilot study.

Reliability (Cronbach's Alpha)

The coefficient of reliability in this study can be seen on Table 2. The result is consistent with previous findings from studies adapting the instrument. The DCI has an overall high internal consistency with value above .90 (α = .93). As for each of the dimensions, it has value ranging from .64 to .95.

Table 2. Coefficient of reliability for DCI

Dimension	Cronbach's Alpha
Stress communication (SC)	0.775
Supportive dyadic coping (SDC)	0.894
Delegated dyadic coping (DDC)	0.644
Common dyadic coping (CDC)	0.776
Negative dyadic coping (NDC)	0.753
Evaluation of dyadic coping (EDC)	0.951

Discussion

The examination on the equivalence of the instrument from a linguistic point of view was done using Sperber (2004) theory. The mean scores of each item are below 2,50, it showed that all items between the original version and the backward-translation version are equivalent in terms of similarity and comparability.

This study also provided evidence for validity based on content and response processes of the instrument. The result of the expert judgement is quite comprehensive, as it evaluated the items by using quantitative measures of the method suggested by Polit and Beck (2007), while also taking into account the qualitative aspects (written feedbacks on both the items and the overall scale). Findings acquired from cognitive interviews added yet another evidence for the instrument's validity. From feedbacks given by the participants and from thematic analysis done according to Willis (2014), we managed to improve the quality of the instrument before conducting tryouts.

The instrument showed a high internal consistency, and so are the dimensions as per the result of the pilot study (Tabel 2). One of the dimensions has coefficient bellow .70, but we did not remove it. Nepomuceno et al. (2022) stated that it is still acceptable though values between .60 to .70 are considered the lower acceptability limit.

Conclusion

Limitations

Since the findings mentioned in this article derived from a pilot study, we suggest further studies to include larger samples (±300) to conduct Confirmatory Factor Analysis/CFA as prove for validity based on internal structure of the instrument. Also, two of the items have inter-items coefficients below 0.30, so we suggest that perhaps some revisions can be made on these items before conducting the large-scale reliability and validity testing (field test).

References

Annur, C. M. (2022). *Kasus Perceraian Meningkat 53%, Mayoritas karena Pertengkaran. Retrieved April 7, 2022*. https://databoks.katadata.co.id/ website: https://databoks.katadata.co.id/datapublish/2022/02/28/kasus-perceraian-meningkat-53-mayoritas-karena-pertengkaran#:~:text=Angka Perceraian di Indonesia (2017-2021)&text=Menurut laporan Statistik Indonesia%2C jumlah,banyak menggugat cerai ketimbang suami.

Bodenmann, G. (1995). A systemic-transactional view of stress and coping in couples. *Swiss Journal of Psychology*, *54*, 34–49.

Bodenmann, G. (2008a). *Dyadic Coping Inventory: Test Manual*. Bern, Switzerland: Huber.

Bodenmann, G. (2008b). *Dyadisches Coping Inventar: Test manual [Dyadic Coping Inventory: Test manual]*. Bern, Switzerland: Huber.

Bodenmann, G., Atkins, D. C., Schar, M., & Poffet, V. (2010). The Association Between Daily Stress and Sexual Activity. *Journal of Family Psychology*, *24*(3), 271–279. https://doi.org/doi: 10.1037/a0019365

Bodenmann, G., Meuwly, N., & Kayser, K. (2011). Two Conceptualizations of Dyadic Coping and Their Potential for Predicting Relationship Quality and Individual Well-Being. A Comparison. *European Psychologist*, *16*(4), 255–266. https://doi.org/DOI: 10.1027/1016-9040/a000068

Bodenmann, G., Randall, A. K., & Falconier, M. K. (Eds.). (2019). *Dyadic Coping: A Collection of Recent Studies*. https://doi.org/DOI: https://doi.org/10.3389/978-2-88963-031-8

Boyce, W. T. (2019). *The Orchid and The Dandelion: Why Sensitive People Struggle and How All Can Thrive*. Colorado: Bluebird.

Crain, W. (2014). *Theories of Development Concepts and Applications* (6th ed.). Harlow: Pearson.

Cutrona, C., Bodenmann, G., Randall, A. K., Clavel, F. K., & Johnson, M. (2018). Stress, Dyadic Coping, and Social Support: Moving toward Integration. In *The Cambridge Handbook of Personal Relationships* (2nd ed., pp. 341–352). https://doi.org/doi:10.1017/9781316417867.027

Dermawan, S., Goei, Y. A., & Kirana, K. C. (2015). Pengaruh Dyadic Coping terhadap Kepuasan Pernikahan pada Pasangan Menikah di Tangerang. *Jurnal Psikologi Ulayat*, *2*(2), 420–433. https://doi.org/DOI: 10.24854/jpu22015-37

Falconier, M. K., Jackson, J. B., Hilpert, P., & Bodenmann, G. (2015). Dyadic coping and relationship satisfaction: A meta-analysis. *Clinical Psychology Review*, *42*, 28–46. https://doi.org/http://dx.doi.org/10.1016/j.cpr.2015.07.002

International Test Commisions. (2017). The ITC Guidelines for Translating and Adapting Tests [Second Edition]. https://doi.org/10.1111/j.1464-0597.1975.tb00322.x

Jayani, D. H. (2021). Perselisihan Jadi Penyebab Utama Perceraian Sepanjang 2020. Retrieved April 7, 2022, from https://databoks.katadata.co.id/ website: https://databoks.katadata.co.id/datapublish/2021/03/19/perselisihan-jadi-penyebab-utama-perceraian-sepanjang-2020

Kayser, K., & Bodenmann, G. (2005). *Couples Coping with Stress: Emerging Perspective on Dyadic Coping* (T. A. Revenson, Ed.). Washington, DC: American Psychological Association.

Lerner, R. M., & Damon, W. (Eds.). (2006). *Handbook of Child Psychology. Volume One: Theoretical Models of Human Development* (6st ed.). New Jersey: John Wiley & Sons, Inc.

Murray, J., & Farrington, D. P. (2005). Parental imprisonment: effects on boys' antisocial behaviour and delinquency through the life-course. *Journal of Child Psychology and Psychiatry*, *46*(12), 1269–1278. https://doi.org/doi: 10.1111/j.1469-7610.2005.01433.x

Nepomuceno, W. R., Goncalves, A., da, S., & Hernandez, J. A. E. (2022). Psychometric Properties of the Dyadic Coping Inventory: Systematic Review. *Interpersona*, *16*(1), 96–119. https://doi.org/https://doi.org/10.5964/ijpr.7597

Nurnaningsih, N., Safitri, J., & Mayangsari, M. D. (2018). PERANAN DYADIC COPING TERHADAP KEPUASAN PERNIKAHAN PADA PASANGAN DI KOMUNITAS ADS KALSEL. *Jurnal Kognisia*, *1*(1), 83–88. https://doi.org/DOI: https://doi.org/10.20527/jk.v1i1.1460

Polit, D. F., & Beck, C. T. (2007). The Content Validity Index: Are You Sure You Know What's Being Reported? Critique and Recommendations. *Research in Nursing & Health*, *29*, 489–497. https://doi.org/DOI: 10.1002/nur.20147

Purba, T. F., & Risnawaty, W. (2021). Dyadic Coping Among Married Couples with Breast Cancer Survivor Wife. *Advances in Social Science, Education and Humanities Research*, *570*, 1006–1012.

Rosa, E. M., & Tudge, J. (2013). Urie Bronfenbrenner's Theory of Human Development: Its Evolution From Ecology to Bioecology. *Journal of Family Theory & Review*, *5*(4), 243–258. https://doi.org/DOI:10.1111/jftr.12022

Scaglioni, S., De Cosmi, V., Ciappolinno, V., Parazzini, F., Brambilla, P., & Agostoni, C. (2018). Factors Influencing Children's Eating Behaviours. *Nutrients*, *10*(6), 706. https://doi.org/doi:10.3390/nu10060706

Sperber, A. D. (2004). Translation and Validation of Study Instruments for Cross-Cultural Research. *Gastroentology*, *126*, S124-8. https://doi.org/10.1053/j.gastro.2003.10.016

Tudge, J., & Rosa, E. M. (2020). Bronfenbrenner's Ecological Theory. In *The Encyclopedia of Child and Adolescent Development*. https://doi.org/DOI 10.1002/9781119171492.wecad251

Willis, G. (2014). *Cognitive Interviewing Methodology* (K. Miller, S. Willson, V. Chepp, & J.-L. Padilla, Eds.). New Jersey: John Wiley & Sons, Inc.

Item development & psychometric testing of work stress scale

Andhita Ramadhania Pristiawati
Laura F. N. Sudarnoto
Angela Oktavia Suryani
Atma Jaya Catholic University of Indonesia, Indonesia

https://doi.org/10.37517/978-1-74286-697-0-04

Andhita Ramadhania Pristiawati is a science psychology master's student at Atma Jaya Catholic University of Indonesia. Currently working as a user experience researcher at a retail company in Indonesia focusing on e-commerce user research.

Abstract

In 2020, the COVID-19 pandemic demanded employees work from home (Ishak & Mangundjaya, 2020). The work-from-home regulation provides several advantages, including high work flexibility. However, this regulation also creates excessive workloads and requires employees to work overtime due to various challenges from superiors with short deadlines. This situation encouraged researchers to develop an accurate psychological scale to identify the employees/ working stress. The scale comprised physical and mental individual responses to managing their uncomfortable work environment. It consists of 48 items measuring four dimensions: role overload, role conflict, role ambiguity, and role responsibility. The participants responded to the items by sending their responses on a 6-point scale, whether s/he felt the sentences described their psychological state from absolutely appropriate (score 1) to absolutely inappropriate (score 6). We recruited 91 employees aged 21 to 36 years old. In the content validity test, three experts judged whether the items were relevant, simple, clear, and unambiguous. The S-CVI was in the range of 0.89 to 0.92, with an S-CVI average of 0.90. Corrected item-total correlation revealed that all items were valid with a coefficient range from 0.40 to 0.79. Cronbach's Alpha was 0.97. We concluded that the instrument was valid and reliable. We also simulated the norm to interpret the participant's obtained scores. We created a standardized interpretation of the scores by applying percentile 25 and 75 to have three categories: low, medium, and high.

Introduction

Since the first announcement in March 2020, the COVID-19 pandemic has continued until 2022. The pandemic pushes the need to adapt and implement social distancing to minimize the Coronavirus spread. A strategy to adjust to the pandemic is to do school and work from home. Working from home is known as WFH.

Implementing the WFH policy provides several advantages for both employees and companies. One of the advantages for employees is high flexibility and saving time for travel. The benefit for the company is that it can save electricity and water budgets.

However, implementing the WFH policy also has drawbacks. The flexibility of time during WFH sometimes impacts the workload. It becomes more excessive, and the employees should work overtime because the workload is higher. In addition, the communication between the supervisor, staff, and coworkers resulted in misunderstandings that affected work. Then, because of poor communication, employees became confused and uncertain about their work because they did not have clear directions from colleagues or superiors. The heavy workload and unclear job description made employees do the job overtime, especially the IT employees. According to Azizmunaji (2018), the employee in the IT department is the most stressful worker because employees who cannot meet the demands of the work assigned to them can cause losses to the company (Abid & Safiih, 2021). Goleman (1997) suggests that emotions play a significant role in action and making "rational" decisions. Work stress in the organization has become a vital symptom to observe since the emergence of demands on employees to be efficient at work (Aisah, 2020).

A survey by PWC in 2020 included 300 Chief Financial Officers in the US showed a decrease in work productivity. This was the third most significant concern for companies after the financial impact and global recession (Ishak & Mangundjaya, 2020). Disorientation and anxiety from being burdened with excessive workload can cause employees to experience stress. We interviewed ten retail company employees in Jakarta in the IT department. We found 7 out of 10 employees reported that they had experienced work stress while undergoing WFH. This phenomenon encourages researchers to develop measuring tools that are valid and reliable and can measure employee work stress while working from home.

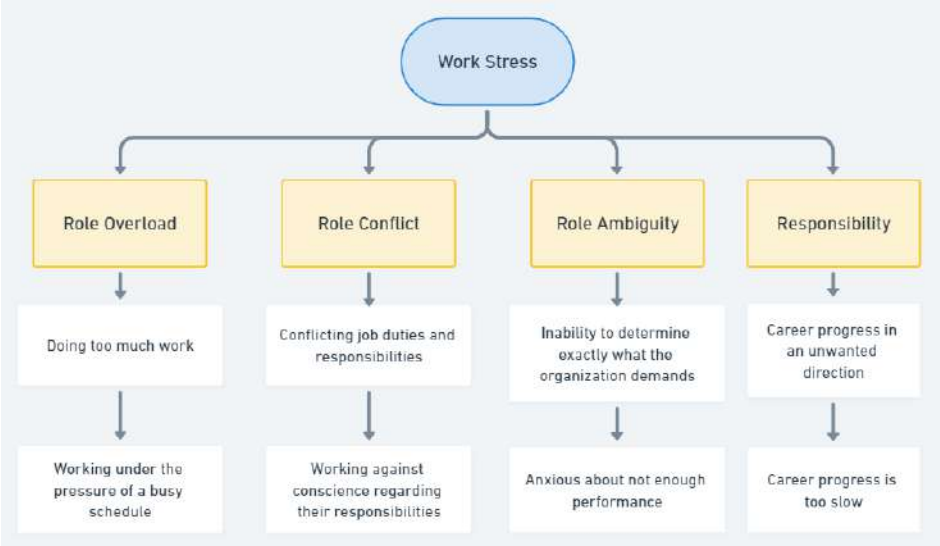
Working stress is the stress experienced by people because of their role (or job) in the organization. Working or job stress is a harmful physical and emotional response when the role requirements do not match employees' abilities, resources, and needs (Sauter et al., 1999). In this rapidly changing environment characterized by intensifying competition and increasing demands for flexibility and adaptability, organizations have made firm decisions such as outsourcing, downsizing, and mergers to adapt to new situations (Hellgren & Sverke, 2003). Job stressors such as workload, working conditions, and expectations from management cause tension and can lead to poor employee health (Beehr & Glazer, 2015). Stress can evoke negative emotions such as fear, frustration, sadness, and anger (Beehr & Glazer, 2015). Demographic variables such as age, gender, occupation, health status, education, and social support can also affect work stress (Matteson & Ivancevich, 1990). Men and women experience many of the same stressors (Barling et al., 2005). Other experts, Kreitner and Kinicki (2005), suggested that stress is an adaptive response linked by individual psychological characteristics and processes, which is a consequence of any external action, situation, or event that places particular mental and physical demands on a person or their job. Individual-level stressors are directly related to personal and work (person-job interface). The most common examples of individual-level stressors are:

1. Role Overload is a condition in which employees have too much work to do or are under pressure from a busy schedule.
2. Role conflict occurs when employees have duties and responsibilities that conflict with one another. This conflict also occurs when employees should do something contrary to their conscience or morals.
3. Role ambiguity occurs when the job itself is not clearly defined. Because employees are unable to determine what the organization is asking of them, they are constantly worried about whether their performance is sufficient or lacking.

4. Responsibility is related to employee career advancement. Career progress that is too slow, too fast, or in an unwanted direction will cause employees to experience high-stress levels. Especially if they have to be responsible for someone else's career will cause the stress level to be higher.

Robbins (2006) defined job stress as a dynamic condition in which individuals face opportunities, constraints, or demands related to their desires. However, the results are uncertain but essential. According to Robbins, indicators or dimensions of work stress are; task demands, role demands, interpersonal demands, organizational structure, and leadership. Considering the review literature, we applied Kreitner and Kinicki's (2005) perspective.

Figure 1. Kreitner and Kinicki's (2005) working stress structure



The instrument consists of 48 items with four domains, and the participants responded to the items by answering in 6 options, whether they feel the statement describes their psychological state from "very describing me" (score 1) to "very not describing me" (score 6).

Table 1. The dimensions and examples of items.

Dimensions	Indicators	Items (examples)
1. Role overload	a. Doing too much work	I am very uncomfortable doing too much work
	b. Working under the pressure of a busy schedule	I am stressed with a busy work schedule
2. Role Conflict	a. Conflicting job duties and responsibilities	I am overwhelmed by the demands of a job that is different from my responsibilities
	b. Working against conscience regarding their responsibilities	I feel the conflict within myself when doing a task that goes against my conscience

3. Role Ambiguity	a. Inability to identify the organization's demands precisely	I am worried about the results of my work if I am not able to provide what my boss asks for
	b. Anxious about not having enough performance	I am worried that my performance will not meet the expectations of my superiors
4. Responsibility	a. Career progress in an unwanted direction	I am very disappointed because the opportunity for career advancement is not paid attention to by my superiors
	b. Career progress is too slow	I feel dissatisfied because my career progress in the company is too slow

Method

Participants

We recruited employees in the IT department of a retail company in Indonesia. We applied census by distributing the instrument via Google Forms to 120 employees in this department. After inspecting the quality of the data, we dropped some incompleted data and had 91 data. The composition of the participants is 59 male and 32 female, age 21 – 31 years old (M = 27, SD =4).

Psychometric testing methods

We implemented three psychometric tests in this study: content and construct validity for the validity testing and reliability testing. In content validity, we involved expert judgment in evaluating the items in qualitative and quantitative approaches. In the qualitative method, the expert delivered their opinion and evaluation on the relevance and statement of the items. Quantitatively, we applied the content validity index (S-CVI = scale-CVI) with scores 1 and 2 indicating the poor item and scores 3 and 4 as good items, whether the items were relevant to the dimension and written in simple, clear, and unambiguous sentences.

In the construct validity testing for the scale, we employed corrected item-total correlation with Pearson correlation. Here we intended to identify whether the items accurately measured the construct. The critical value of this analysis was 0.174 (df = 89, alpha = 0.05). For the reliability testing, we used a standard of 0.70 to interpret the coefficient reliability (Anastasia & Urbina, 2004).

We also designed a simulation of a norm, a standard, to interpret the scores of the scale. We implemented the within-group norm method by dividing the score distribution into three groups: low, medium, and high. We applied percentiles 25, 50, and 75 for categorizing.

Result

Content validity

In the qualitative analysis, three experts with psychology backgrounds gave inputs to revise some sentences such as hating work to being anxious and unhappy being under pressure from work. In the quantitative analysis, we conducted a CVI analysis and calculated the scale level of the CVI. The result of S-CVI is presented in table 2. The result revealed that the content of the scale was valid.

Table 2. The SCVI coefficients of the Working Stress Scale

Content validity Index	Relevance	Clarity	Simplicity	Unambiguity
S-CVI/UA	0,77	0,71	0,71	0,75
S-CVI/AVE	0,92	0,90	0,89	0,92

**S-CVI/UA = Scale-CVI universal agreement of experts; S-CVI/AVE = Scale-CVI average.*

Construct validity

Corrected item-total correlation revealed that all items were valid, with coefficients ranging from 0.40 to 0.79. Table 3 describes the list of the items.

Table 3. The list of the final 48 items

No	Bahasa Indonesia	English	item-total (total score of 48 items)
1	Saya sangat tidak nyaman melakukan pekerjaan yang terlalu banyak	I am very uncomfortable doing too much work	0.621
2	Pekerjaan lebih dari satu membuat saya kewalahan	Multiple jobs overwhelm me	0.470
3	Saya merasa kesal melakukan banyak pekerjaan yang diberikan oleh atasan	I feel annoyed doing a lot of work given by my superiors	0.671
4	Saya merasa terbebani ketika mengerjakan lebih dari satu pekerjaan	I feel overwhelmed when I do more than one job	0.584
5	Saya merasa terbebani karena atasan memberikan pekerjaan yang melebihi tanggung jawab saya	I feel burdened because my boss gives me work that exceeds my responsibility	0.622
6	Pekerjaan saya terlalu banyak sehingga membuat saya kelelahan secara psikologis	I have so much work that it makes me psychologically exhausted	0.651
7	Saya terbebani ketika harus bekerja dengan jadwal kerja yang padat	I am burdened when I have to work with a busy work schedule	0.624
8	Saya tertekan dengan jadwal pekerjaan yang padat	I am stressed with a busy work schedule	0.677
9	Saya kewalahan dengan jadwal pekerjaan yang padat.	I am overwhelmed with a busy work schedule.	0.578
10	Jadwal pekerjaan yang padat membuat saya tidak dapat beristirahat	Busy work schedule makes me unable to rest	0.741

11	Tekanan jadwal yang padat membuat saya tidak dapat berkonsentrasi	The pressure of a busy schedule keeps me from concentrating	0.744
12	Jadwal pekerjaan yang padat membuat saya kehilangan waktu pribadi	Busy work schedule makes me lose personal time	0.631
13	Saya bingung dengan jenis tugas yang tidak sesuai	I'm confused by the incompatible task type	0.574
14	Tanggung jawab yang tidak sesuai dengan pekerjaan membuat saya terbebani	Responsibilities that are not by work burden me	0.670
15	Saya kewalahan dengan tuntutan pekerjaan yang berbeda dengan tanggung jawab saya	I am overwhelmed by the demands of a job that is different from my responsibilities	0.631
16	Saya gelisah ketika tanggung jawab pekerjaan saya saling bertentangan	I get restless when my job responsibilities conflict with each other	0.662
17	Tugas yang bertentangan dengan tanggung jawab membuat saya merasa konflik batin	Tasks that conflict with responsibilities make me feel inner conflict	0.646
18	Saya merasa cemas hasil pekerjaan saya tidak berkualitas karena bertentangan dengan tanggung jawab saya	I feel worried that my work is not of good quality because it conflicts with my responsibilities	0.576
19	Saya merasakan konflik dalam diri sendiri ketika mengerjakan tugas yang bertentangan dengan hati nurani saya	I feel the conflict within myself when doing a task that goes against my conscience	0.507
20	Saya merasa konflik batin ketika mengerjakan pekerjaan yang berbeda dengan keiinginan saya	I feel inner conflict when doing work that is different from what I want	0.653
21	Konflik moral yang saya rasakan dalam bekerja membuat saya tidak nyaman melakukan tugas-tugasnya	The moral conflict that I feel at work makes me uncomfortable doing the tasks	0.727
22	Saya sangat kecewa kepada atasan ketika diberikan tugas yang tidak sesuai dengan bidang kompetensi saya	I am very disappointed with my superiors when given assignments that are not in accordance with my area of competence	0.789
23	Saya lelah untuk mengerjakan pekerjaan yang bertentangan dengan hati nurani saya	I'm tired of doing work that goes against my conscience	0.723
24	Saya menjadi mudah marah ketika pekerjaan saya tidak sesuai dengan apa yang saya inginkan	I get angry easily when my work is not the way I want it to be	0.621
25	Saya bingung mengerjakan pekerjaan saya karena peran saya ambigu	I am confused about doing my job because my role is ambiguous	0.780

26	Saya cemas atas hasil pekerjaan saya apabila tidak mampu memberikan yang diminta atasan	I am worried about the results of my work if I am not able to provide what my boss asks for	0.403
27	Saya tidak mampu menentukan apa yang sebenarnya diharapkan oleh perusahaan	I am unable to determine what the company actually expects	0.762
28	Saya mudah patah semangat ketika saya tidak mampu memahami permintaan atasan	I get discouraged easily when I can't understand my boss's requests	0.599
29	Saya kesal terbebani menerima umpan balik buruk dari atasan terkait pekerjaan saya	I resent the burden of receiving bad feedback from my boss regarding my work	0.523
30	Saya cemas ketika sulit memahami tugas yang diberikan atasan	I worry when it is difficult to understand the tasks given by superiors	0.765
31	Saya khawatir kinerja yang saya lakukan tidak sesuai dengan ekspektasi atasan	I am worried that my performance will not meet the expectations of my superiors	0.583
32	Saya cemas ketika kinerja saya belum memenuhi indikator pencapaian minimal	I am worried when my performance does not meet the minimum achievement indicators	0.738
33	Saya gelisah ketika memikirkan kinerja yang saya lakukan belum maksimal	I get nervous when I think that my performance is not optimal	0.698
34	Saya sangat kecewa pada diri sendiri karena kinerja saya belum maksimal	I was very disappointed in myself because my performance was not optimal	0.718
35	Saya cemas ketika pekerjaan yang saya lakukan tidak tepat dengan yang diharapkan oleh atasan	I am worried when the work I do is not in accordance with the expectations of the boss	0.664
36	Saya kesulitan tidur nyenyak di malam hari karena cemas memikirkan pekerjaan	I have trouble sleeping well at night because I worry about work	0.682
37	Saya sangat kecewa karena kesempatan kenaikan jenjang karir kurang diperhatikan atasan	I am very disappointed because the opportunity for career advancement is not paid attention to by my superiors	0.729
38	Saya cemas karena tidak ada jenjang karir yang menjanjikan pada perusahaan	I am worried because there is no promising career path in the company	0.735
39	Saya kesal karena tidak mendapatkan promosi naik jabatan	I was upset that I didn't get a promotion	0.725
40	Saya merasa tertekan ketika ditolak saat mengajukan kenaikan jabatan	I felt pressured when I was turned down when I applied for a promotion	0.725
41	Saya cemas pada perkembangan karir saya	I am worried about my career development so I want to resign from this company	0.633

	sehingga ingin mengundurkan diri dari perusahaan ini		
42	Saya merasa tertekan dengan ketidakjelasan perkembangan karir di perusahaan	I feel pressured by the unclear career development in the company	0.653
43	Saya merasa tidak puas karena kemajuan karir di perusahaan terlalu lambat	I feel dissatisfied because my career progress in the company is too slow	0.534
44	Saya merasa gelisah dengan kebijakan karir di perusahaan	I feel uneasy about the company's career policy	0.576
45	Saya kecewa karena tidak memiliki kemajuan karir yang sesuai dengan kinerja saya	I was disappointed because I didn't have career advancement that matched my performance	0.532
46	Saya cemas karena prestasi kerja saya tidak memberikan kemajuan pada karir saat ini	I am worried that my work performance is not advancing my current career	0.538
47	Saya menyesal karena kurangnya prestasi kerja membuat kemajuan karir saya lambat	I regret that the lack of work performance has slowed my career progress	0.717
48	Saya tertekan ketika melihat rekan kerja memiliki karir yang lebih maju dari saya	I get depressed when I see co-workers who have more advanced careers than me	0.779

Reliability

The reliability testing showed that the Cronbach Alpha for the scale was 0.974. Table 4 explains the reliability coefficients of each dimension. According to Anastasia and Urbina (2004), the scale is reliable.

Table 4. Reliability Coefficients of Working Stress Scale for each dimension

Dimensions	Cronbach Alpha
Role Overload	0.917
Role Conflict	0.923
Role Ambiguity	0.917
Role Responsibility	0.925

The norm for score interpretation (a simulation)

The distribution score was from 73 to 263 (M = 148.71, SD = 45,56). With P_{25} = 110 (Low Stress); P_{50} = 140 (Medium Stress); dan P_{75} = 179.5 (High Stress). We assumed that individuals in the medium category could cope well with work stressors such as role overload, role conflict, role ambiguity, and responsibility. There are some stressors that cannot be handled well by individuals, but individuals can handle other work stressors well. For example, employees very uncomfortable doing too much work and employees feel worried about their performance. In the low-stress category, we think employees can cope with pressure in the work environment and with coworkers. In the high-stress group, we suppose employees cannot cope with their emotions, and the pressure is too intense.

Based on this category, we made a simulation of the interpretation. We found 32 participants in the low group, 40 in the medium group, and 19 in the high group. This simulation showed that working from home is not stressful for some employees. We hypothesized it is due to flexible working hours; therefore, the employees can adjust their working hours according to their needs. They can stretch and take a break between working hours to reduce stress. This study's results align with research by Utami et al. (2021) on Jakarta employees. They found that working from home made employees have normal stress levels.

References

Abid, M., & Safiih, A. R. (2021). Pengaruh Motivasi Kerja Terhadap Kinerja Karyawan Pada Pt. Putra Mulia Telecommunication Selama Memberlakukan Wfh (Wrok From Home). *Jurnal Arastirma*, *1*(2), 216. https://doi.org/10.32493/arastirma.v1i2.12360

Aisah, S. N. (2020). Pengaruh Gaya Kepemimpinan terhadap Kinerja Karyawan. *Bulletin of Management and Business*, *1*(2), 42–50. https://doi.org/10.31328/bmb.v1i2.100

Anastasia, A., & Urbina, S. (2004). *Psychological Testing* (9th ed.). Prentice Hall.

Azizmunaji, A. (2018). *Risiko Kesehatan Menghantui Pekerja IT? Ini Solusinya*. Kompas.Com. https://www.kompasiana.com/ahmadazizmunaji5372/5b1f29cfcaf7db5b07608b72/resiko-kesehatan-menghantui-pekerja-it-ini-solusinya?page=2&page_images=1

Barling, J., Kelloway, E., & Frone, M. (2005). *Handbook of Work Stress*. SAGE Publications, Inc. https://doi.org/10.4135/9781412975995

Beehr, T. A., & Glazer, S. (2015). A cultural perspective of social support in relation to occupational stress. In *Exploring Theoretical Mechanisms and Perspectives* (pp. 97–142). https://doi.org/10.1016/S1479-3555(01)01011-3

Goleman, D. (1997). *Emotional inteligence* (3rd ed.). Gramedia Pustaka Utama.

Hellgren, J., & Sverke, M. (2003). Does job insecurity lead to impaired well-being or vice versa? Estimation of cross-lagged effects using latent variable modelling. *Journal of Organizational Behavior*, *24*(2), 215–236. https://doi.org/10.1002/job.184

Ishak, M., & Mangundjaya, W. L. (2020). Pengelolaan Stres dan Peningkatan Produktivitas Kerja Selama Work From Home pada Masa Pandemi COVID-19. *Jurnal Psikologi Udayana*, *7*(2), 93–109. https://doi.org/10.24843/JPU.2020.v07.i02.p.09

Kreitner, R., & Kinicki, A. (2005). *Organizational Behavior* (5th ed.). Salemba Empat.

Matteson, M. T., & Ivancevich, J. M. (1990). Controlling Work Stress: Effective Human Resource and Management Strategies. *Organization Studies*, *11*(1), 151–152. https://doi.org/10.1177/017084069001100117

PricewaterhouseCoopers Survey (2020). When everyone can work from home, what's the office for? [PWC's US Remote Work Survey – June 25, 2020]. PwC. https://pwc-return-t-work-survey.pdf

Robbins, S. P. (2006). *Perilaku Organisasi* (R. Saraswati & F. Sirait (eds.)). Salemba Empat.

Sauter, S., Murphy, L., Colligan, M., Swanson, N., Joseph Hurrell, J., Frederick Scharf, J., & Grubb, R. S. P. (1999). Stress at Work. *Centers for Disease Control and Prevention, US Department of Health and Human Services*, 99–101. https://www.cdc.gov/niosh/docs/99-101/default.html

Utami, D., Latifah, N., Andriyani, A., & Fajrini, F. (2021). Gambaran Tingkat Stres dalam Pelaksanaan Work From Home Selama Masa Pandemi COVID19 di DKI Jakarta. *Muhammadiyah Public Health Journal*, *1*(2), 40–51.

Adaptation of three psychological measurement instruments assessing impact of the COVID-19 pandemic: Fear of COVID-19 scale, coronavirus anxiety scale, and altruism

Angela Oktavia Suryani
Laura Francisca N. Sudarnoto
Benedicta Prihatin D. Riyanti

Atma Jaya Catholic University of Indonesia

https://doi.org/10.37517/978-1-74286-697-0-05

Angela Oktavia Suryani is an assistant professor at the Faculty of Psychology, Atma Jaya Catholic University of Indonesia, specializing in psychological measurement and cross-cultural psychology research methods. Her researches are on psychological test constructions and adaptation, as well as cross-cultural psychological aspects comparisons.

Abstract

The COVID-19 pandemic has been going on for more than two years and has caused problems for the world community. One of the impacts of COVID-19 is psychological barriers, including decreased well-being, increased stress, anxiety, depression, loss of family members, and being laid off. Is there still energy in a difficult situation like this to give empathy and have a sense of wanting to help without reward (altruism)? The psychological condition of the impact of COVID-19 is the background for the need to carry out valid and reliable psychological measurements. This study aims to identify the psychometric properties of three psychological instruments adapted to the Indonesian culture. These instruments measure the psychological impact related to the context of COVID-19, including Fear of COVID-19, Anxiety toward COVID-19, and Altruism. This research is part of an international umbrella research themed: "International and Multidimensional Perspectives on the Impact of COVID-19 across Generations (IMPACT-C19)," which involves more than 16 countries. This study implemented a convenience sampling method to recruit 176 participants. The adaptation stages included translating the instruments from English into Indonesian (using the translator's discussion method), back-translation from Indonesian to English, harmonizing the translation results, and statistical testing. We applied Confirmatory Factor Analysis (CFA) to test the instrument's validity and the McDonald's Omega method to test reliability. The results showed that the Fear of COVID-19 scale was valid with adequate goodness of fit values, namely χ^2 (21) = 375.47, p = 0.10, TLI = 0.98 (recommended 0.95), CFI = 0.97 (recommended = 0.95)), RMSEA = 0.06 (suggested < 0.08), with a reliability coefficient of 0.81. The anxiety toward the

COVID-19 scale is also valid with excellent goodness of fit values, namely χ^2 (10) = 336.59, p = 0.34, TLI = 1.00, CFI = 1.00, RMSEA = 0.03, with a reliability coefficient of 0.83. The results of the Altruism CFA test showed that all items were valid with good goodness of fit values, namely χ^2 (10) = 147.37, p = 0.22, TLI = 0.99, CFI = 0.96, RMSEA = 0.05, with a reliability coefficient of 0.70. The research results showed that the three instruments are valid with good goodness of fit and adequate reliability coefficient.

Introduction

WHO (World Health Organization) officially declared Coronavirus disease 2019 (Coronavirus Disease 2019, shortened to COVID-19) as a pandemic (WHO, 2019) on 12 March 2020. The impact of COVID-19 has caused severe health problems and death. Actions to prevent transmission are maintaining social distance and independent isolation (Cameron et al., 2020). These efforts have been carried out by governments worldwide, including Indonesia, by implementing social distancing policies, namely imposing Large-Scale Social Restrictions (LSSR) (Government of Indonesia, 2020).

Regulation of LSSR implementation is limited to regional scope as an addition to health quarantine at homes and hospitals. This policy follows Law No. 6 of 2018 concerning health quarantine in Indonesia (Government of Indonesia, 2018). Even though it prevents more people from getting infected, the implementation of this procedure significantly impacted the community's social and economic life. Some industrial organizations cannot survive and lay off their employees (Taufik & Ayuningtyas, 2020). This problem also occurs in other countries (Vaziri et al., 2020). In addition to social and economic impacts, there are also psychological impacts experienced by the world community, such as decreased mental health and psychological well-being, increased stress, anxiety, and depression (Hamza et al., 2020; Ifdil et al., 2020; Marmarosh et al., 2020; Rodríguez-Hidalgo et al., 2020). Other psychological impacts are negative attitudes, stereotyping, and stigma for individuals infected with COVID-19 (Earnshaw et al., 2020). The results of Özdin & Özdin's research (2020) show that the groups most affected psychologically are women, individuals with a history of psychiatric illness, individuals who live in urban areas, and those with chronic co-morbidities. This study's results align with Pedraza's research (2020) that women are more likely to feel anxious than men during the COVID-19 pandemic; older people feel less anxious even though they belong to a group that is vulnerable to transmission of COVID-19.

Lakhan et al. (2020) found that the general population's prevalence rates of depression, anxiety, stress, sleep problems, and psychological distress were higher during the COVID-19 pandemic. This deplorable atmosphere of life needs special attention and service from psychologists. Therefore, in response to this pandemic situation, a research team from Indonesia joined in umbrella research with the theme: "International and Multidimensional Perspectives on the Impact of COVID-19 across Generations (IMPACT-C19)". The study is cross-cultural research on measuring psychological aspects related to COVID-19 organized by The Research Initiatives Work Group (RIWG) and the American Psychological Association (APA) COVID-19 Task Force, led by Radosveta Dimitrova from Stockholm University, Sweden, and Rita Rivera from Albizu University Miami, USA. This cross-cultural research involved researchers from 16 countries. This research represents Indonesia. In this cross-cultural research, the participating countries use the same measurement tools. Before measuring and comparing test scores between cultures, researchers should examine the instrument to what extent the participants in

those countries understand the item in the same psychological meaning considering that culture shapes human behavior (Segal, 1986; Triandis & Suh, 2002). Therefore, we adapted three instruments measuring the psychological impact of COVID-19: fear of COVID-19, anxiety toward COVID-19, and altruism.

Anxiety towards COVID-19

Anxiety is a state of apprehension or worry that complaints that something terrible will happen soon (Nevid et al., 2017). Anxiety is a natural reaction that can be experienced by anyone when seeing various situations or events that can trigger anxiety. Kaplan et al. (2010) stated that anxiety is a human response to a threatening situation. It occurs ordinarily in human development when s/he faces a change in new experiences and finds self-identity and the meaning of life. At low levels, anxiety helps individuals to stay alert, to take steps to prevent danger and minimize the impact of that hazard. But if anxiety is very high, it will be disturbing. This excessive anxiety will create behavior that is not adaptive and may endanger the person. For that reason, anxiety cannot be considered a mild disorder. The symptoms can lead to anxiety disorder if we do not prevent and cure it. According to Adwas et al. (2019), anxiety disorders manifest in mood disorders, thinking, behavior, and physiological activity. General symptoms of anxiety that accompany disturbances in sleep, concentration, and social and or occupational functioning. Anxiety is associated with restlessness, difficulty concentrating, going blank, muscle tension, irritability, and feeling tense and tired quickly. Causes of anxiety include stress, physical conditions, genetics, and environmental factors.

Individuals' negative experiences and exposure to negative issues in mass or social media regarding the growing health crisis can add to their level of anxiety and fear (Kumar & Somani, 2020; Lee, 2020b; Shuja et al., 2020). It happens because individuals do not think clearly and rationally when reacting to COVID-19, followed by very high levels of anxiety (Ahorsu et al., 2020; Lee, 2020a). Nikčević and Spada (2020) noted several studies showing that the pandemic situation affects psychological conditions beyond the pandemic disease (stress, post-traumatic stress, health anxiety, and suicidal ideation). Puspita et al. (2021) conducted an exploratory study involving teenagers in Indonesia that showed that the majority (80%) of the participants were at a mild level of anxiety, a small number were at a moderate level (9%), and severe level (11%).

Attention to anxiety about COVID-19 has prompted several experts to develop anxiety measurement tools for COVID-19, such as The Fear of COVID-19 Scale (Ahorsu et al., 2020), the first measuring tool to identify fear of COVID-19 in general. Another researcher is Lee (Lee, 2020a, 2020b), who developed the Coronavirus Anxiety Scale (CAS). This measuring tool is reliable and valid for correlation with other psychological variables related to COVID-19, such as fear of COVID-19. CAS also has good discrimination power in measuring psychological-functional disorders; therefore, it is valid for research on COVID-19 and clinical practice (Nikčević & Spada, 2020).

Fear of COVID-19

Individuals experiencing a fearful situation, especially situations of unforeseen uncertainties, understandably give rise to physiologically and psychologically tense moods. Usually, they sense fear at an ordinary (short-term) level as a response to a stressor. It is a biologically based protective mechanism to avoid threats to the body (Hardi, in Ermolaev et al., 2020). The fear experienced related to COVID-19 can be more aggressive and have a negative impact due to excessive information in society, even the

existence of incorrect information (hoaxes) circulating widely on social media. Researchers from various countries have studied and researched the phenomenon of fear (fear) of the COVID-19 pandemic. For example, Wakashima et al. (2020) examined a sample of 450 Japanese citizens. Results show the factor structure of the Japanese FCV-19S (the fear of COVID-19 scale), including seven items, and one factor was equivalent to the original FCV-19S. The scores of FCV-19S have a positive association with anxiety and depression and with perceived susceptibility to infection (Wakashima et al., 2020). Other researchers, Ermolaev et al. (2020), have examined the psychological characteristics of Russian citizens' social fear in the context of the COVID-19 pandemic. The study results reveal that social fear associated with fear of failure and defeat, rejection and suppression, loss, communication, and independence is dominant among citizens actively broadcasting information and news about COVID-19.

Altruism (related COVID-19 pandemic)

Altruism is an essential behavior related to the existence and survival of various species worldwide, including humans. The survival and maintenance of life and social control are highly dependent on the altruism of members at multiple levels. Altruism refers to any organism's behavior to increase others' welfare at their own expense, even at the risk of their own life. Altruism is unconditional kindness without expecting that kindness to happen to him again. Altruistic people will assist and achieve satisfaction by helping others (Nielsen, 2010; Penner et al., 2005; Smith et al., 2006). Penner et al. (2005) explain the existence of internal motivation as a characteristic of altruistic behavior. Altruism is a moral behavior necessary to maintain social balance and survival.

Altruistic behavior develops from the moral teachings of religion. Altruism also develops due to cultural and moral influences. The psychoanalysis theory views the superego as responsible for filtering human behavior, including altruistic behavior. It collaborates with culture, norms, and social values to direct human behavior (Nielsen, 2010). Another perspective shows that altruistic behavior is proven to be influenced by parenting styles. Longitudinal research from Hastings et al. (2000) found that authoritative parenting was associated with more prosocial behavior. Other researchers found a link between genes, personality, social environment, and helping behavior (Penner et al., 2005).

Related to the COVID-19 pandemic, Grimalda et al. (2021) study in US and Italy found that infected people donated more to others than those who did not infect. In Indonesia, the concept of *gotong royong* during the pandemic is salient among urban residents. *Gotong royong* is a joint activity/working together to achieve the desired result without expecting any return. Helping behavior, such as giving donations, information, food, etc., increased during the pandemic (Faedlulloh, 2021).

Research questions

Considering the need to explore the psychological impact of the COVID-19 pandemic, we need to have valid and reliable psychological instruments to measure the effect. Here, we formulated the research question regarding that need: what is the validity and reliability of the adapted instrument?

Method

Procedure

The procedure for adapting measuring instruments in this study follows the guidelines for translation and equivalence testing of measuring instruments used in cross-cultural research (Abubakar et al., 2013). The procedure includes the following five stages: forward translation, back-translation, harmonization step, quality check dan statistical evaluation.

Instruments

Fear of COVID-19 Scale

In this study, we adapted the fear of COVID-19 Scale developed by Ahorsu et al. (2020). This instrument consists of seven statements regarding fear and worries about the dangerous COVID-19. Indicators include symptoms of fear, such as thoughts about dying from contracting the virus, worry about hearing news about this virus, palpitations, difficulty sleeping at the thoughts of this virus, etc.

Corona Anxiety Scale

The adaptation involved Coronavirus Anxiety Scale (CAS) developed by Lee (2020a). This measuring instrument contains five statements that measure anxiety symptoms reflected in the body, including (1) Dizziness (dizziness/headache), (2) Sleep disturbance, (3) Tonic immobility (feeling paralyzed suddenly), (4) Appetite loss (loss of appetite), and (5) Abdominal distress (tension/abdominal pain). Participants should respond to what extent they experienced these symptoms during the past two weeks, from not at all (score 0) to nearly every day (score 4).

Altruism

We adapted a subscale of the Prosocial Personality Battery, self-report altruism, constructed by Penner et al. (2002). The scale consists of five items comprising helping behavior to neighbors and strangers. The participants should state whether they do those behaviors frequently, from never (score 1) to very often (score 5).

Participants

We recruited participants via a convenience sampling method utilizing students and lecturer networks in the university. Participants in this study were 176 students aged 18-63 years (M = 23.91, SD = 8.62), dominated by women as much as 77.3%. Based on religious demographic data, the participants are primarily Christian (Catholics, 45.5%; Protestants, 27.3%) since most participants were students at our university. At the time of data collection, most participants (93.8%) stated that they had never had an infection by COVID-19.

Analysis methods

The psychometric tests carried out in this study included validity and reliability tests. The instrument validity test comprised content validity: expert judgment, readability, face validity, and construct validity test (confirmatory factor analysis). The goodness of fit criteria used in this study include the chi-square coefficient with $p > 0.05$, CFI > 0.90, TLI > 0.95 and RMSEA < 0.08 (Hu & Bentler, 1999). The reliability test applied to the McDonald Omega formula. We utilized JASP 0.14.3 to analyze the data.

Results

Translation

The forward and back translations from English to Bahasa Indonesia included six scholars in psychology and two translators with a psychology background. Seven people did the forward translation, and one did the back translation. The translator worked independently.

Harmonization step

We employed a group discussion to evaluate the translation. When we disagreed on a translated word, for example, that did not match the Indonesian culture, we returned to the original term and discussed its psychological meaning. We ended the harmonization step when we reached a consensus for each item.

Cognitive interview

We asked participants to fill out the instrument and evaluate whether the participants related to the items, understood the statement easily, and unambiguity. We selected two students who reported that all items relevant to them and the statement was clear.

Statistic testing

Fear of COVID-19 Scale

The results of the CFA testing showed that all items measuring fear of COVID-19 were valid with adequate goodness of fit values, namely $\chi 2$ (21) = 375.47, p = 0.10, TLI = 0.98, CFI = 0.97, RMSEA = 0.06. Reliability with McDonald's Omega shows a very good coefficient of 0.81. Table 1 describes the factor loading of each item in this scale.

Table 1. The factor loading of adapted Fear of COVID-19 items

No.	Original	Bahasa Indonesia	Factor loading
Fear1	I am most afraid of COVID-19	Saya sangat takut terhadap COVID-19	0.47***
Fear2	It makes me uncomfortable to think about COVID-19	Saya merasa tidak nyaman ketika memikirkan COVID-19	0.62***
Fear2	My hands become clammy when I think about COVID-19	Tangan saya menjadi basah/lembab ketika memikirkan COVID-19	0.55***
Fear4	I am afraid of losing my life because of COVID-19	Saya takut kehilangan nyawa karena terkena COVID-19	0.74***
Fear5	When watching news and stories about COVID-19on social media, I become anxious	Ketika menyimak berita dan kisah-kisah mengenai COVID-19 di media sosial, saya menjadi gelisah	0.70***
Fear6	I cannot sleep because I am worrying about getting COVID-19	Saya tidak dapat tidur karena kuatir terkena COVID-19	0.47***

Fear7	My heart races or palpitates when I think about getting COVID-19	Detak jantung saya sampai berdebar-debar atau berdetak cepat ketika memikirkan terjangkit COVID-19	0.72***

****sig p < 0.001*

Coronavirus Anxiety Scale

The results of the CFA testing showed that the five items were valid in the construct validity test with excellent goodness of fit values, namely $\chi 2$ (10) = 336.59, p = 0.34, TLI = 1.00, CFI = 1.00, RMSEA = 0.03 (recommendation < 0.08). Reliability with McDonald's Omega shows a very good coefficient of 0.83. Table 2 displays the factor loading of the items.

Table 2. Factor loading of CSA-adapted items

No.	Original	Bahasa Indonesia	Factor loading
CAS1	I felt dizzy, lightheaded, or faint	Saya merasa pening, sakit kepala, atau merasa lemah	0.81***
CAS2	I had trouble falling or staying asleep	Saya mengalami kesulitan tidur	0.72***
CAS3	I felt paralyzed or frozen	Saya merasa seperti lumpuh atau membeku	0.55***
CAS4	I lost interest in eating	Saya menjadi kehilangan selera makan	0.63***
CAS5	I felt nauseous or had stomach problems	Saya merasa mual atau sakit perut	0.77***

Self-report Altruism

The results of the CFA testing showed that all items were valid with very good goodness of fit values, namely $\chi 2$ (10) = 147.37, p = 0.22, TLI = 0.99, CFI = 0.96, RMSEA = 0.05. Reliability with McDonald's Omega shows an adequate coefficient of 0.70. Table 3 illustrates the factor loading of each item of this scale.

Table 3. Factor loading of adapted Self-report Altruism items

No.	Original	Bahasa Indonesia	Factor loading
SRA1	I have helped carry a stranger's belongings (e.g., books, parcels, etc.)	Saya pernah membantu seorang yang tidak saya kenal membawa barang-barangnya (seperti buku, bingkisan, dsb)	0.78***
SRA2	I have allowed someone to go ahead of me in a line (e.g., supermarket, copying machine, etc.)	Saya pernah mengijinkan seseorang mendahului saya dalam sebuah antrian (seperti di supermarket, ketika *foto-copy*, dsb)	0.54***
SRA3	I have let a neighbor whom I didn't know too well borrow an item of some value (tools, dish, etc.)	Saya pernah mengijinkan tetangga yang tidak terlalu saya kenal meminjam sebuah barang (perkakas, piring, dsb)	0.81***

SRA4	I have, before being asked, voluntarily looked after a neighbor's pets or children without being paid	Saya pernah secara suka rela (sebelum diminta) menjaga binatang peliharaan atau anak-anak tetangga tanpa upah	0.67***
SRA5	I have offered to help a handicapped or elderly stranger across a street	Saya pernah menawarkan bantuan kepada seorang asing penyandang disabilitas atau lansia untuk menyeberangi jalan	0.77***

Conclusion

This study aims to examine the validity and reliability three adapted instrument measuring the psychological impact of the COVID-19 pandemic, the fear of COVID-19 Scale, Coronavirus Anxiety Scale, and Self-Report Altruism. The adaptation process follows the stages proposed by Abubakar et al. (2013). Participants in this study were Atma Jaya Catholic University students, both from undergraduate and postgraduate levels. The data collection process uses the Qualtric online media carried out by the IMPACT-C19 team. Data collection using this media reached 301 participants, but only 176 participants could process complete data. However, the amount of data is still sufficient to be processed using the CFA method.

The results of the validity test showed that all measuring instruments were valid with satisfactory goodness of fit CFA values. In general, it can be concluded that all items are relevant for use in the Indonesian group. However, this research has not involved participants from other parts of Indonesia who are assumed to have a culture that is different from the conditions of the COVID-19 pandemic in the Greater Jakarta area where most of the participants in this study came from. Recommendations for further research, measuring instruments can be tested on a wider group that covers the complete territory of Indonesia (Western, Central and Eastern Indonesia). The results of reliability testing using McDonald's Omega show that all measuring instruments are reliable with good coefficients. This shows that the items of the measuring instrument consistently measure the construct to be measured. It can be concluded that this measuring tool is ready to be used in subsequent studies.

References

Abubakar, A., Dimitrova, R., Adams, B. G., Jordanov, V., & Stefenel, D. (2013). *Procedures for translating and evaluating equivalence in questionnaires for use in cross-cultural studies.* Bulletin of the *Transilvania, 6(56), 79–86.*

Adwas, A. A., Jbireal, J.M., & Azab, A. E. (2019). Anxiety: Insights into Signs, Symptoms, Etiology, Pathophysiology, andTreatment. The South African journal of medical sciences October 2019. http://www.easpublisher.com/easjms/

Ahorsu, D. K., Lin, C. Y., Imani, V., Saffari, M., Griffiths, M. D., & Pakpour, A. H. (2020). The fear of COVID-19 Scale: Development and initial validation. *International Journal of Mental Health and Addiction, 27*, 1–9. https://doi.org/10.1007/s11469-020-00270-8

Brown, S., Doom, J. R., Lechuga-Peña, S., Watamura, S. E., & Koppels, T. (2020). Stress and parentingduring the global COVID-19 pandemic. Child Abuse Negl 2020 Aug 20:104699. doi: 10.1016/j.chiabu.2020.104699

Cameron, C. D., Hutcherson, C. A., Ferguson, A. M., Scheffer, J. A., Eaton, L. A., & Kalichman, S. C. (2020). Social and behavioral health responses to COVID- 19: lessons learned from four decades of an HIV pandemic. *Journal of behavioral medicine*, *43*(3), 341–345. https://doi.org/10.1007/s10865-020-00157-y.

Earnshaw, V., Brousseau, N., Hill, E., Kalichman, S., Eaton, L., & Fox, A. (2020). Anticipated stigma, stereotypes, and COVID-19 testing. *Stigma and Health. 5*, *4*, 390–393.doi: 10.1037/sah0000255.

Ermolaev, V. V., Sorokoumova, E. A., Voroncova, J., Nasonova, D. K., & Chetverikova, A. I. (2020). Psychological features of social fears associated with the COVID-19 content of news feed inRussia. Eurasia J Biosci 14: 2403-2409.

Grimalda, G., Buchan, N. R., & Ozturk, O. D. (2021). Exposure to COVID-19 is associated with increased altruism, particularly at the local level. *Sci Rep* 11, 18950. https://doi.org/10.1038/s41598-021-97234-2

Hadjiandreou, E., & Inzlicht, M. (2019). Empathy is hard work: People choose to avoid empathy because of itscognitive costs. *Journal of Experimental Psychology:General, 148*, 962-976. doi:/10.1037/xge0000595

Hamza, C. A., Ewing, L., Heath, N. L., & Goldstein, A. L. (2020). When social isolation is nothing new: A longitudinal study psychological distress during COVID-19 among university students with and without preexisting mental health concerns. *Canadian Psychology/Psychologie canadienne.* Advance online publication. doi:10.1037/cap0000255

Hastings, P. D., Zahn-Waxler, C., Robinson, J., Usher, B., & Bridges, D. (2000). The development of concern for others in children with behavior problems. *Developmentalpsychology*, *36*(5), 531–546. https://doi.org/10.1037/0012-1649.36.5.531

Ifdil, I., Yuca, V., & Yendi, F. M. (2020). Stress and anxiety among late adulthood in Indonesia during COVID-19 outbreak. *Jurnal Penelitian Pendidikan Indonesia*, *6*, 31-46. doi:10.29210/02020612

Indonesian Government [Pemerintah Indonesia]. (2018). *Undang-Undang Nomor 6 Tahun 2018 tentang Kekarantinaan Kesehatan* (PDF), Jakarta: Kementerian Sekretariat Negara Republik Indonesia

Indonesian Government [Pemerintah Indonesia]. (2020). *Peraturan Pemerintah Nomor 21 Tahun 2020 tentang Pembatasan Sosial Berskala Besar dalam rangka Percepatan Penanganan Corona VirusDisease 2019 (COVID-19)* (PDF), Jakarta: Kementerian Sekretariat Negara Republik Indonesia

Innamorati, M., Ebisch, S. J. H., Gallese, V., & Saggino, A. (2019). A bidimensional measure ofempathy: Empathic Experience Scale. *PLoS ONE* 14(4): e0216164. https://doi.org/10.1371/journal. pone.0216164

Kaplan, H. I., Sadock, B. J., & Grebb, J. A. (2010). Kaplan-Sadock Sinopsis Psikiatri Ilmu Pengetahuan Perilaku Psikiatri Klinis. Dalam *Gangguan Kecemasan.* Tangerang: Binarupa Aksara.

Kubb, C., & Foran, H. M. (2020). Measuring COVID-19 Related Anxiety in Parents: Psychometric Comparison of Four Different Inventories. *JMIR mental health*, *7*(12), e24507.https://doi.org/10.2196/24507

Kumar, A., & Somani, A. (2020). Dealing with Corona virus anxiety and OCD. *Asian Journal of Psychiatry, 51, 102053*. doi:https://doi.org/10.1016/j.ajp.2020.102053

Lakhan, R., Agrawal, A., & Sharma, M. (2020). *Prevalence of Depression, Anxiety, and Stress during COVID-19 Pandemic. Journal of Neurosciences in Rural Practice.*https://www.thieme-connect.com/products/ejournals/html/10.1055/s-0040-1716442

Lee, S. A. (2020a). How much "Thinking" about COVID-19 is clinically dysfunctional? *Brain,behavior, and immunity*, *87*, 97–98. https://doi.org/10.1016/j.bbi.2020.04.067

Lee, S. A. (2020b). Coronavirus Anxiety Scale: A brief mental health screener for COVID-19 related anxiety. Death Studies, 44(7), 393–401. https://doi.org/10.1080/07481187.2020.174848116

Marmarosh, C. L., Forsyth, D. R., Strauss, B., & Burlingame, G. M. (2020). The psychology of the COVID-19 pandemic: A group-level perspective. *Group Dynamics: Theory, Research,and Practice, 24*, 122-138. doi:10.1037/gdn0000142

Maximiano-Barreto, M. A., Fabrício, D. M., Luchesi, B. M., & Chagas, M. H. N. (2020). Factorsassociated with levels of empathy among students and professionals in the health field: a systematic review. *Trends Psychiatry Psychother*.;42(2):207-215. http://dx.doi.org/10.1590/2237-6089-2019-0035.

Muncer, S. J., & Ling, J. (2006). *Psychometric analysis of the empathy quotient (EQ) scale,40(6), 1111–1119.* doi:10.1016/j.paid.2005.09.020

Nevid, J. S., Rathus, S. A., & Greene, B. (2017). Abnormal Psychology in a Changing world. 10th edition. New Jersey: Prentice-Hall

Nielsen, M. L. (2010). Incorporating a concept of altruism into psychoanalytic theory*. International Journal of Applied Psychoanalytic Studies, 7(4), 262–280.* doi:10.1002/aps.248

Nikčević, A. V., & Spada, M. M. (2020). The COVID-19 anxiety syndrome scale: Development and psychometric properties. *Psychiatry Research, 292, 113322.* https://doi.org/10.1016/j.psychres.2020.113322.

Özdin, S., & Özdin, Ş. B. (2020). Levels and predictors of anxiety, depression, and health anxiety during COVID-19 pandemic in Turkish society: The importance of gender. *International Journal of Social Psychiatry*, 66(5), 504–511 https://journals.sagepub.com/doi/full/10.1177/0020764020927051

Pedraza, P. (2020). *Life dissatisfaction and anxiety in COVID-19 pandemic*. Academia.edu.https://www.academia.edu/89787545/Life_Dissatisfaction_and_Anxiety_in_COVID_19_pandemic

Penner, L. A., Dovidio, J. F., Piliavin, J. A., & Schroeder, D. A. (2005). Prosocial Behavior: Multilever Perspective. Annu. Rev. Psychol.56:365–92. doi: 10.1146/annurev.psych.56.091103.070141.

Penner, L. A., Fritzsche, B. A., Craiger, J. P., & Freifeld, T. S. (1995). Measuring the prosocialpersonality. In J.N. Butcher & C.D. Spielberger (Eds.), *Advances in Personality Assessment,10,* 147-163.

Puspita, I. M., Rozifa, A. W., & Nadhiroh, A. M. (2021). Gambaran Kecemasan Dan Kepatuhan remaja Putri Terhadap Kebiasaan Baru di Masa Pandemi COVID-19. *Journal of Midwifery Science)*, 5(1), 52–61. http://jurnal.univrab.ac.id/index.php/jomis/article/view/1492

Rodríguez-Hidalgo, A. J., Pantaleón, Y., Dios, I., & Falla, D. (2020). Fear of COVID-19, Stress, and Anxiety in University Undergraduate Students: A Predictive Model for Depression. *Frontiers Psychology,11*:591797. doi:10.3389/fpsyg.2020.591797

Segal, M. H. (1986). Culture and behavior: psychology in global perspective. *Annual Reviewof Psychology, 37,* 523-564. doi: 10.1146/annurev.ps.37.020186.002515

Senese, V. P., De Nicola, A., Passaro, A., & Ruggiero, G. (2018). The Factorial Structure ofa 15-Item Version of the Italian Empathy Quotient Scale. *European Journal of Psychological Assessment, 34*, 344-351. https://doi.org/10.1027/1015-5759/a000348

Shuja, K. H., Aqeel, M., Jaffar, A., & Ahmed, A. (2020). COVID-19 pandemic and impending global mental health implications. *Psychiatria Danubina, 32*, 32–35. https://doi.org/10.24869/psyd.2020.32

Smith, S. W., Smith, S. L., Pieper, K. M., Yoo, J. H., Ferris, A. L., Downs, E., & Bowden, B. (2006). Altruism on American Television: Examining the Amount of, and Context Surrounding, Acts of Helping and Sharing**.** *Journal of Communication, 56*(4), 707–727. https://doi.org/10.1111/j.1460-2466.2006.00316.x

Taufik, T., & Ayuningtyas, E. A. (2020). Dampak pandemi COVID 19 terhadap bisnis dan eksistensi platform online. *Jurnal Pengembangan Wiraswasta*, *22, 1*, 21-32. Doi: 10.33370/jpw.v22i01.389.

Triandis, H. C., & Suh, E. M. (2002). Cultural influence on personality. *Annual Review of Psychology, 53,* 133-160. doi: 10.1146/annurev.psych.53.100901.13520

Vaziri, H., Casper, W. J., Wayne, J. H., & Matthews, R. A. (2020). Changes to the work-familyinterface during the COVID-19 pandemic: Examining predictors and implications using latent transition analysis. *Journal of Applied Psychology, 105, (10)*, 1073-1087. doi: 10.1037/apl0000819.

Wakashima, K., Asai, K., Kobayashi, D., Koiwa, K., Kamoshida, S., & Sakuraba, M. (2020). The Japanese version of the Fear of COVID-19 scale: Reliability, validity, and relation to copingbehavior. *PLoS ONE* 15(11): e0241958. https://doi.org/ 10.1371/journal.pone.0241958

World Health Organization (WHO). (2019). *WHO announces COVID-19 outbreak as pandemic.* Retrieved January 4th, 2021. https://www.euro.who.int/en/health-topics/health-emergencies/coronavirus-COVID-19/news/news/2020/3/who-announces-COVID-19-outbreak-a-pandemic.

Confirmatory factor analysis of the Perception Academic Stress Scales (PAS)

Damajanti Kusuma Dewi
Miftakhul Jannah
Ira Darmawanti

Surabaya State University, Indonesia

https://doi.org/10.37517/978-1-74286-697-0-06

Damajanti Kusuma Dewi was born in Madiun on October 27, 1970. She was graduated from the University of Putra Bangsa Surabaya majoring in psychology. Then she had a magister's degree in psychology on psychometry from Gadjah Mada University Yogjakarta Indonesia in 1998. in 2020 received a doctorate degree from Malang State University, with a concentration in Educational Psychology.

Currently a representative of the psychology program at Surabaya State University. Before joining the psychology program, she joined the educational technology program. Digital storytelling was her project in 2012, which was presented at Universitat Valencia Spain; Cooperative learning to improve math skills, developing science modules to improve academic achievement are two of her publications in 2011. She is also active in social engagement by having mentorship for traditional play among children in Surabaya.

Dewi was active in the psychology association in Indonesia. She also had some research on online guidance and counselling. She made a module on project-based learning to improve higher-order thinking in 2015 and create a profile for teachers of early childhood education. She is one of the committees in the early childhood education association specialized in curriculum development.

Abstract

Stress is a condition that must be experienced by humans, including students. This happens because every student will always be faced with a condition to meet the demands of the environment. Students will experience stress if they feel they do not have the ability to meet these academic demands. Therefore, an instrument is needed that can measure academic stress felt by students. This study aims to test the construct validity of the Perceived Academic Stress Scale (PASS)-Adaptation of Bedewy & Gabriel. Before conducting factor analysis, the adaptation and modification process of the instrument was carried out. The adaptation process was carried out to change the language, from English to Indonesian. The modification process was carried out to adjust the items to the situation and conditions of Indonesian students. The research subjects came from Surabaya State University students, who were in semesters 1, 3 and 5. The validity of the PAS construct was sought using Confirmatory Factor Analysis (CFA). To obtain the factors underlying the PAS-Bedewy & Gabriel Adaptation, it was done through

2 stages. Based on the five stages of extraction, 3 supporting theoretical factors were obtained. Factor 1: which relates to perceived academic load, supported by items no. 3, 6, 7 and 8; Factor 2: which relates to examinations, supported by items no. 10 and 11. Factor 3: which relates to confidence in completing assignments or lectures, supported by items no. 5, and 13. The study supports Bedewy & Gabriel's first theoretical construct, which is based on 3 factors namely Factor 1: pressures to perform; Factor 2: perceptions of workload; Factor 3: academic self-perceptions.

Introduction

Everyone will always face a challenging environment, whether as a child or an adult. Each environmental challenge will be faced by each person in their own way. When a person faces environmental challenges, they will also perform assessment activities. Whether the environmental challenge is stressful or not.

A person who perceives that the environmental challenge is very high, so that it exceeds the capacity of the ability to overcome it, will experience stress [1], [2]; or the ability to adjust to it [3]. Stress as an imbalance between environmental demands and the individual's ability to meet these demands [4]. The greater the gap between environmental demands and the ability to cope with these demands, the greater the impact on the person [5]; including tension, distress, and anxiety [5]

In recent years, "stress has become an important topic in academic circles" [2], [6]. Academic stress has been identified as a detrimental problem across different countries, cultures, and ethnic groups [7]. Academic stress is students' perception of the pressure they face, time constraints to complete assignments, academic workload, and their academic self-perception [8]. Academic stress is students' perceptions of the pressures they face, time constraints to complete assignments, academic workload, and their academic self-perception [8]

Research shows that there are many factors that cause academic stress. Academic stress caused by academic activities, including inappropriate studying habits [9]; time management problems [9]; academic assignments (Pascoe et al., 2020), including tasks or projects that are very difficult. In addition to the stress associated with academic tasks, academic stress is also caused by social and emotional factors, including conflict or situations faced with professors or peers [10][2], and other people living in the same domestic premises [2]; the feeling of not belonging, and lack of emotional support [10]. Another factor that can cause academic stress is the amount of time spent on academic mobility [9].

Actually, academic stress is a natural condition experienced by every student. However, if it is not handled from the start, it will worsen the student's condition. therefore, it is necessary to identify the symptoms of academic stress early, which can be seen in anxiety, depression, decrease exercise, changes in eating habits, and sleep disturbance (Backović et al., 2012); digestive disorder (Ongori, 2007). These symptoms are experienced by students more than once [11]; in addition, the symptoms increase every month [12]

While some experts say that academic stress can have a positive effect, others say the opposite [13]. The main impact of academic stress is a decrease in academic performance [14] , which leads to a decrease in academic achievement [15]. Students who experience academic stress, will consider that all academic tasks are too overwhelming [16]; this has the impact of feeling overwhelmed by their academic tasks [17]and feeling worried about the success of their studies.[14], [17], [18][19]

Given the magnitude of the impact of academic stress on student academic performance, an early treatment is needed that can prevent academic stress from getting worse. One way to detect academic stress early on is to periodically measure students' academic stress levels. Therefore, a standardized academic stress measurement instrument is needed.

Several studies developed academic stress scales [20], including:

1. Scale for Assessing Academic Stress (SAAS) was developed by Sinha, Sharma, and Mahendra [1]
2. Perceived Stress Scale (PSS), was developed by Cohen [21]
3. Student-Life Stress Inventory (SSI) was developed by Gadzella [22]
4. Lakaev Academic Stress Response Scale (LARS) was developed by Lakaev [23]
5. Depression Anxiety Stress Scale (DASS) was developed by Lovibond and Lovibond [24]
6. Perception of Academic Stress Scale (PAS) was by Bedewy and Gabriel [8]

The purpose of the study was to develop a measurement scale through the process of adaptation and modification of PAS - Bedewy and Gabriel (2015) for Indonesian students.

Method

In this study, PASS - Bedewy and Gabriel (2015) was developed through two stages, including:

1. Translation stage, i.e., from a. from English to Bahasa Indonesia; and b. from Bahasa Indonesia to English. At this stage, two translators were needed. The first translator is a lecturer from the English department, the second translator is a lecturer from the Psychology department who has proficiency in English. The purpose of the back translation was to verify whether the initial version was equivalent to the original scale. The result at this stage was the PAS – The First Format Adaptation.

Table 1. PAS – The First Format Adaptation

No	Original Version	Translation 1 (Indonesian Version)	Translation 2 (English Version)
	Stresses related to academic expectations		*Stresses related to academic expectations*
1	Competition with my peers for grades is quite intense	*Persaingan dengan teman-teman saya untuk nilai cukup ketat*	The competition between I and friends is quite tight for getting a good score.

2	My teachers are critical of my academic performance	*Guru saya kritis terhadap akademik saya*	My teacher is critical of my academics
3	Teachers have unrealistic expectations of me	*Guru memiliki harapan yang tidak realistis tentang saya*	The teacher has unrealistic expectations about me
4	The unrealistic expectations of my parents stress me out	*Harapan yang tidak realistis dari orang tua saya membuat saya stress*	The unrealistic expectations from my parents stress me out
	Stresses related to faculty work and examinations		*Stresses related to faculty work and examinations*
5	The time allocated to classes and academic work is enough	*Waktu yang dialokasikan untuk kelas dan pekerjaan akademik sudah cukup*	The time allocated for class and academic work is enough
6	The size of the curriculum (workload) is excessive	*Ukuran kurikulum (beban kerja) berlebihan*	The load of Curriculum is excessive
7	I believe that the amount of work assignment is too much	*Saya percaya bahwa jumlah penugasan terlalu banyak*	I believe that the number of assignments is too much
8	Am unable to catch up if getting behind my work	*Saya tidak dapat mengejar ketinggalan pekerjaan*	I can't catch up with work
9	I have enough time to relax after work	*Saya punya cukup waktu untuk bersantai setelah bekerja*	I have enough time to relax after work
10	The examination questions are usually difficult	*Pertanyaan ujian biasanya sulit*	The questions in the examination are usually difficult
11	Examination time is short to complete the answers	*Waktu ujian singkat untuk menyelesaikan jawaban*	The time allotment in the examination is too short to complete the answers
12	Examination times are very stressful to me	*Waktu ujian sangat menegangkan bagi saya*	The time during examination is very stressful for me
	Stresses related to students' academic self-perceptions		*Stresses related to students' academic self-perceptions*
13	Am confident that I will be a successful student	*Saya yakin bahwa saya akan menjadi siswa yang sukses*	I am sure that I will become a successful student
14	Am confident that I will be successful in my future career	*Saya yakin bahwa saya akan sukses dalam karir masa depan saya*	I am sure that I will succeed in my future career

15	I can make academic decisions easily	*Saya dapat membuat keputusan akademis dengan mudah*	I can make academic decisions easily
16	I fear failing courses this year	*Saya takut gagal kursus tahun ini*	I am afraid of failing the course this year
17	I think that my worry about examinations is weakness of character	*Saya pikir kekhawatiran saya tentang ujian adalah kelemahan karakter*	I think my concern about the exam is character flaws
18	Even if I pass my exams, am worried about getting a job	*Bahkan jika saya lulus ujian, saya khawatir tentang mendapatkan pekerjaan*	Even if I pass the exam, I'm worried about getting a job

2. Modification Stage: the next step is to distribute the first test format to limited subjects, to students who have the same characteristics as the research subjects. The results of the limited try out were reorganized into the Second Format Adaptation Scale.

Table 2. PAS – The Second Format Adaptation

No	Translation (Indonesian Version)	Translation (English Version)	Modification
1.	Persaingan dengan teman-teman saya untuk nilai cukup ketat	*The competition between I and friends is quite tight for getting a good score.*	Saya dan teman-teman bersaing sangat ketat untuk mendapatkan nilai yang bagus.
2.	Guru saya kritis terhadap akademik saya	*My teacher is critical of my academics*	Dosen saya memonitoring dan mengevaluasi nilai saya
3.	Guru memiliki harapan yang tidak realistis tentang saya	*The teacher has unrealistic expectations about me*	Saya merasa tidak sanggup memenuhi target akademik yang ditetapkan dosen saya
4.	Harapan yang tidak realistis dari orang tua saya membuat saya stress	*The unrealistic expectations from my parents stress me out*	Saya merasa stress jika dituntut orang tua meraih prestasi akademik
5.	Waktu yang dialokasikan untuk kelas dan pekerjaan akademik sudah cukup	*The time allocated for class and academic work is enough*	Saya merasa waktu yang ditetapkan oleh dosen dalam tugas terstruktur dan mandiri cukup
6.	Ukuran kurikulum (beban kerja) berlebihan	*The load of Curriculum is excessive*	Saya merasa beban semester berlebihan
7.	Saya percaya bahwa jumlah penugasan terlalu banyak	*I believe that the number of assignments is too much*	Saya menganggap jika tugas-tugas dari dosen berlebihan

8.	Saya tidak dapat mengejar ketinggalan pekerjaan	*I can't catch up with work*	Saya tidak bisa menyelesaikan tugas-tugas dari dosen
9.	Saya punya cukup waktu untuk bersantai setelah bekerja	*I have enough time to relax after work*	Saya merasa kewalahan dengan tugas yang banyak karena sering menunda mengerjakan
10.	Pertanyaan ujian biasanya sulit	*The questions in the examination are usually difficult*	Menurut saya soal-soal yang dibuat dosen dalam ujian sulit
11.	Waktu ujian singkat untuk menyelesaikan jawaban	*The time allotment in the examination is* too short *to complete the answers*	Menurut saya waktu dalam mengerjakan ujian singkat/kurang
12.	Waktu ujian sangat menegangkan bagi saya	*The time during examination is very stressful for me*	Menurut saya setelah ujian adalah waktu bebas
13.	Saya yakin bahwa saya akan menjadi siswa yang sukses	*I am sure that I will become a successful student*	Saya merasa yakin bahwa IPK saya akan naik
14.	Saya yakin bahwa saya akan sukses dalam karir masa depan saya	*I am sure that I will succeed in my future career*	Saya tidak yakin saya akan bekerja menjadi apa setelah lulus
15.	Saya dapat membuat keputusan akademis dengan mudah	*I can make academic decisions easily*	Saya sudah memiliki rencana pengerjaan skripsi
16.	Saya takut gagal kursus tahun ini	*I am afraid of failing the course this year*	Saya takut IPK semester ini turun
17.	Saya pikir kekhawatiran saya tentang ujian adalah kelemahan karakter	*I think my concern about the exam is character flaws*	Menurut saya ujian harus dihadapi bukan dikhawatirkan
18.	Bahkan jika saya lulus ujian, saya khawatir tentang mendapatkan pekerjaan	*Even if I pass the exam, I'm worried about getting a job*	Saya merasa kelulusan tidak menjanjikan langsung mendapatkan pekerjaan

3. Try Out Stage: At this stage, using 1553 research subjects, the correlation coefficient moved from 0.935 to 0.944. The results of the perceived academic stress scale reliability test showed a coefficient of 0.942 alpha Cronbach.

4. Factor Analysis Stage: the subjects used in this study are different from the subjects used in the try-out stage. The details are as follows:

Table 3. Distribution of Subjects Based on Gender

	Gender	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Male	392	20.09	20.09	20.09
	Female	1559	79.91	79.91	100.00
	Total	1951	100.00	100,00	

Table 4. Distribution of Subjects Based on Semesters

	Semester	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	5	1156	59.25	59.25	59.25
	3	487	24.96	24.96	84.21
	1	308	15.79	15.79	100,00
	Total	1951	100.00	100.00	

Table 5. Distribution of Subjects Based on Faculty

	Department	Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Arts and Language	204	10.46	10.46	10.46
	Business and Economy	302	15.48	15.48	25.94
	Math and Natural Science	230	11.70	11.70	37.72
	Engineering	271	13.80	13.80	51.61
	Law and Social Science	149	7,64	7,64	59.25
	Education	795	40.75	40.75	100.00
Total		1951	100.00	100.00	

Factor analysis on PASS - Bedewy & Gabriel Adaptation was conducted in 5 stages. The table below illustrates the items that failed at each stage, because they did not meet the criteria.

Table 6. Stage 1-4 of PAS Adaptation Factors Analysis

Factor	Item	No	Stage 1	Stage 2	Stage 3	Stage 4
Academic Expectation	The competition between I and friends are quite tight	1	.504	.439		
	My teacher is critical of my academics	2	.539	.538	.630	.629
	My teacher has unrealistic expectations of me	3	.525	.593	.611	.630

	The unrealistic expectations from my parents stress me out	4	.375			
Workload and Examination	The time allocated for class and academic work is enough	5	.529	.555	.612	.645
	The load of curriculum is excessive	6	.577	.563	.622	.637
	I believe that the number of assignments is too much	7	.621	.625	.640	.667
	I can't catch up with work	8	.595	.589	.640	.640
	I have enough time to relax after work	9	.298			
	The questions in the examination are usually difficult	10	.583	.523	.627	.683
	The time allotment in the examination is too short to complete the answers	11	.547	.533	.688	.737
	The time during examination is very stressful to me	12	.347			
University Students' Academic Self-perception	I am sure that I will become a successful student	13	.586	.651	.695	.824
	I am sure that I will succeed in my future career	14	.492			
	I can make academic decision easily	15	.481			
	I am afraid of failing the course this year	16	.346			
	I think my concern about this exam is character flaws	17	.601	.478		
	Even if I pass the exam, I'm worried about getting a job	18	.585	.611	.488	
	Total	18				

Based on table 6 above, it appears that only 8 out of 18 items proceed to stage 5.

Tabel 7. KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.746
Bartlett's Test of Sphericity	Approx. Chi-Square	2503.119
	Df	28
	Sig.	.000

Based on the results above, the KMO MSA value is 0.746. Because the KMO value is> 0.5, factor analysis can be continued because it meets the requirements for the KMO MSA value.

Tabel 8. Communalities

	Initial	Extraction
PAS3	1.000	.574
PAS5	1.000	.519
PAS6	1.000	.580
PAS7	1.000	.539
PAS8	1.000	.642
PAS10	1.000	.571
PAS11	1.000	.669
PAS13	1.000	.708

Extraction Method: Principal Component Analysis

Based on the results above, all items have an extraction value greater than 0.5, so it can be concluded that all items can be used to explain the factors.

Table 9. Rotated Component Matrix[a]

	Component		
	1	2	3
PAS3	.735	.097	.155
PAS5	-.058	.313	.646
PAS6	.694	.296	-.108
PAS7	.581	.448	.028
PAS8	.795	-.024	.099
PAS10	.239	.716	.023
PAS11	.064	.812	.078
PAS13	.191	-.155	.805

Based on the results table 9 above, it can be seen that all factor loading values are greater than 0.5. Based on these results, it can be concluded that:

- Factor 1, consisting of items 3, 6. 7 and 8
- Factor 2, consisting of items 10 and 11
- Factor 3, consisting of items 5 and 13

Conclusion

This research resulted in 3 factors, namely:
1. Factor 1: related to perceived academic load, supported by items no. 3, 6, 7 and 8;
2. Factor 2: related to exams, supported by item no. 10 and 11
3. Factor 3: which relates to the confidence of completing assignments or lectures, supported by item no. 5, and 13.

The results of this study support the theoretical construct used by Bedewy and Gabriel to develop the PASS in college students, which consists of 3 factors, namely: (1) the academic expectations subscale (four items), (2) the workload and examinations subscale (eight items), and (3) the students' academic self-perceptions sub-scale (six items). These items related to academic stress were converted into a 5-point Likert-type questionnaire of 18 items, resulting in PAS.

However, the results of research in the field do not match the theoretical concept. Where based on factor analysis, 4 supporting factors are obtained, namely:
Factor 1: pressures to perform.
Factor 2: perceptions of workload.
Factor 3: academic self-perceptions.
Factor 4: time restraints

In addition, Bedewy and Gabriel also revised their PAS, which can be seen in the table below:

Tabel 9. Comparison of PAS with PAS-Revised

No	Original Version	The final version of the Perceptions of Academic Stress (PAS) scale
1	Competition with my peers for grades is quite intense	Am confident that I will be a successful student
2	My teachers are critical of my academic performance	Am confident that I will be a successful in my future career
3	Teachers have unrealistic expectations of me	I can make academic decisions easily
4	The unrealistic expectations of my parents stress me out	The time allocated to classes and academic work is enough
5	The time allocated to classes and academic work is enough	I have enough time to relax after work
6	The size of the curriculum (workload) is excessive	My teachers are critical of my academic performance
7	I believe that the amount of work assignment is too much	I fear failing courses this year
8	Am unable to catch up if getting behind my work	I think that my worry about examinations is weakness of character
9	I have enough time to relax after work	Teachers have unrealistic expectations of me
10	The examination questions are usually difficult	The size of the curriculum (workload) is excessive

11	Examination time is short to complete the answers	I believe that the amount of work assignment is too much
12	Examination times are very stressful to me	Am unable to catch up if getting behind the work
13	Am confident that I will be a successful student	The unrealistic expectations of my parents stress me out
14	Am confident that I will be successful in my future career	competition with my peers for grades is quite intense
15	I can make academic decisions easily	The examination questions are usually difficult
16	I fear failing courses this year	Examination time is short to complete the answers
17	I think that my worry about examinations is weakness of character	Examination times are very stressful to me out
18	Even if I pass my exams, am worried about getting a job	Even if I pass my exams, am worried about getting a job

In addition to moving the position of item numbers, one of the revisions made was the absence of information about theoretical constructs.

The factor analysis comparison between PAS - Bedewy & Gabriel 2015 and PAS - Adaptation is as follows:

Table 10. Comparison of PAS 2015 results with PAS Adaptation

No.	Factor	PASS – Bedewy & Gabriel 2015				PASS - Adaptation		
		Factor 1	Factor 2	Factor 3	Factor 4	Factor 1	Factor 2	Factor 3
	Stresses related to academic expectations							
1	Competition with my peers for grades is quite intense	1 (0.75)						
2	My teachers are critical of my academic performance	1 (0.41)						
3	Teachers have unrealistic expectations of me				4 (0.52)	1		
4	The unrealistic expectations of my parents stress me out	1 (0,72)						

	Stresses related to faculty work and examinations							
5	The time allocated to classes and academic work is enough				4 (0.59)			3
6	The size of the curriculum (workload) is excessive		2 (0.78)			1		
7	I believe that the amount of work assignment is too much		2 (079)			1		
8	Am unable to catch up if getting behind my work				4 (0.46)	1		
9	I have enough time to relax after work				4 (0.63)			
10	The examination questions are usually difficult		2 (0.47)				2	
11	Examination time is short to complete the answers				4 (0.47)		2	
12	Examination times are very stressful to me	1 (0.50)						
	Stresses related to students' academic self-perceptions							
13	Am confident that I will be a successful student			3 (0.67)				3
14	Am confident that I will be successful in my future career			3 (0.71)				

15	I can make academic decisions easily			3 (0.45)				
16	I fear failing courses this year			3 (0.48)				
17	I think that my worry about examinations is weakness of character	1 (0.44)						
18	Even if I pass my exams, am worried about getting job		2 (0.52)					

References

Agolla, J. E., & Ongori, H. (2009). "An Assessment of Academic Stress among Undergraduate Students: The Case of University of Botswana," *Educational Research and Review*, vol. 4, no. 2, pp. 63–070, [Online]. Available: http://www.academicjournals.org/ERR

Aina, Q., & Wijayanti, P. H. (2019). "Coping the Academic Stress: The Way the Students Dealing with Stress," in *International Seminar on Language, Education and Culture*, Mar. vol. 3, no. 10, p. 212. doi: 10.18502/kss.v3i10.3903.

Backović, D., Živojinović, J., Maksimović, J., & Maksimović, M. (2012). Gender differences in academic stress and burnout among medical students in final years of education. Psychiatria Danubina, 24(2), 175–181.

Bedewy, D., & Gabriel, A. (2015). "Examining perceptions of academic stress and its sources among university students: The Perception of Academic Stress Scale," *Health Psychology Open*, vol. 2, no. 2. SAGE Publications Inc., Nov. 03. doi: 10.1177/2055102915596714.

Busari, A. O. (2012). "Identifying Difference in Perceptions of Academic Stress and Reaction to Stressors Based on Gender among First Year University Students," *Int J Humanit Soc Sci*, vol. 2, no. 14, pp. 138–146, [Online]. Available: www.ijhssnet.com

Cohen, S. (1994). "Perceived Stress Scale,". [Online]. Available: www.mindgarden.com

Eswi, A. S., Radi, S., & Youssri, H. (2013). "Stress/ stressors as perceived by baccalaureate Saudi nursing students," *Middle East J Sci Res*, vol. 14, no. 2, pp. 193–202, doi: 10.5829/idosi.mejsr.2013.14.2.734.

Fairbrother, K., & Warn, J. (2003). "Workplace dimensions, stress and job satisfaction," *Journal of Managerial Psychology*, vol. 18, no. 1–2, pp. 8–21, doi: 10.1108/02683940310459565.

Gadzella, B. M. (2022). "Student-Life Stress Inventory: Identification of and Reactions to Stressors," *Psychol Rep*, vol. 74, pp. 395–402, 1994, Accessed: Nov, 18. [Online]. Available: https://sci-hub.hkvisa.net/10.2466/pr0.1994.74.2.395#

Hill, M. R., Goicochea, S., & Merlo, L. J. (2018). “In their own words: stressors facing medical students in the millennial generation,” *Med Educ Online*, vol. 23, no. 1, Jan. doi: 10.1080/10872981.2018.1530558.

Huan, V. S., See, Y. L., Ang, R. P., & Har, C. W. (2008). “The Impact of adolescent Concerns on Their Academic Stress,” *Educ Rev (Birm)*, vol. 60, no. 2, pp. 169–178, May, doi: 10.1080/00131910801934045.

Kania, S. K. (2014). “The Relationship Between Gender Differences and Stress,” *The Huron University College Journal of Learning and Motivation*, vol. 52, no. 1, pp. 92–101.

Kaplan, D. S., Liu, R. X., & Kaplan, H. B. (2022). “School Related Stress in Early Adolescence and Academic Performance Three Years Later_The Conditional Influence of Self Expectations,” *Social Psychology of Education*, vol. 8, pp. 3–17, 2005, Accessed: Nov. 18. [Online]. Available: https://link.springer.com/content/pdf/10.1007/s11218-004-3129-5.pdf

Lakaev, N. (2022). “Validation of an Australian Academic Stress Questionnaire,” 2009. Accessed: Nov. 18. [Online]. Available: https://sci-hub.hkvisa.net/10.1375/ajgc.19.1.56

Lovibond, P. F., & Lovibond, S. H. (2022). “The Structure of Negative Emotional States: Comparison of the Depression Anxiety Stress Scales (DASS) with The Beck Depression and Anxiety Inventories,” *BehaP. Res. Ther*, vol. 33, no. 3, pp. 335–343, 1995, Accessed: Nov, 18. [Online]. Available: https://www.academia.edu/3070169/The_structure_of_negative_emotional_states_Comparison_of_the_Depression_Anxiety_Stress_Scales_DASS_with_the_Beck_Depression_and_Anxiety_Inventories

Lundgren, H. P. (2016). “Academic and General Stress among Students: The Role of Implicit Theory of Intelligence and Gender,” Thesis, University of Gothenburg.

Murberg, T. A., & Bru, E. (2004). “School-Related Stress and Psychosomatic Symptoms among Norwegian Adolescents,” *Sch Psychol Int*, vol. 25, no. 3, pp. 317–332, Aug, doi: 10.1177/0143034304046904.

Ollfors, M., & Andersson, S. I. (2007). “Ability of stress, sense of control, and self-theories to predict Swedish high school students’ final grades,” *Educational Research and Evaluation*, vol. 13, no. 2, pp. 143–169, Apr, doi: 10.1080/13803610701434241.

Ongori, H. (2007). “A review of the literature on employee turnover,” *African Journal of Business Management*, pp. 49–054, [Online]. Available: http://www.academicjournals.org/ajbm

Pascoe, M. C., Hetrick, S. E., & Parker, A. G. (2020). “The impact of stress on students in secondary school and higher education,” *International Journal of Adolescence and Youth*, vol. 25, no. 1. Routledge, pp. 104–112, Jan. 02. doi: 10.1080/02673843.2019.1596823.

Rustam, H. K., & Tentama, F. (2020). “Creating Academic Stress Scale and The Application For Students: Validity And Reliability Test In Psychometrics,” *International Journal of Scientific & Technology Research*, vol. 9, no. 01, pp. 661–667, [Online]. Available: www.ijstr.org

Schraml, K., Perski, A., Grossi, G., & Makower, I. (2012). “Chronic Stress and Its Consequences on Subsequent Academic Achievement among Adolescents,” *J Educ Develop Psychol*, vol. 2, no. 1, Apr, doi: 10.5539/jedp.v2n1p69.

Topper, E. F. (2007). “Stress in the library workplace,” *New Library World*, vol. 108, no. 11–12, pp. 561–564, doi: 10.1108/03074800710838290.

Wong, P. T. P., Wong, L. C. J., & Lonner, W. J. (2006). *Handbook of multicultural perspectives on stress and coping*. Springer.

Zia-Ur-Rehman, M., & Sharif, R. (2014). "How Perceived Stress Can Influence Academic Performance: Analyzing The Role Of Some Critical Stressors," *Journal of Contemporary Studies,* vol. 3, no. 1, pp. 51–62.

Simulation of comparing the sensitivity between response model and response time model detecting aberrant behavior

Danang Kamal Musthafa Suprananto
UIN Syarif Hidayatullah Jakarta

https://doi.org/10.37517/978-1-74286-697-0-14

Danang is a graduate in Psychometrics from the UIN Syarif Hidayatullah Jakarta. He is an experienced data analyst with deep interest in psychological assessment. He excels in R Program and other statistical software such as SPSS, JASP, and LISREL. He currently works as Research Assistant at the Australian Council for Educational Research (ACER) Indonesia.

Abstract

This study aims to determine whether there is a difference in detection rates between response model and response time model to detect aberrant behavior. Also, to determine examinee's ability estimation accuracy of each model along with checking the strengths and weakness when there is an aberrant behavior in testing data especially CBT. This research is a simulation study where concentrate on test-length, sample-size, and aberrant level with 50 replications. Analyzing parameter recovery from replicated data to check the strengths and weaknesses of each model. Further, comparing the l_z person-fit that using response's data and response time's data to see which one the most sensitive is from both models. Moreover, the estimation of examinee's ability also compared to see the accuracy of each model. The result shows that increment of aberrant level would make item parameter estimation more bias for both models. Detection rates using response time (or response time model) indicates more sensitivity than detection rates using response (or response model). Both models also showed there was an increment estimation for examinee who doing aberrant behavior, so becomes bias and invalid if used as decision-making.

Introduction

Theoretically, when a person cheats to get correct answers during exam is called aberrant behavior. The presence of aberrant behavior caused by several things such as anxiety during the test, boredom, out of time to answer all items, low motivation, guessing, random responding, cultural bias, discrepancy of response intentionally, and misunderstand of test instruction (Birenbaum, 1985; Seo & Weiss, 2013). When this behavior occurs, there's a great consequence for test developer or test taker itself (Drasgow et al., 1987; Fox & Marianti, 2017; Karabatsos, 2003; Man et al., 2018). For test developers, aberrant behavior will reduce the accuracy of the test. Mousavi and Cui (2020) showed that test information function and the

accuracy of item parameter estimation have decreased because of aberrant behavior. A study (Liu et al., 2019) that was conducting a simulation under Multidimensional Item Response Theory (IRT) also showed that aberrant behavior have impacts on model fit, such as the construct validity and reliability of the test. Therefore, methods to detect aberrant behavior are urgently required.

Detection of aberrant behavior normally conducted in IRT approach using person-fit statistics (Lord, 1980; Man et al., 2018). The goal is identifying an examinee who has a response pattern differ to expected response from a model, hence it can distinguish an examinee who classified as aberrant and non-aberrant is (Karabatsos, 2003; Reise, 1990). However, detection of aberrant behavior using person-fit statistics is analyzing unexpected response vector, so it known well as analysis of person-misfit (Meijer & Sijtsma, 1995; Reise, 1990). That's it, a person-fit that analyzing a response vector only shows a person which his/her response deviates but unable to know what type of aberrant behavior s(he) did during a test. In order to detect aberrant behavior more accurately and a transition paper-based test (PBT) into computer-based test (CBT) in current era, which technology grows rapidly (Fox, 2012, p. 227; Lee & Chen, 2011), since 1983, response time modelling as extension of IRT model has been developed as well as detection of aberrant behavior using response time (RT) data.

RT has been studied in IRT approach as an extension of IRT which accounts speed and response accuracy in a model (Thissen, 1983). Since the presence of response time modelling, aberrant behavior is expected to be more accurately detected with the availability of RT. Qian et al. (2016) whom analyzed the residual of log-response time illustrated there was an indication item pre-knowledge. The study statistically showed that aberrant behavior was significantly more in 2012 than 2010 which assumed as item pre-knowledge. Another study under lognormal response time modeling (Marianti et al., 2014) which conducted using simulation and real data also showed that person-fit was detecting 20%. When time discrimination accounted, resulting an increment around 34%. Thus, RT will help to detect aberrant behavior during a test.

Although several studies showed an increment on detection rates of aberrant behavior, cheating behavior is still the most difficult behavior to detected (Karabatsos, 2003). As a result, the score of examinee obtained from cheating would be questionable its validity since incongruent with a measurement model and becomes unfair if used as decision-making (de la Torre & Deng, 2008; Reise & Due, 1991). Therefore, this study aims to compare detection rates between response model and response time model as well as seeking the strengths and weaknesses of each model when examinees have aberrant behavior during a test.

Aberrant Behavior

When an examinee gives responses that differ from a model, there will be a discrepancy between observed response and expected response from the model. This discrepancy is considered as misfitting response, aberrant behavior, aberrant response, unexpected response, improbability response, anomalous behavior, etc. However, from several names , the most common name used is aberrant behavior (Meijer, 2003; Meijer & Sijtsma, 2001; Molenaar & Hoijtink, 1990; Mousavi & Cui, 2020; Wright & Stone, 1979, p. 66).

Imagine a person with a high ability taking a test with 30 items and the items on the test have variation of difficulty level, from the easiest to the hardest. Assume those items are sorted from the easiest to hardest, the person mentioned previously, unfortunately incorrectly answers five easy items but correctly answers the rest items. It is impossible if a person with a

high ability giving incorrect answer on easy item, and this circumstance known as aberrant (Meijer & Sijtsma, 1995). Hence, the consequence is the score will be spurious (spuriously high or spuriously low) (Karabatsos, 2003). There are several aberrant behavior types (Karabatsos, 2003; Lee & Chen, 2011; Meijer, 1996; van der Linden & Guo, 2008; Wang et al., 2018):

1. *Sleeping behavior*
   When a test begins from easy to hard, a person that did not check again his/her answer on easy items and gives incorrect answer unintentionally. Therefore, the proportion of correct response on easy items is smaller than on medium and hard items.
2. *Cheating*
   Cheating is a behavior that unfairly gets correct answer on an item that s(he) actually unable to answer correctly. If a person has low ability and cheats (such as copying answer for his/her neighbor) so s(he) will correctly answer on item that hard for him/her.
3. *Careless responding*
   An examinee's behavior occurs when incorrectly answers an item intentionally and s(he) knows the correct answer.
4. *Creative responding*
   When an examinee incorrectly answers on easy items because s(he) interprets those items uniquely and resulting creative way to answer it. This behavior frequently appears when a competent person perhaps finding items that too easy to him/her, also assume that those items are too easy to be answered (or too simple to be true).
5. *Lucky guessing*
   An examinee luckily and correctly answers on items where s(he) doesn't know the answer or unable to gives the correct answer. In general, this behavior is identical giving a quick answer but has low accuracy.
6. *Plodding*
   This behavior can be exemplified in Guttman model where items have sorted from easy to hard. There's a chance that an examiner will slowly and methodically answer the items. Also, rejecting to answer next item until s(he) can solve an item after find the correct answer.
7. *Alignment Errors*
   Discrepancy of response pattern because of mistake giving correct answer. The assumption is paper-based test (PBT), while the answer sheet and the questioned sheet are separated. There's a chance that an examiner forgets s(he) skipped a certain item and hasn't solved it yet. Unfortunately, s(he) gives answer on the skipped item. As a result, several items will incorrectly answer.
8. *Random responding*
   There will be a situation when a test is administered, and an examinee randomly selects the answer on each item especially test in multiple-choice.
9. *Deficiency of Ability*
   This behavior can be exemplified when a test consists of two sub-ability such as sub-ability A dan sub-ability B. If the easy items measure A and hard item measure B, a person that has knowledge of A would have correct answer's proportion higher than correct answer's proportion on B.
10. *Memorization*
    This behavior doesn't occur directly during a test but would happen when a test is repeated so s(he) memorizes the items that have already been given to him/her.

11. *Item Pre-knowledge*
    When there's an unusual combination of answering items and response time. The combination means an examinee correctly answers an item in a relatively quick whereas the item has small probability of correct answer (hard item).
12. *Pacing*
    A situation where an examinee only focuses on items that s(he) can answer and hopes to maximize the total score. This behavior often leaves certain item that hasn't answer yet resulting empty response or omit.

Person-fit

Person-fit is a statistical method that is used in IRT approach identifying to what extent the discrepancy between examinee's observed response and expected response from a chosen model. Person-fit statistics can be distinguished into two groups (Karabatsos, 2003; Meijer & Sijtsma, 1995, 2001), such as parametric and non-parametric where we can see on Table 1.

Table 1. 36 Person-fit Statistics

Non-Parametric Person-Fit Statistics (11)	Parametric Person-Fit Statistics (25)
G (Guttman, 1944, 1950)	*U* (Wright & Stone, 1979)
*G** (van der Flier, 1977)	*ZU* (Wright, 1980)
r_{pbis} (Donlon & Fischer, 1968)	*lnU* (Wright & Stone, 1979)
C (Sato, 1975)	*W* (Wright, 1980)
MCI (Harnisch & Linn, 1981)	*ZW* (Wright, 1980)
U3 (van der Flier, 1980)	*lnW* (Wright & Stone, 1979)
ZU3 (van der Flier, 1982)	*ECI1, ECI2, ECI3, ECI4, ECI5, ECI6, ECIlz, ECIl2z, ECI4z, ECIl6z* (Tatsuoka, 1984)
H^T (Sijtsma, 1986; Sijtsma & Mejer, 1992)	*l* (Levine & Rubin, 1979)
A, D, E_i (Kane & Brennan, 1980)	l_z (Drasgow, Levine, & Williams, 1985)
	M (Molenaar & Hojtink, 1990)
	M(*p-value*) (Bedrick, 1997)
	Item-Grouping Person-Fit Statistics
	D(θ) (Trabin & Weiss, 1983)
	l_{zm} (Drasgow, Levine, & Mclaughlin, 1991)
	UB (Smith, 1986)
	ZUB (Smith, 1986)
	lnUB (Wright & Stone, 1979)

Sources: Karabatsos (2003)

The most obvious difference between parametric and non-parametric approach is in the measurement scale while a parametric leads to measurement on interval or ratio, whereas a non-parametric leads to measurement on ordinal scale (Meijer & Sijtsma, 1995).

Method

This study used simulation where data was generated under IRT approach which would be compared in several conditions. The IRT model selected for response model and response time model was 2PL which consists item difficulty dan item discrimination. Condition in this study two IRT approach (response model and response time model) with aberrant level of N (5%, 10%, and 20% from sample-size), two different test-lengths (20 items and 40 items), also two different sample-sizes (N = 500 and N = 2.000). Therefore, the simulation study has a design 2 x 3 x 2 x 2 = 24 model data with 50 replications as well.

Manipulation of aberrant behavior conducted by taking random sample of ability θ < 0 based on previous study (Maeda & Zhang, 2020) and considering aberrant level. The randomly sampled person considered a person who was doing aberrant behavior during a test. Next, items which have item difficulty ≥ 1.0 would be chosen and five items randomly sampled as item with aberrant behavior. The response from the generated model changes into a correct response on five selected items. The assumption is a person with low ability on five hard items selected previously would answer them correctly. So, the probability of correct answer would become 100% which an indication of cheating. To make it more realistic, following on previous study (Man et al., 2018) that a certain value of time for selected items was added using uniform distribution around 30 to 120 seconds. Value that generated by uniform distribution added into response time of aberrant person on selected items. This step is assumed that aberrant persons need more addition time to copy answers from his/her neighbor (Man et al., 2018).

In order to generate data and estimate the model, this study used R Program version 4.1.2 (Team, 2021) with several packages such as "mirt" (Chalmers, 2012) for estimating response model, "PerFit" (Tendeiro, 2021) for analysis of person-fit, and "LNIRT" (Fox et al., 2021) for estimating response time model. The estimation of models was using Bayesian approach with MCMC algorithm via Gibbs Sampler for response time model. Two chains of 10.000 iterations were run and the burn-in cycle of each iteration was 2.500. For response model, "mirt" package provides MHRM as Bayesian approach. Also, two chains of 10.000 iterations and 2.500 burn-in periods were run. But, when the maximum change of iteration reaches < 0.0001, the iteration was terminated which indicated as the model converged. All processing was done in R Program.

Severals methods of generating data like standard normal distribution, lognormal distribution, uniform distribution, and so on, were used in this study. Item discrimination generated using lognormal distribution with 0.0 of meanlog and 0.3 of standard deviation, whereas item difficulty generated using standard normal distributiton. All item parameters generated at two test-length (20 and 40 items). Multivariate normal distribution with 0.0 of mean dan 1.0 of variance was used for ability parameter

Using l_z statistics (Drasgow et al., 1985) for person fit in response model, while response time model was person-fit proposed by Marianti et al. (2014) The cut-off point for response model was l_z < -1.645 (one-tailed, α = 0.05) based on previous studies (Reise & Due, 1991; Seo & Weiss, 2013). Meanwhile in the response time model, cut-off point with significant level at 5% (α = 0.05) was used. Also, Type I and Type II error rates were analyzed, where Type I error refers to what extent the proportion of non-aberrant persons detected as aberrant (False Positive), whereas Type II error refers to what extent the proportion of aberrant persons detected as non-aberrant (False Negative) (Cizek & Wollack, 2017, p. 12; Maeda & Zhang, 2020). Thus, the detection rates is an index that calculated by 1 – Type II error (Man et al., 2018).

Nevertheless, evaluation criteria that is used to compare true parameter dan estimated parameter known as parameter recovery. This step also wants to identify to what extent the difference between generated data and estimated data. Several indices were used such as bias dan mean absolute difference (MAD) (Bulut, 2015; Feinberg & Rubright, 2016; Mousavi & Cui, 2020) shown on equation (1) dan (2).

$$Bias = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_i - \theta_{True}\right)}{n}, \text{dan} \quad (1)$$

$$MAD = \frac{\sum_{i=1}^{n}|\hat{\theta}_i - \theta_{True}|}{n} \quad (2)$$

Where $\hat{\theta}_i$ is estimated parameter's value, θ_{True} is generated parameter's value, whereas *n* is the number of replications. Bias and MAD can be used for each parameter by means change θ to α for item discrimination, β for item difficulty, and so on. Mousavi and Cui (2020) suggested that range of -0.20 to 0.20 of recommended bias, whereas MAD is no more than 0.6. Each model was identified by both indices where value of *bias* and MAD that close to 0.0 which considered as better estimate.

Result and Discussion

Descriptive Statistics of Generated Data

Descriptive statistics conducted to check whether generated data was according to simulation method. One of replicated data was chosen as representation of all replications. The result of generated data will be shown on Table 2 down below.

Table 2. Descriptive Statistics of Generated Data

K	AL	Item Discrimination				Item Difficulty				Ability			
		Mean	SD	Min	Max	Mean	SD	Min	Max	Mean	SD	Min	Max
20	5	1.02	0.19	0.68	1.37	0.00	1.03	-1.84	1.72	0.00	1.00	-3.06	3.55
	10	1.02	0.20	0.44	1.35	0.00	1.10	-2.84	1.62	0.00	1.00	-3.49	2.73
	20	1.03	0.23	0.60	1.43	0.00	1.03	-1.47	1.47	0.00	1.00	-2.52	2.62
40	5	1.03	0.25	0.57	1.58	0.00	0.98	-1.77	2.68	0.00	1.00	-4.02	2.88
	10	1.05	0.29	0.24	1.83	0.00	0.78	-1.78	1.72	0.00	1.00	-3.35	2.83
	20	1.03	0.24	0.36	1.46	0.00	0.77	-1.70	1.57	0.00	1.00	-2.97	2.76

Note. K = test-length, AL = aberrant-level (%), Mean = average value, SD = standard deviation.
Sources. Personal Data (2022).

As we can see, true parameter of item discrimination which generated by lognormal distribution, has mean around 1.0 along with 0.2 of standard deviation. Highest value of item discrimination is 1.83, whereas the lowest was 0.24 at 40 items with aberrant level of 10%. At the moment, item difficulty that generated by standard normal distribution, has 0.0 of mean for all conditions along with around 1.0 of standard deviation. The lowest value of tem difficulty was -2.84 whereas the highest value was 2.68. Meanwhile for ability parameter that generated by multivariate normal distribution, shows 0.0 of mean with 1.0 of standard deviation. The lowest value was -4.02 and the highest value was 3.55.

Analysis of Parameter Recovery

Parameter Recovery on Sample-size 500 (N = 500)

Data which had been simulated and manipulated on each condition will be compare among estimated parameter and true parameter as evaluation on response model and response time model. The analysis will be used to diagnose the impact of aberrant on each model. To clarify the result, it will be separated into two analyses based on sample-size (N = 500 and 2.000). The following is the result of parameter recovery of each model on sample-size 500 respondents which has shown on Table 3.

Table 3. Parameter Recovery pada Sample-size 500

K	AL (%)	N	Response Model				Response Time Model			
			Bias		MAD		Bias		MAD	
			a	b	a	b	a	b	a	b
20	5	500	-0.062	0.055	0.131	0.389	0.032	-0.106	0.180	0.175
	10	500	-0.093	-0.049	0.162	0.405	0.105	-0.166	0.338	0.230
	20	500	-0.130	-0.163	0.211	0.332	**0.247**	**-0.229**	**0.611**	0.293
40	5	500	0.018	0.089	0.122	0.342	0.027	-0.071	0.141	0.121
	10	500	-0.025	0.084	0.142	0.342	0.078	-0.102	0.233	0.154
	20	500	-0.015	-0.049	0.161	0.275	0.198	-0.134	0.408	0.190

*Note. K = test-length, AL = aberrant level (%), a = item discrimination, b = item difficulty, MAD = mean absolute difference, **bold** = exceed criteria index. Sources: Personal Data (2022).*

According to Table 3, it can be seen that response model has fulfilled the criteria index. However, along aberrant level increased, the estimation of item discrimination and item difficulty have underestimated. It means that the higher aberrant's presence on a certain testing data, the lower item parameter estimation will be. When item discrimination decrease, the item's discriminating power to distinct low and high ability person will be lose . Whereas, the lower item difficulty estimation, the items will represent easier than they should.

In response time model, it can be seen that only at aberrant level of 20% and test-length of 20 items hasn't met evaluation criteria. In contrast with response model where item parameters decreased along with aberrant level raised, item discrimination on response time model has overestimated. Though item discrimination raised, the estimation result doesn't reflect what they should be. On item difficulty, it was consistent with response model as the estimation became underestimated and indicated that items became easier than they had to be.

Parameter Recovery on Sample-size 2000 (N = 2000)

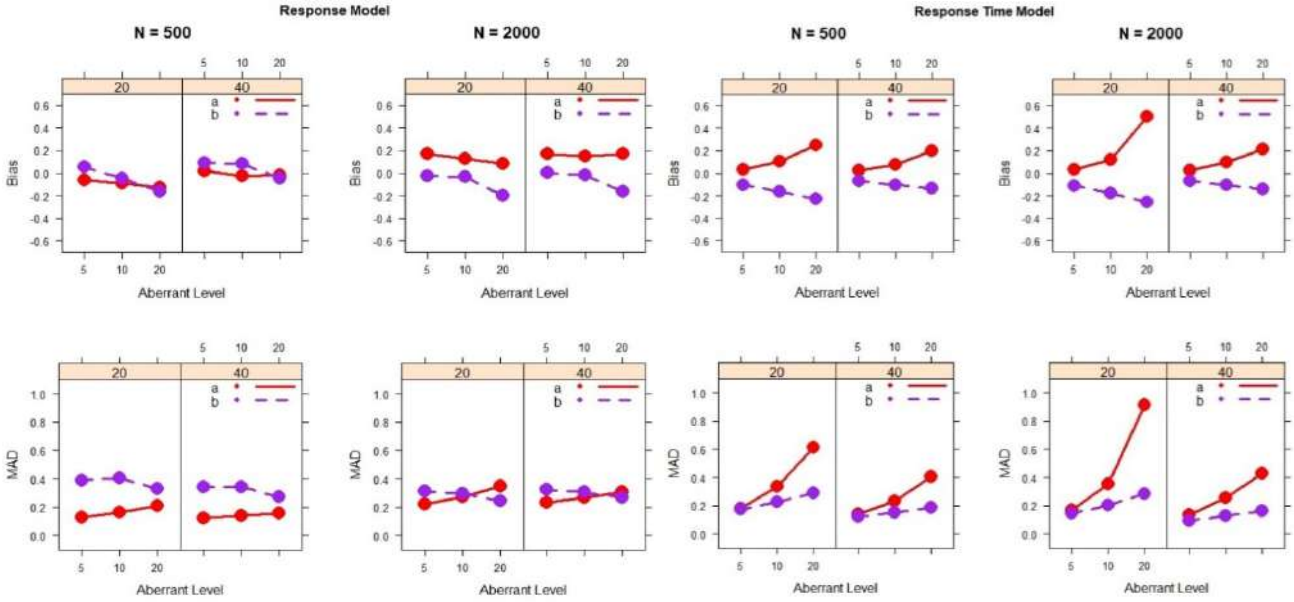
Table 4. Parameter Recovery on Sample-size 2.000

K	AL (%)	N	Response Model				Response Time Model			
			Bias		MAD		Bias		MAD	
			a	b	a	b	a	b	a	b
20	5	2000	0.173	-0.026	0.220	0.342	0.035	-0.114	0.172	0.145
	10	2000	0.127	-0.030	0.271	0.342	0.117	-0.175	0.355	0.203
	20	2000	0.084	-0.196	0.350	0.275	**0.504**	**-0.259**	**0.913**	0.284
40	5	2000	0.173	0.005	0.232	0.314	0.027	-0.071	0.133	0.095
	10	2000	0.152	-0.020	0.268	0.295	0.096	-0.106	0.258	0.130
	20	2000	0.168	-0.163	0.310	0.246	**0.211**	-0.139	0.429	0.166

*Note. K = test-length, AL = aberrant percentage (%), a = item discrimination, b = item difficulty, MAD = mean absolute difference, **bold** = exceed criteria index. Sources: Personal Data (2022)*

On Table 4, response model has fulfilled the criteria indices (bias and MAD) on all conditions. However, item discrimination estimation has the maximum evaluation criteria. Along with the increment of aberrant level, thus the item discrimination dan item difficulty also decreased on 20 items or 40 items. The result is congruent to parameter recovery on sample-size 500 that has presented on Table 3.

Figure 1. Interaction Plot of Parameter Recovery between Response Model and Response Time Model



Sources: Personal Data (2022)

On response time model, wher crimination hasn't fulfilled the criteria index like bias < 0.2 and MAD < 0.6 on model at 20 items as well as at 40 items. But at 40 items, item discrimination only exceeded bias index. The result was also consistent to sample-size 500 on Table 3 where item discrimination suffers overestimation while item difficulty suffers underestimation.

To make it easier to understand Table 3 and Table 4, interaction plot of aberrant level with criteria indices of each model will be presented on Figure 1. As we can see the higher aberrant level on response model, the more underestimates would be on both item parameters. Whereas on response time model, it clearly has shown the higher aberrant level affects underestimation on item difficulty but overestimation on item discrimination.

Analysis of Detection Rates

The analysis of detection rates represents the sensitivity of each model detecting aberrant which replicated 50 times. The result of comparison of sensitivity between response model dan response time model to detect aberrant behavior shown on Table 5.

Table 5. Comparison of Detection Rates between Response Model and Response Time Model

AL (%)	N	Detection Rates											
		Response Model						Response Time Model					
		K = 20			K = 40			K = 20			K = 40		
		D	Type I	Type II	D	Type I	Type II	D	Type I	Type II	D	Type I	Type II
5	500	0.997	0.007	0.003	0.988	0.011	0.012	1.000	0.008	0.000	1.000	0.011	0.000
10	500	0.979	0.010	0.021	0.936	0.017	0.064	0.976	0.010	0.024	0.972	0.014	0.028
20	500	0.836	0.020	0.164	0.813	0.021	0.187	0.878	0.013	0.122	0.868	0.017	0.132
5	2000	1.000	0.008	0.000	0.999	0.012	0.001	1.000	0.008	0.000	1.000	0.011	0.000
10	2000	0.985	0.012	0.015	0.957	0.017	0.043	0.978	0.010	0.022	0.979	0.013	0.021
20	2000	0.860	0.020	0.140	0.833	0.021	0.167	0.887	0.015	0.113	0.876	0.017	0.124

Note. AL = aberrant level (%), N = sample-size, K = test-length, D = detection rates, Type I = non-aberrant detected as aberrant, Type II = aberrant detected as non-aberrant. Sources: Personal Data (2022).

According to Table 5, it has shown that detection rates of both model decrease along with aberrant level increases. It means that the higher aberrant level, the more difficult aberrant behavior would be detected. Further, the detection rates on both models are vary where response time model is more sensitive to detect aberrant behavior rather than response model. The highest difference on both models is 5.5% where we can see on sample-size 500

at 40 items. Also, it can be seen that the higher aberrant level, the higher type i error on both models. This is an indication when the more aberrant persons present in a testing data, not only detection rates would be low but the chance of misclassifying non-aberrant persons as aberrant will increase as well. In simply words, it will be more difficult to decide who the aberrant persons are, and vice versa.

Analysis of Distance between Theta and Item Difficulty

The addition analysis in this study is analysis of distance between *ability* (θ) and item difficulty. The comparison between unmanipulated model and estimated model from manipulated data containing aberrant. Data was randomly chosen, and 3rd condition (N = 500, test-length = 20 items, and aberrant level = 20%) was selected from last replication which can be seen on Table 6.

Table 6. Distance between Ability and item Difficulty

Model	θ - b					
	AI		Non-AI		All Item	
	Mean	SD	Mean	SD	Mean	SD
True	-1.468	1.000	-1.441	1.000	0.000	1.027
RM	-1.088	1.088	-1.976	1.088	0.188	1.272
RTM	-0.319	0.615	-1.467	0.615	0.230	0.809

Note. AI = aberrant item, Non-AI = non-aberrant item, True = true parameter, RM = response model, RTM = response time model. Sources: Personal Data (2022).

According to Table 6, we can see there was a shift between ability and item difficulty. Before the data was manipulated, the distance between ability and item difficulty was -1.468 (b = 1.468). It means when the value from the distance accounts in IRT model, selected samples have probability of correct answer under 50% on selected items. However, when the response of selected samples has been estimated on both models, it can be seen that the distance between ability and item difficulty became closer to zero. On response model, the distance was -1.088 whereas response time model's distance was -0.319 for all sample's mean. It means, aberrant behavior would affect the distance between ability and item difficulty. In other words, it becomes hard to detect by person-fit statistics.

Analysis of Ability Estimation's Accuracy

The accuracy of ability estimation was also analyzed by separating the aberrant and all samples (total) using criteria indices (bias and MAD). In this step, each model was compared to see which model has smallest bias and MAD as indicator of better accuracy. Also, the result will be separated into two tables which sample-size 500 and sample-size 2.000.

Accuracy Comparison on Sample-size 500 (N = 500)

Table 7. Ability Estimation on Sample-size 500

K	AL	N	Response Model				Response Times Model			
			Aberrant		Total		Aberrant		Total	
			Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
20	5	500	0.807	0.815	0.059	0.355	0.660	0.665	0.000	0.301
	10	500	0.652	0.671	0.033	0.366	0.521	0.531	0.000	0.335
	20	500	0.491	0.533	0.063	0.373	0.414	0.437	0.000	0.392
40	5	500	0.250	0.323	0.070	0.335	0.338	0.357	0.000	0.211
	10	500	0.167	0.292	0.079	0.346	0.281	0.306	0.000	0.227
	20	500	0.087	0.259	0.056	0.310	0.266	0.295	0.000	0.275

Note. K = test-length, AL = aberrant level (%), Aberrant = aberrant samples, Total = all samples, N = sample-size, MAD = mean absolute difference. Sources: Personal Data (2022).

Based on Table 7, for aberrant sample at 20 items, it showed that ability estimation of response time model is lower than response model. It can be interpreted as the estimation of aberrant person's ability on response model became more spuriously high (SH) than response

time model when the sample-size was small. Moreover, it also can be seen for all samples (total) ability's estimation of response time model had surprisingly small bias under three digits round which represented to zero. This phenomenon indicated that the response time model had excessively smaller difference between true ability parameter and estimated ability parameter than response model. Nonetheless, there was a different ability estimation when aberrant level reached 20% where MAD of response model smaller than response time model. It means, when data contains so many aberrant persons, did not affect so much to all samples (total) which contains non-aberrant persons.

Furthermore, at 40 items, it was supported with previous result. On this condition, for aberrant samples, response model showed better accuracy of ability estimation than response time model. It was proven by the smaller criteria indices as better accuracy. It means, response model will have better accuracy of ability estimation for aberrant persons in a longer test. However, for all samples (total), response time model showed better ability estimation than response model based on both indices which gives smaller value.

Accuracy Comparison on Sample-size 2000 (N = 2000)

Table 8. Ability Estimation on Sample-size 2.000

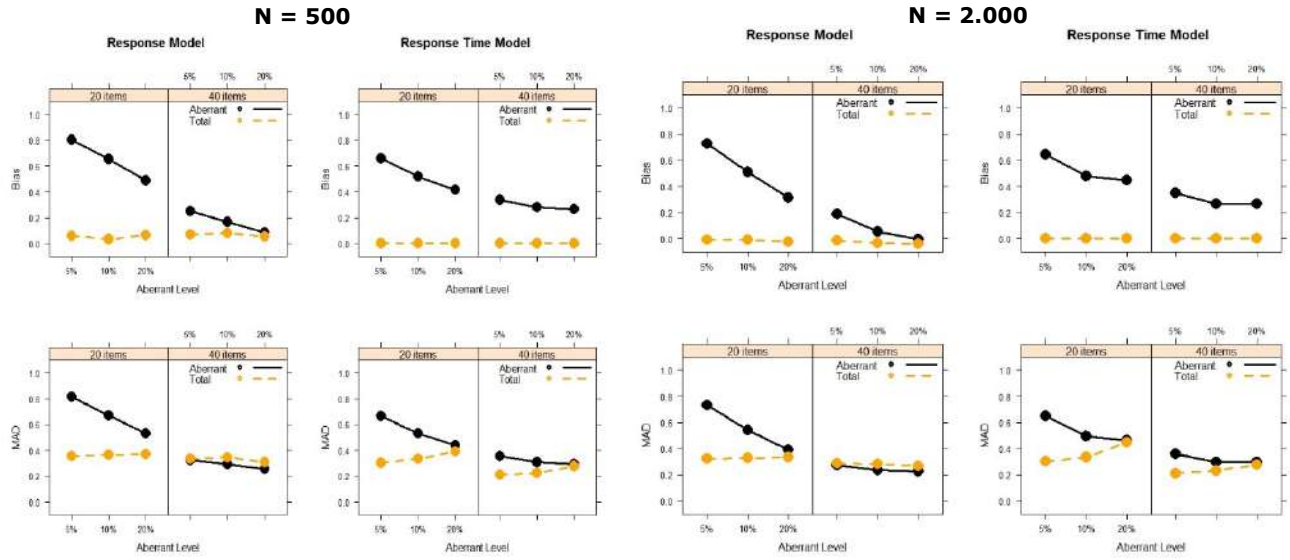
K	AL	N	Response Model				Response Times Model			
			Aberrant		Total		Aberrant		Total	
			Bias	MAD	Bias	MAD	Bias	MAD	Bias	MAD
20	5	2000	0.728	0.735	-0.010	0.325	0.646	0.650	0.000	0.301
	10	2000	0.511	0.540	-0.009	0.331	0.479	0.494	0.000	0.333
	20	2000	0.310	0.393	-0.022	0.332	0.445	0.463	0.000	0.447
40	5	2000	0.187	0.273	-0.013	0.289	0.350	0.362	0.000	0.210
	10	2000	0.053	0.236	-0.036	0.285	0.268	0.297	0.000	0.233
	20	2000	-0.002	0.227	-0.042	0.266	0.266	0.296	0.000	0.277

Note. K = test-length, AL = aberrant level (%), Aberrant = aberrant samples, Total = all samples, N = sample-size, MAD = mean absolute difference. Sources: Personal Data (2022).

Considering Table 8, at 20 items on aberrant samples, the response model has higher value than response time model on shorter test. Yet, on condition where aberrant level reached 20%, response model literally has smaller value than response time model. It can be interpreted as a response model had better ability estimation than response time model on aberrant level 20% when the sample-size is big enough even with shorter test. The result was also consistent for all samples (total), where response time model had better estimation when aberrant level 5% to 10% only. Even though bias showed small value, MAD on aberrant level 20% showed response model smaller value which indicated as better accuracy.

Thereafter, at 40 items as longer test, response model had better estimation which smaller bias and MAD than response time model on aberrant samples. It means that when the test was longer and sample-size was big enough, the response model did not give enlargement estimation than response time model. However, it would be a different scenario when referring on ability estimation of all samples (total) where response time model had smaller value. Even though there was escalation of ability estimation on aberrant persons, still non-aberrant could be estimated approaching true ability parameter which means representing the true ability of examinee. At 20 % of aberrant level, there was slightly difference between response model and response time model where response model had smaller value of MAD. The interaction plot of accuracy is presented in Figure 2 down below for convenient.

Figure 2. Interaction Plot of Accuracy between Response Model and Response Time Model



Sources: Personal Data (2022)

Conclusion

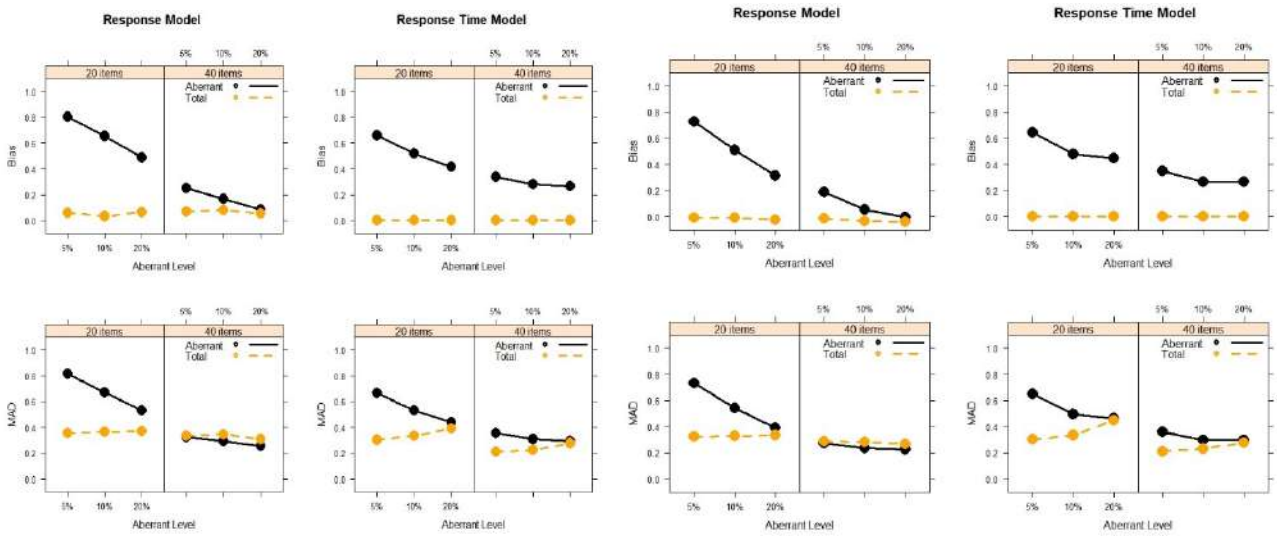
The aim of this study was to determine whether there is a difference of sensitivity detection rates between response model and response time model. Turned out, 9 out of 12 showed that response time model was more sensitive to detect aberrant behavior. It means, counting response time (RT) in response time model, resulting the detection rates became more sensitive than response model. However, the highest difference between both models was only 5.5%. According to the analysis of distance between ability and item difficulty, the smaller the difference ($|\theta - b|$), the more difficult to detect aberrant behavior especially if the value is less than two ($|\theta - b| < 2$) (Smith, 1986; Wright & Stone, 1979, p. 75).

In terms of the item parameter estimation, the finding support Maeda and Zhang (2020) where aberrant will directly affect item parameter estimation. As seen that when aberrant level reached 20%, many items would be estimated easier than they should be. The finding also showed that the ability estimation of both models suffered by the occurrence of aberrant behavior. As a result, the accuracy would be low and the aberrant persons have a spuriously high ability, which means would be questionable if it used as a decision-making in terms of testing (de la Torre & Deng, 2008; Molenaar & Hoijtink, 1990; Reise & Due, 1991).

In general, the finding of this study only as an illustration how the sensitivity of both models handling data containing aberrant behavior. Also, it would be unwise if the person-fit statistics were used as a decision-making of someone who is cheating or guesses. Indeed, a person-fit statistics will show us someone has a discrepancy between his/her observed response (or response time) patterns and expected response patterns (or response time) on a certain IRT model. If the person-fit statistics shows us that someone has a significant, it doesn't mean s(he) doing aberrant behavior during a test. Clearly, we need more evidence such as audit of the test security, directly monitoring or via cctv, proctoring, etc. Because, person-fit statistics is only a tool that help us to find aberrant behavior which sometimes can be statistically mistaken and never be sufficient as primary evidence of someone doing cheating or guessing, or whatsoever on a test.

Limitation and Suggestion

Among the limitations on current study are first, the fact that manipulation on this simulation study would never reflect the real situation of aberrant behavior during a real test. Second, the number of items that assumed aberrant behavior would appear were fixed as five items. It means that a short or longer test would have different proportion of aberrant (5 of 20 items or 5 of 40 items). Third, the addition time of aberrant behavior was also fixed for all aberrant persons. In reality, the addition time could be varied depending on a person, which sometimes could be faster or slower than in this simulation. Last, the aberrant level in this study was proportion of sample (% of N), in fact that a real test would be a different situation.



Further studies, especially a simulation study that related to response time model, may increase the variation of test-length such as range from 10 items to 80 items. Also, may varying other aberrant behavior such as plodding, random response, pacing, etc. Moreover, the aberrant level may be constructed by gradation which refers to how many items of each aberrant person has aberrant behavior on items depending on his/her ability. The variation of manipulating response time (RT) also may be randomly generated to make it more realistic. An analysis of ROC curve (Receiver Operating Characteristic) also may be conducted as well to determine which model has a better classification of aberrant. Other suggestions may apply to other IRT models (i.e., Rasch Model, 1PL, 3PL, etc.) that can be adjusted following research assumption. Among other things, many other person-fit statistics, which have a powerful detection rate beside l_z, may be utilized as well. Therefore, the detection rates of person-fit statistics could be considered.

References

Birenbaum, M. (1985). Comparing the Effectiveness of Several IRT Based Appropriateness Measures in Detecting Unusual Response Patterns. *Educational and Psychological Measurement*, *45*(3), 523–534. https://doi.org/10.1177/001316448504500309

Bulut, O. (2015). Applying Item Response Theory Models to Entrance Examination for Graduate Studies: Practical Issues and Insights. *Journal of Measurement and Evaluation in Education and Psychology*, *6*(2), 313–330. https://doi.org/10.21031/epod.17523

Chalmers, R. P. (2012). Mirt: A multidimensional item response theory package for the R environment. *Journal of Statistical Software*, *48*, 1–195. https://doi.org/10.18637/jss.v048.i06

Cizek, G. J., & Wollack, J. A. (2017). *Handbook of Quantitative Methods for Detecting Cheating on Tests*. Routledge.

de la Torre, J., & Deng, W. (2008). Improving person-fit assessment by correcting the ability estimate and its reference distribution. *Journal of Educational Measurement*, *45*(2), 159–177. https://doi.org/10.1111/j.1745-3984.2008.00058.x

Drasgow, F., Levine, M. V., & McLaughlin, M. E. (1987). Practical with Optimal and Indices Appropriateness. *Applied Psychological Measurement*, *11*(1), 59–79.

Drasgow, F., Levine, M. V., & Williams, E. A. (1985). Appropriateness measurement with polychotomous item response models and standardized indices. *British Journal of Mathematical and Statistical Psychology*, *38*(1), 67–86. https://doi.org/10.1111/j.2044-8317.1985.tb00817.x

Feinberg, R. A., & Rubright, J. D. (2016). Conducting Simulation Studies in Psychometrics. *Educational Measurement: Issues and Practice*, *35*(2), 36–49. https://doi.org/10.1111/emip.12111

Fox, J.-P. (2012). *Bayesian Item Response Modeling: Theory and Applications*. Springer.

Fox, J.-P., Klotzke, K., & Entink, R. K. (2021). *Package ' LNIRT .'* 1–18. https://doi.org/10.18637/jss.v020.i07>.License

Fox, J.-P., & Marianti, S. (2017). Person-Fit Statistics for Joint Models for Accuracy and Speed. *Journal of Educational Measurement*, *54*(2), 243–262. https://doi.org/10.1111/jedm.12143

Karabatsos, G. (2003). Comparing the Aberrant Response Detection Performance of Thirty-Six Person-Fit Statistics. *Applied Measurement in Education*, *16*(4), 277–298. https://doi.org/10.1207/S15324818AME1604_2

Lee, Y.-H., & Chen, H. (2011). A review of recent response-time analyses in educational testing. *Psychological Test and Assessment Modeling*, *53*(3), 359–379. http://www.psychologie-aktuell.com/fileadmin/download/ptam/3-2011_20110927/06_Lee.pdf

Liu, T., Sun, Y., Li, Z., & Xin, T. (2019). The Impact of Aberrant Response on Reliability and Validity. *Measurement: Interdisciplinary Research and Perspectives*, *17*(3), 133–142. https://doi.org/10.1080/15366367.2019.1584848

Lord, F. M. (1980). *Application of Item Response Theory to Practical Testing Problems*. Lawrence Erlbaum Associates.

Maeda, H., & Zhang, B. (2020). Bayesian Extension of Biweight and Huber Weight for Robust Ability Estimation. *Journal of Educational Measurement*, *57*(1), 51–70. https://doi.org/10.1111/jedm.12240

Man, K., Harring, J. R., Ouyang, Y., & Thomas, S. L. (2018). Response Time Based Nonparametric Kullback-Leibler Divergence Measure for Detecting Aberrant Test-Taking Behavior. *International Journal of Testing*, *18*(2), 155–177. https://doi.org/10.1080/15305058.2018.1429446

Marianti, S., Fox, J.-P., Avetisyan, M., Veldkamp, B. P., & Tijmstra, J. (2014). Testing for Aberrant Behavior in Response Time Modeling. *Journal of Educational and Behavioral Statistics*, *39*(6), 426–451. https://doi.org/10.3102/1076998614559412

Meijer, R. R. (1996). Person-Fit Research : An Introduction. *Applied Measurement in Education*, *9*(1), 3–8. https://doi.org/10.1207/s15324818ame0901

Meijer, R. R. (2003). Diagnosing Item Score Patterns on a Test Using Item Response Theory-Based Person-Fit Statistics. *Psychological Methods*, *8*(1), 72–87. https://doi.org/10.1037/1082-989X.8.1.72

Meijer, R. R., & Sijtsma, K. (1995). Detection of Aberrant Item Score Patterns: A Review of Recent Developments. *Applied Measurement in Education*, *8*(3), 261–272. https://doi.org/10.1207/s15324818ame0803_5

Meijer, R. R., & Sijtsma, K. (2001). Methodology review: Evaluating person fit. *Applied Psychological Measurement*, *25*(2), 107–135. https://doi.org/10.1177/01466210122031957

Molenaar, I. W., & Hoijtink, H. (1990). The many null distributions of person fit indices. *Psychometrika*, *55*(1), 75–106. https://doi.org/10.1007/BF02294745

Mousavi, A., & Cui, Y. (2020). The effect of person misfit on item parameter estimation and classification accuracy: A simulation study. *Education Sciences*, *10*(11), 1–15. https://doi.org/10.3390/educsci10110324

Qian, H., Staniewska, D., Reckase, M., & Woo, A. (2016). Using Response Time to Detect Item Preknowledge in Computer-Based Licensure Examinations. *Educational Measurement: Issues and Practice*, *35*(1), 38–47. https://doi.org/10.1111/emip.12102

Reise, S. P. (1990). A Comparison of Item- and Person-Fit Methods of Assessing Model-Data Fit in IRT. *Applied Psychological Measurement*, *14*(2), 127–137. https://doi.org/10.1177/014662169001400202

Reise, S. P., & Due, A. M. (1991). The Influence of Test Characteristics on the Detection of Aberrant Response Patterns. *Applied Psychological Measurement*, *15*(3), 217–226. https://doi.org/10.1177/014662169101500301

Seo, D. G., & Weiss, D. J. (2013). lz Person-Fit Index to Identify Misfit Students With Achievement Test Data. *Educational and Psychological Measurement*, *73*(6), 994–1016. https://doi.org/10.1177/0013164413497015

Smith, R. M. (1986). Person Fit in The Rasch Model. *Educational and Psychological Measurement*, *46*(2), 359–372. http://hjb.sagepub.com.proxy.lib.umich.edu/content/9/2/183.full.pdf+html

Team, R. C. (2021). *R: A language and environment for statistical computing*. https://www.r-project.org/index.html

Tendeiro, J. N. (2021). *Package "PerFit."* 1–56.

Thissen, D. (1983). Timed Testing: An Approach Using Item Response Theory. In *New Horizons in Testing*. Academic Press, Inc. https://doi.org/10.1016/b978-0-12-742780-5.50019-6

van der Linden, W. J., & Guo, F. (2008). Bayesian procedures for identifying aberrant response-time patterns in adaptive testing. *Psychometrika*, *73*(3), 365–384. https://doi.org/10.1007/s11336-007-9046-8

Wang, C., Xu, G., Shang, Z., & Kuncel, N. (2018). Detecting Aberrant Behavior and Item Preknowledge: A Comparison of Mixture Modeling Method and Residual Method. *Journal of Educational and Behavioral Statistics*, *43*(4), 469–501. https://doi.org/10.3102/1076998618767123

Wright, B. D., & Stone, M. H. (1979). *Best Test Design*. Mesa Press. https://doi.org/10.1016/s1567-8032(00)80032-4

A formative evaluation of the program for student representatives in a university discipline panel

Diana Genevive Layag

De La Salle University, Philippines

https://doi.org/10.37517/978-1-74286-697-0-07

The author has a background in developmental and educational psychology, as well as in student affairs. She has conducted research in youth spiritual development, student discipline formation, student conduct, and collaborative online international learning.

Abstract

The university employs a representation of different sectors in the academic community when dealing with major student conduct violations. Among them is the representative of the students. At the time of the study, the student representatives were qualified students who were recruited and trained to become part of the university discipline case conference panel. Upon successful completion of the training, these students were appointed to be members of a program for the pool of student representatives who sit as part of the university adjudicating body, which makes decisions for the university student discipline cases. Terminally, the researcher wanted to improve the engagement and retention of these trained students in the program. Thus, the purpose of conducting this study was to explore the experiences of the students in handling their first case of student conduct violation as part of the university discipline panel. This utilized qualitative methodology in conducting a formative evaluation of the program. Students were asked to submit written reflections on the first student conduct case that they handled. Of the 19 student representatives, eight participated in this study. Line-by-line analysis was done to develop the themes. The findings of this study were four themes that surfaced from the students' experiences: competence, decision-making, vicarious learning, and emotions. This study supports previous research on the importance of supportive adult relationships, which positively impact young people's engagement, as well as the involvement of students in the disciplinary process.

Introduction

The university employs a representation of different sectors in the academic community when dealing with major student conduct violations. Among them is the representative of the students. At the time of the study, the student representatives were qualified students who were recruited and trained to become part of the university discipline case conference panel. Upon successful completion of the training, which is a combination of

conceptual and simulation exercises, these students were appointed to be members of a program for the pool of student representatives who sit as part of the university adjudicating body, which makes decisions for the university student disciplinary cases. Upon appointment, the student representatives become qualified to sit as a member of the university discipline panel and are assigned their first disciplinary case.

Terminally, the researcher wanted to improve the engagement of the student representatives in the program as well as the retention rates of these trained students. These students attended a series of training for the program, and the researcher felt it apt to look into how the program affects the students through the unique experiences the program provides. Thus, the purpose of conducting this study was to explore the experiences of the students in handling their first case of student conduct violation as part of the university discipline panel.

As a program implemented in the university, drawn from the lens of student development theory, this endeavor aimed to enhance the development of student representatives who make decisions in the boardroom, the researcher deemed it wise to gather the individual experiences of students to understand from their perspective how they perceive the experience of being part of a university body that makes decisions on the behavior of their fellow students. King and Howard-Hamilton (2000) also stress the importance of gathering such information to help the university guide the development of students and improve the quality of its programs and services.

In creating a development program for major discipline case student representatives, Stimpson and Stimpson (2008) records that the development of a program in the conduct process may be the responsibility of the one in charge of it but that those with a role in the process should be included (Cordner & Brooks, 1987); espouses different methods to educate them which includes “lectures, experiential activities, observations and mock hearings”; and advocates continuing the training throughout the year. Learning and development for students already involved in the conduct process still have to continue. The actual exposure to the hearings and proceedings requires specially developed critical thinking skills. The inclusion of students in the student disciplinary process has potential benefits for the development of students as well as in the conduct of student discipline on campus. Authors urge universities and colleges to rethink student discipline approaches by integrating the academic and non-academic facets of student development. The good news about this integration is the reinforcement of the institution’s vision and mission, as well as codes of conduct which entail review and clarification and serve as a starting point to its student disciplinary process (Dannells, 1997).

Adults can facilitate learning such as mentoring (Bhatti & Viney, 2010), coaching (Devine et al., 2013), feedback (Allen et al., 2010), and debriefing (Ostovar et al., 2018), and the cognitive and emotional engagements of these trained students rely as much on the presence of encouraging adults around them as on the program itself. With the guidance of supportive adults in the community, students involved in the disciplinary process are found to increase in confidence, civic empowerment, and critical consciousness. Kraus promotes that supportive adult relationships positively impact young people’s engagement. Hence, it is also noteworthy that adults, especially those involved in the conduct process, be involved in the learning and development of student jurors.

Methodology

Research Design

The design of this research is a qualitative formative evaluation. A formative evaluation is conducted during the implementation of a program or its development (Mathison, 2005; Hall et al., 2014). Such evaluation helps develop an understanding of the program participants' interpretations as well as their experiences (Hall et al., 2014). Rectifications may be made while the program is running without waiting for the end point where a summative evaluation may then be done. Thus, presented here is the qualitative formative evaluation of the student representatives in the first cases that they handled.

Participants

Participants for this study were drawn purposively from the19 newly appointed members representing different colleges. Eight (8) newly appointed student representatives for the university discipline panel participated (4 males, 4 females; 42%).

Instrument

The instrument used for the evaluation of the program was the reflective essays on the first case taken as official major discipline case student representatives, the participants in this study.

Data Gathering

After the appointment of the new student members to the university discipline panel, these student representatives began to take their schedule for the case conferences. The time the participants take their first case depends on their availability to handle the case as well as the availability of a scheduled case. Thus, a student discipline representative may take his or her first case at any time within the academic year.

Being finally seated as a student discipline representative in the panel is a culmination of one's preparatory activities for the role and shows the student member what he or she has learned in the training preparations and further development program activities. The students were asked to write a reflective essay about their first case experience.

Data Analysis

Line-by-line analysis (Thomas & Harden, 2008) was used to develop themes for the data. The essays were first numbered per line. The researcher conducted an initial analysis of the essay and marked the responses that describe the experiences, thoughts, and feelings of the participant. These were then transferred verbatim to a table. This table was created to develop the descriptive themes from the general data set. The researcher continued to analyze the encoded responses from the transferred notes. These were labeled from a conceptual perspective. Using the first case as the main source of themes, the other participant essays were analyzed. New labels were generated as new ideas emerged from the data. This iterative process was continually done until the last case. The researcher continued the iterative process by reanalyzing the data in different timeframes.

Findings and Discussion

The researcher's observations on the first case experience of the participants showed a chronological order of presentation of thoughts and recount of experiences: before, during and after one's case conference.

While most of the participants focused their reflections on the skills that they were able to apply in the case conference room as well as the areas where they needed to improve further, some focused their reflections on the way they had presented themselves in the room or the emotions they had felt while taking the role as the student representative of the cases they have handled.

Competence

The need for competence, as found in the results when the development program was crafted, was seen in the findings of this evaluation study. The competence needs of the student representative to the discipline panel as seen in their reflections can be gleaned in the areas in which they want to improve, such as in developing their questions shown in the statements,

> *"…to perform better with asking questions and especially setting the mood and praying…" (Student #1, Lines 6-7),* and
>
> *"Formulating the questions required careful consideration of how the discussion would flow." (Student #2, Lines 4-5)*

Getting a coherent picture of the case scenario was also considered a challenge and has also been a competence driver for the students.

> *"Processing the information required serious attention to detail." (Student #2, Line1)*
>
> *"I was trying to think of questions to ask, but with only that information, it was hard for me to think of what I could ask…" (Student #6, Lines 11-12)*

However, competence in the boardroom is not only limited to the cognitive aspect of the case. Students in their first case also had to consider the perspective of the main person involved – the student-respondent. Empathy, the emotional phenomenon in which "one individual, through observation of another, comes to experience some change in his or her thoughts or feelings" (Davis, 2006), is also something that students realize on the process and need to actively work on, before, during and after the case conference.

> *"Putting myself in the shoes of the evaluated student brought me to the student's situation…" (Student #2, Line 11)*
>
> *"I realized that I need to be more careful in treating different cases with same violations...since different persons undergo different circumstances...and that although they may have the same type of violation, they may not have reasons as to why they did so." (Student #3, Lines 8-11)*
>
> *"To be able to help the respondent, the key is listening and understanding." (Student #5, Line 19)*

Additionally, the student representative in the boardroom needs to be able to process the social environment. Student representatives also become keen on taking the perspective of others in the room, especially the focal person in the student-respondent.

> *"...the respondent…was caught in the situation where he was not able to think properly because of the circumstance…Not all violations committed by students are purposely done or fully intentional." (Student #4, Lines 3-4, 12)*

> *"It is to dig deeper and find insights that aren't written in the report. It is to know the respondent and understand his/her situation more and what led him/her to commit the offense." (Student #5, Lines 11-13)*

As such may be the case for the need for competence in fulfilling the role, some of these student representatives also showed concerns about self-presentations in the panel and in how they performed,

> *"Will I be able to ask the right questions that would lead into a meaningful discussion?" (Student #6, Lines 6-7)*

> *"…how you brand or present yourself to others should be how you would want them to perceive you…" (Student #8, Lines 17-18)*

Thus, the surfacing of competence as a recurring theme for this formative evaluation revealed that student representatives strive to present themselves well in front of their older, professional fellow panel members as well as with their peers who are the responding student charged with the violation. Also, they are particular with the cognitive, social, and emotional aspects of the case that they handled.

Decision-Making

Inevitably, the case conference with the student respondent had to culminate with a resolution. This is the integration of the student representatives' preparations for the case as well as the product of his/her evaluation of the meeting. Making a decision – the case resolution - that is just, and fair is an ultimate goal for one's presence in the panel/board. As members of the adjudicating body, the student representative to the discipline panel speaks of this part as,

> *"…deliberate in an informed manner as I strived to balance emotion and rationality." (Student #2, Lines 12-13)*

> *"This part was really challenging for me because I had to understand where the student was coming from." (Student #6, Line 20-21)*

> *"I also learned about the importance of being objective about the case I'm handling." (Student #3, Line 11-12)*

For the student representative, decision-making in the boardroom needs to consider many things such as what is going on in him/herself – feeling, awareness of biases, and being purposive of one's presence in the university panel.

> *"… to ensure that I will make the best decision in order to help form the student..." (Student #3, Line 14-15)*

"I always had in my mind that I should be able to help the student, get the idea on what is the real reason why she was there, and try to help her be better…I had to place myself into her shoes to really assess the situation and see what she needs to learn. I had to consider her school workloads and at the same time the additional work that we are about to give her due to her violations – on what would be the effect on her." (Student #6, Lines 18-24)

"…our responsibility is to determine the best solution in helping the students become better versions of themselves." (Student #2, Lines 22-23)

It is interesting to consider that students were able to unify the different skill components for a student discipline representative in fulfilling his/her role to make a just and fair case resolution. This statement sums it up well,

"Each person goes through unique experiences, and that's why the resolution for each case varies from person to person." (Student #5, Lines 17-18)

Vicarious Learning

Vicarious learning has been defined as "being able to observe or 'listen in' on experts or peers as they discuss a new topic, learning through the experiences of another…or from a competent other" (Roberts, 2010). The student discipline representatives also experience vicarious learning as the members of the board share words of wisdom with the respondents. Student representatives are able not only to learn from the case the chronology of its events and the technical repercussions of the behavior, based on the student handbook but also the wisdom learned as they reflect on their own personal experiences while in the boardroom.

"During the hearing, I gained a lot of insight and learning points from the professionals in the panel. As they were giving advice to the student, I was able to pick up some life lessons which I can apply to myself." (Student #2, Lines 16-18)

Learning from adults in the panel showed that students are observing the ways of the adults. Just as the senior student representatives have shared about learning from the mentors, the student representatives were keen on learning from the adults in the process.

"During the mock session with (the coordinator) …I realized that my questions were not enough to make the respondent reflect on his actions. The session helped me a lot with formulating better questions and in delivering my ideas." (Student #5, Lines 4-6)

Indeed, students learn a lot from the adults in their environment.

Some of the student representatives focused their self-reflections on the student-respondent of the case. It is striking to note that what stands out to these students is the vicarious learning that they have with the student charged with the case. How the experience of the other person is being processed socially and cognitively by the student representative reflecting on the case shows that the manner in which different student representatives, members of the adjudicating body, varies greatly in how they perceive the student-respondents in the case conference room. For this particular student representative member, there seemed to be an attempt to understand a fellow student,

> "...*the respondent... was caught in the situation where he was not able to think properly because of the circumstance." (Student #4, Lines 3-4)*

However, as if the resolution is later implied,

> *"In the end, they regret this choice once they realize the mistake they have done and the consequences that came after it." (Student #4, Lines 13-15).*

Emotions

It is also noteworthy to discuss the different emotions that student representatives reported in their first case experience. From feeling nervous and uncertain at the start of the case to disbelief in the nature of the case.

> *"I was quite nervous… …I wasn't sure if I knew the flow of the case conference well enough." (Student #5, Lines 1, 8)*

> *"I was pretty nervous the whole week thinking about what will actually happen during the actual case…. I personally have never imagined someone who could…" (Student #6, Lines 3-4, 9)*

Being overwhelmed with the experience, feeling pressured and a sense of a difference in reality from what one has prepared oneself for also describes the feelings of these first timers. Indeed, one learns from actual experience.

> *"...no amount of training and visualizations can compare to the real thing…since some things might not go the way that I was expecting them to be…" (Student #3, Lines 3-5)*

> *"…it was an overwhelming experience…" (Student #2, Line 9)*

> *"There is a lot of pressure as a student representative. We are always the first to speak up...to ask questions and to give opinions." (Student #5, Lines 15-16)*

There were also moments of self-doubts as they take on their first cases. Issues arose as they personally critiqued the performance of the skills that they have been trained in or questioned the way they presented themselves as competent student representatives in the adjudicating body.

> *"…my questions weren't as good in extracting info and reflecting purposes…" (Student #1, Lines 3-4)*

> *"'Will I be confident enough to speak up my real concerns and questions? 'Will the other student be accommodating enough to answer my questions?' 'Will I be able to ask the right questions that would lead into a meaningful discussion?'" (Student #6, Lines 4-7)*

Regardless of how one felt at the start, or even during the disciplinary proceeding, the case conferences of the students appeared to have ended well as students felt optimistic and grateful after their first cases.

> *"I'd like to take more (repeated violations) cases, so I'd have a specialty case and have practice in the same game." (Student #1, Lines 7-8)*

"I do hope that I improve more and become better in my role as a (student representative). I'm truly grateful to have been given this opportunity…I hope what I do can really bring help to others." (Student #5, Lines 20-22)

"It was a really grateful opportunity, having to be part of a life of a person, knowing that at some point, you were able to help them become better. As a student, I myself is also thankful for the school for having this initiative of giving us an opportunity to help our fellow students… knowing that the satisfaction that I get from helping them will be priceless…" (Student #6, Lines 28-30, 33-34)

Researcher Reflections

It is interesting to note that the chronological reflections stated by the students as they handled their first disciplinary cases have been of different colors and hues. While many have shared competence, as the themes have shown, the differences in these perceived competencies seem to lie in a spectrum from self-doubt to an increased level of confidence in trying to feel the social and emotional atmosphere inside the boardroom, as well as in personally assessing the reasonings of the student charged with the violation.

While inside the boardroom, the students' observations around them, particularly the adults who work with them in handling the case are very important as they glean important professional and life lessons from them as well. As they share their experiences and perspectives as students themselves, they are also inspired by the people they work with and feel the importance of personally applying the lessons that others have experienced. In terms of decision-making, students have learned to take other perspectives, use critical thinking abilities, show awareness of their biases, and strove for impartiality. Similar to the hues and colors of their competence needs, the emotions they felt before, during, and after the case hearings were as varied. And just as they feel, I am optimistic that despite the roller coaster ride of these feelings, as they mature in their roles, this program will form them to emerge with the graduate attributes our university envisions them to be.

Conclusion

The findings of this study were four themes that surfaced from the students' experiences: competence, decision-making, vicarious learning, and emotions. This study supports previous research on the importance of supportive adult relationships, which positively impact young people's engagement, as well as the involvement of students in the disciplinary process.

It's amazing how much one can glean from the first case experiences of the student discipline case representatives. This study reiterates in its finding what Stoner II (2000) assents in involving students in the disciplinary process because they can understand their peers better. This also not paves an understanding of others, as accorded in the development of an empathic stance towards the situation but also in the vicarious learning that students experience – both from their peers and adults. The findings have also shown the importance of mentoring, coaching, and feedback on the performance of students. The need to perform competently in different facets of the case conference has been evident in the results as student representatives engage cognitively, emotionally,

and socially in the process to derive at a fair and just resolution as they make decisions about the case. This study also promotes Kraus findings on the importance of supportive adult relationships, which positively impact young people's engagement. Further program enhancement is recommended through continual formative evaluation using qualitative and quantitative methodologies.

It is important to realize the role that adults play in the development of our students, not only in skills but also in the many unspoken things the adults around them do that they observe and catch. Truly, student programs ought to be more intent on the accompaniment that the university accords to these young people. After all, this is what we, educators, do best.

References

Allen, T. D., Shockley, K. M., & Poteat, L. (2010). Prote´ge´ anxiety attachment and feedback in mentoring relationships. *Journal of Vocational Behaviour* 77: 73–80.

Bhatti, N., & Viney, R. (2010). Coaching and mentoring. *BMJ* 2010; 341 (Published 14 July 2010) doi: https://doi.org/10.1136/bmj.c3427

Cordner, P., & Brooks, T. F. (1987). Training Techniques for Judicial Systems. (Volume 1987, Issue 39 p. 31-42) https://doi.org/10.1002/ss.37119873905

Dannells, M. (1997). ASHE-ERIC Higher Education Report Volume 25-2, From Discipline to Development: Rethinking Student Conduct in Higher Education.

Davis, M. H. (2006). Empathy. Handbook of the Sociology of Emotions. ISBN: 978-0-387-30713-8. https://link.springer.com/chapter/10.1007/978-0-387-30715-2_20

DeFosset, A. R., Schooley, T. S., Abrams, L. S., Kuo, T., & Gase, L. N. (2017). Describing theoretical underpinnings in juvenile justice diversion: A case study explicating Teen Court program theory to guide research and practice. *Children and Youth Services Review.* Volume 73, February, Pages 419-429 www.elsevier.com/locate/childyouth.

Devine, M., Meyers, R., & Houssemand, C. (2013). 3rd World Conference on Learning, Teaching and Educational Leadership (WCLTA-2012). How can coaching make a positive impact within educational settings? Procedia - Social and Behavioral Sciences 93 1382 – 1389. www.sciencedirect.com

Gehring, D. D. (2001). The objectives of student discipline and the process that's due: Are they compatible? *NASPA Journal,* 38:4, 466-481.

Hall, J., Freeman, M., & Roulston, K. (2014). Right timing in formative program evaluation. Evaluation and Program Planning, 45, 151–156. https://doi.org/10.1016/j.evalprogplan.2014.04.007

King, P., & Howard-Hamilton, M. F. (2000). Using Student Development Theory to Inform Institutional Research. New Directions for Institutional Research, 2000(108), 19–36. https://doi.org/10.1002/ir.10802

Mathison, S. (Ed.). (2005). Encyclopedia of evaluation. Thousand Oaks, CA: Sage Publications.

Ostovar, S., Allahbakhshian, A., Gholizadeh, L., Dizaji, S. L., Sarbakhsh, P., & Ghahramanian, A. (2018). Comparison of the effects of debriefing methods on psychomotor skills, self-confidence, and satisfaction in novice nursing students:

A quasi-experimental study. *Journal of Advanced Pharmaceutical Technology & Research* | Published by Wolters Kluwer – Medknow

Roberts, D. (2010). Vicarious learning: A review of the literature. *Nurse Education in Practice* 10 13–16. https://www.sciencedirect.com/science/article/pii/S1471595309000183

Stimpson, M. T., & Stimpson, R. L. (2008). Twenty-seven years of student conduct literature: Implications for practice, scholarship, and ASJA. *Journal of Student Conduct Administration,* Vol. 1, no. 1.

Stoner, E. N. (2000). Reviewing your Student Discipline Policy: A Project worth the investment. United Educators Insurance Risk Retention Group, Inc., Chevy Chase, MD.; National Association of Student Personnel Administrators, Inc. Produced with Reed Smith Shaw & McClay.

Thomas, J., & Harden, A. (2008). Methods for the thematic synthesis of qualitative research in systematic reviews. *BMC Medical Research Methodology* 2008, 8:45 doi:10.1186/1471-2288-8-45

Validation of the Child and Adult Social Support Scale (CASSS) which measures social support in Indonesian version

Endah Mastuti
Fajrianthi
Fitri Andriani

Faculty of Psychology, Airlangga University

https://doi.org/10.37517/978-1-74286-697-0-08

Endah Mastuti is a lecturer at the Faculty of Psychology, Airlangga University.

Abstract

Social support has an important role, so that students in online learning can reduce the various problems they face. Social support here comes from parents, teachers, classmates, close friends, and the school itself as an institution. Based on this, it is necessary to do research related to social support to get an overview of the support from whom students need in online learning. Furthermore, this can be used as input to intervene in the problems faced by students. To conduct this research, it is necessary to have a measuring tool to conduct social support research. One of the comprehensive measuring tools in measuring social support is the Child and Adult Social Support Scale (CASSS) which has been made by (Malecki & Demaray, 2002).

This study aims to: (1) adapt the Indonesian version of the CASSS measuring instrument, (2) test the psychometric properties of the measuring instrument that has been adapted in Indonesian. The results show that the CASSS measuring instrument that has been adapted shows that the loading factor value of all items is above 0.7. Based on this, it can be said that all items (60 items) in the CASSS scale which are the observed variables have good validity on the latent variables on each dimension of the Social support construct. Meanwhile, each dimension also has a good factor load on the Social Support construct. This shows that all CASSS items have good validity values so that they can be used for further research. The results of the reliability calculation show that all subscales of the CASSS Scale have a reliability value above 0.9 using the alpha technique.

Introduction

A lot of research on student issues related to social support. Various problems have occurred among Indonesian students including the emergence of academic stress, decreased achievement, decreased motivation to learn, fatigue, etc. (Agustina & Wisnumurti, 2019), Wibowo and Susanto (2014)

There are several factors that can reduce various things that can help students to reduce these conditions, including social support. Based on this, it is necessary to carry out various studies related to social support so as to get an overview of the support from whom students need in learning. To conduct this research, it is necessary to have a measuring tool to conduct social support research. One of the comprehensive measuring tools in measuring social support is the Child and Adult Social Support Scale (CASSS) which has been made by Malecki et al. (2019)

The definition of social support used in the development of CASSS (2000) is closely related to Tardy's (1985) model and is widely interpreted. Tardy's (1985) model includes five conceptual problems: direction, disposition, description/evaluation, content, and network. Referral refers to the notion that social support is given and received. Disposition is the availability of support in terms of what people have access to, versus the enactment of support or the actual utilization of support resources. Social support is usually described and/or evaluated. Description of social support involves investigating what types of social support are received and from whom they receive it, while evaluation of social support is a measurement of an individual's satisfaction with the support he or she receives. The content of social support can be separated into four types, which vary depending on the situation: emotional (i.e., trust, love, and empathy), instrumental (i.e., helpful behavior), informational (i.e., providing advice), and appraisal support (i.e., offering evaluative feedback). The fifth conceptual problem identified by Tardy's (1985) is social support networks. The network is made up of various people who provide support. Individual networks can include family, friends, neighbors, coworkers, and community professionals. Based on this model, we view social support as individuals' perceptions of general support or specific support behaviors (available or reacted) from people in their social networks, which enhance their functioning and/or may protect them from adverse outcomes. Sources of support (social networks) as measured by CASSS (2002) include parents, teachers, classmates, close friends, and school. In addition, CASSS (2002) utilizes four types of supportive behavior from each source including emotional, instrumental, informational, and appraisal support.

The urgency of the research, the large number of needs to measure social support both in the context of education and development, so that a comprehensive standard measuring tool is needed in identifying social needs. This measuring tool can not only be used for research purposes but can also be used as a psychological assessment tool, regarding the perceived social support of children and adolescents and how they consider it important for children and adolescents.

Methods

The subjects in this study were 250 high school students in East Java aged 15-19 years. This study uses a survey technique by distributing questionnaires as a data collection

instrument. with The Child and Adult Social Support Scale (CASSS). Administration is done by providing a link to fill out the questionnaire via an electronic device.

Table 1. Blueprint of CASSS Scale

	Emotional Support	Instrumental Support	Informational Support	Appraisal Support
Parent	1.2.3	10.11.12	4. 5. 6	7. 8. 9
Teacher	1.2.3	10.11.12	4. 5. 6	7. 8. 9
Classmate	1.2.3	10.11.12	4. 5. 6	7. 8. 9
Close Friend	1.2.3	10.11.12	4. 5. 6	7. 8. 9
School	1.2.3	10.11.12	4. 5. 6	7. 8. 9

Data Analysis

This study begins with an adaptation of the scale used for research. The scale adapted is the Child Adolescence Social Support Scale CASSS from (Malecki & Demaray, 2002). The process of adapting measuring instruments is carried out in accordance with the applicable procedures according to the ITC [2005]:

1. Preparation: At this stage, according to suggestions from the ITC, researchers are expected to license the measuring instruments used.
2. Translation Process
   a. *Forward Translation* with two translator
   b. Sintesis result of *Forward Translation*
   c. *Backward Translation* with two translator
   d. Sintesis result of *Backward Translation*
3. Reviewing Process
4. Tryout Item
5. Property psychometric analysis
   a. Validity evaluation
   b. Reliability evaluation

Furthermore, the data from the test items were analyzed using CFA (Confirmatory Factor Analysis). CFA is a type of structural equation model (SEM) that examines the hypothesized relationship between the indicator (item) and the latent variable that the indicator wants to measure. CFA analysis was carried out to determine whether the scale used in this study really represented the underlying theory or not [6].

Table 2. Data Analysis

Analysis		Value Received
Validity	*Content Validity Index* (CVI)	CVI > 0,8
	Confirmatory Factor Analysis (CFA)	*Loading factor* > 0,5
	p-value	> 0,05 declared *fit*
	χ2/df	Between 2 to 5
	Standardized Root Mean Square (SRMR)	Close 0 more *fit*, 0,06-0,08 as *marginal fit*
	Root Mean Square Error of Approximation (RMSEA)	Close 0 more *fit*, 0,06-0,08 as *marginal fit*
	Comparative Fit Index (CFI)	> 0,8 as *marginal fit*

Reliability	*Cronbach's alpha*	> 0,6 as good
	Construct Reliability (CR)	> 0,7 as good

Result and Discussion

Validity

Evidence Based on Test Content

In this process, researcher ask five experts to review and rate :

1. The expert reviewer evaluates the two components, namely, aspects of similarity (level of similarity in meaning, even though the terms are different) and comparability (similarity of language, phrases, terms, words, and sentences) in the back-translation process with the original scale. In this process, expert reviewers are asked to rate them on a rating scale (range 1–7). The purpose of this process is to produce items that have the same language and meaning.
2. The researcher asked the expert to conduct an assessment of the aspects of the test content that needed to be evaluated. The aspects are: (1) Relevance or relevance. namely whether the content of the test has conformity with the specific content domain to be measured, (2) Clarity or clarity. that is, whether the content of the test clearly reflects the specific content domain to be measured. is quite clear and easy to understand (3) Importance or the level of importance is how important the item is when it is associated with the research construct and context (Goodwin & Leech, 2003). The results mean score of Similarity is 1,13 and comparability is 1,01. According to Sperber (2004), equivalent items are items that have a mean score < 4.The results of the I-CVI calculation on the scale show that the I-CVI is at a value of 1.00. This means that all items on the scale adapted in this study have a good I-CVI and all items can be used in the study. If the I-CVI value is higher than 0.79. then the item is good and usable. (Zamanzadeh et al., 2015). The evidence above shows that this scale has evidence based on test content.

Evidence based on Internal Structure

The results of the Confirmatory Factor Analysis test show that the results of the Goodness of Fit test for the CASSS scale are shown in table 4. Based on the six criteria, good of fit indicates good fit, namely RMSEA, GFI, CFI, and NFI, while the AGFI and p-values indicate poor fit.

Table 3. *Goodness of Fit* for CASSS scale

Kriterion *Goodness of fit*	Acceptable Level	Result	Information
p-value	≥ 0.05	0.000	*Poor fit*
RMSEA	≤ 0.08	0.048	*Good fit*
GFI	≥ 0.90	0.99	*Good fit*
AGFI	≥ 0.90	0.67	*Poor fit*
CFI	≥ 0.90	0.99	*Good fit*
NFI	≥ 0.90	0.97	*Good fit*

Meanwhile, the results of the second order CFA loading factor and t-value can be seen in Table 5. The results show that the loading factor value is more than 0.3 and the t-value is more than 1.96. The loading factor value of all items is above 0.7. Based on this, it can be concluded that all items (60 items) in the CASSS scale which are observed variables have good validity on the latent variables on each dimension of the Social Support construct. Meanwhile, each dimension also has a good factor load on the Social Support construct. This shows that all CASSS items have good validity values so that they can be used for further research.

Table 4. Loading factor & T-Value CASSS Scale

Item	Loading Factor	T-Value
1	0.65	
2	0.72	11.90
3	0.72	13.55
4	0.82	13.86
5	0.83	15.15
6	0.85	14.65
7	0.79	14.95
8	0.79	14.54
9	0.67	12.76
10	0.76	12.84
11	0.78	12.29
12	0.68	12.12
13	0.77	
14	0.78	18.23
15	0.81	18.39
16	0.80	17.27
17	0.83	20.03
18	0.85	19.57
19	0.84	18.70
20	0.85	19.55
21	0.80	17.01
22	0.83	19.04
23	0.79	17.51
24	0.76	15.01
25	0.75	
26	0.74	17.77
27	0.81	19.20
28	0.88	19.24
29	0.89	18.89
30	0.89	19.50
31	0.86	18.55
32	0.87	19.06
33	0.83	17.80
34	0.85	18.28
35	0.85	18.31
36	0.80	17.90

Item	Loading Factor	T-Value
37	0.86	
38	0.85	33.34
39	0.85	33.10
40	0.87	28.62
41	0.90	30.90
42	0.89	44.11
43	0.82	25.34
44	0.90	41.67
45	0.88	41.49
46	0.87	23.51
47	0.77	19.96
48	0.88	38.11
49	0.85	
50	0.81	29.54
51	0.86	32.05
52	0.86	34.43
53	0.86	31.02
54	0.87	33.02
55	0.87	32.34
56	0.89	33.42
57	0.89	35.60
58	0.88	35.62
59	0.84	16.42
60	0.84	28.96

Reliability

The researchers calculated reliability using the alpha technique and the reliability of each construct was measured by Construct Reliability (CR). The value of construct reliability that is set is greater than 0.7. so, if the results of the contract reliability test show a number above 0.7 then it is reliable. The researcher also looks for the Variance Extracted and Reliability scores. The reliability of the construct to test the reliability can use the value of AVE (Variance Extracted). The AVE value is calculated as the total square of the standardized factor loading divided by the total square of the standardized loading plus the total variance of the error. An AVE value equal to or above 0.5 indicates a good convergence (Hair et al., 2014).

Table 5. Reliability CASSS

Construct	Alpha	CR	AVE
Social Support			
Parent	0.943	0,941	0.574
Teacher	0.962	0.968	0.655
Classmate	0.968	0.961	0.675
Close Friend	0.976	0.968	0.722
School	0.976	0.969	0.726

Conclusion

Based on the discussion, it is concluded that the CASSS with 5 subscales: Social support form parent, teacher's Social support, Classmate's Social support, Close friend's Social support and School 's Social support were proven to meet the criteria of goodness of fit, validity and construct reliability. Furthermore, the scale had fulfilled reliability criterion comprehensively. Therefore, the developed scale was feasible to be used in collecting data to measure the student's social support in Indonesia. Based on the findings, it is concluded that the CASSS in Indonesian Version have good psychometric properties.

References

Agustina, L., & Wisnumurti, A. (2019). Dukungan Sosial Dan Motivasi Belajar Siswa Sma Masehi 2 Psak Semarang. *Personifikasi*, Vol.10 No.1 Mei 2019.

Clark, Kelly N., and Christine K. Malecki. 2019. "Academic Grit Scale: Psychometric Properties and Associations with Achievement and Life Satisfaction." Journal of School Psychology 72: 49–66. doi:10.1016/j.jsp.2018.12.001.

Goodwin, L. D., & Leech, N. L. (2003). "The Meaning of Validity in the Bew Standards for Educationl and Psychological Testing: Implications for Measurement Courses," Meas. Eval. Couns. Dev., vol. 36, pp. 181–191.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2014). *Multivariate Data Analysis* (Seventh Ed). Pearson Education, Inc.

International Test Commission. (2005). "ITC Guidelines for translating and adaptating tests," pp. 1–9.

Jeanrie, C., & Bertrand, R. (1999). "Translating Tests with the International Test Commission ' s Guidelines : Keeping Validity in Mind," *Eur. J. Psychol. Assess.*, vol. 15, no. 3, pp. 277–283, doi: 10.1027//1015-5759.15.3.277.

Malecki, Christine & Demaray, Michelle. (2002). Measuring perceived social support: Development of the child and adolescent social support scale (CASSS). Psychology in the Schools. 39. 1 - 18. 10.1002/pits.10004.

Rios, J., & Wells, C. (2014). "Validity evidence based on internal structure," *Psicothema*, vol. 26, no. 1, pp. 108–116, doi: 10.7334/psicothema2013.260.

Sperber, A. D. (2004). "Translation and Validation of Study Instruments for Cross-Cultural Research," *Am. Gastroenterol. Assoc.*, vol. 126, pp. 124–128, doi: 10.1053/j.gastro.2003.10.016.

Wibowo, M., & Susanto, S. (2014). Dinamika Dukungan Sosial pada Prestasi Siswa Sekolah Dasar Berbasis Pendekatan *Indigenous Psychology Jurnal Psikologi Tabularasa.* Volume 9, No.1, April 2014: 30-36.

Zamanzadeh, V., Ghahramanian, A., Rassouli, M., Abbaszadeh, A., Alavi-Majd, H., & Nikanfar, A. R. (2015). Design and Implementation Content Validity Study: Development of an instrument for measuring Patient-Centered Communication. *Journal of Caring Sciences*, *4*(2), 165–178. https://doi.org/10.15171/jcs.2015.017./jcs.2015.017.

Development and psychometric assessment of pilot's wellbeing scale: A study on Indonesian commercial pilot

Patricia Yora Wenas
Angelina Oktavia Suryani
Laura Fransisca Sudarnoto

Atma Jaya Catholic University of Indonesia

https://doi.org/10.37517/978-1-74286-697-0-09

Patricia Yora Wenas is a professional commercial pilot with seven years of airline experience. She is also a pilot instructor and a recruiter for commercial airlines. Her academic background is in law and advocacy. She is currently a graduate student at the Atma Jaya Catholic University of Indonesia, majoring in psychology.

Abstract

Flight success and accidents rely on the pilot's condition. To date, there is no psychometrically tested instrument to measure the psychological well-being of pilots. This study aims to develop items for measuring the well-being of commercial pilots in Indonesia. The constructs, dimensions, and indicators were formulated through literature reviews and interviews with pilots about their well-being. We develop well-being as a state in which a person feels positive, healthy, prosperous, comfortable, and valued in their work, able to control and contribute to the social environment. The instrument conveyed six dimensions, namely (1) positive emotion, (2) health, (3) competence, (4) recognition, (5) work security, and (6) social relationship. We implemented summated rating scale model (Likert's scale) to write 80 items on a 6-point scale. The scoring was from 1 (never/strongly not agree) to 6 (almost always/strongly agree); the scores for the unfavourable items were reversed. We applied CVI and SCVI methods to examine the instrument's content validity, corrected-item total correlation to test the item homogeneity, and Cronbach's alpha formula to assess the scale reliability. In content validity, we included two psychometrics experts and one well-being expert to evaluate the items' quality via qualitative and quantitative methods. In the qualitative approach, the expert appraised the item's sentence: whether it derived from the indicators, dimensions, and construct, and whether the sentences were comprehended. In quantitative form, the experts assigned scores from 1 to 4, indicating the item's relevance with the construct, clarity, simplicity, and freedom of ambiguity. We recruited 117 commercial pilots via convenience sampling. The content validity testing suggested that 31 items should be revised. The quantitative test showed satisfying results with a CVI average for the relevance aspect of 0.92 and an SCVI/UA of 0.93. Based on the content validity assessment, we decided to retain 62 items. Next, the corrected item-

total correlation revealed that 61 items were homogenously measuring the construct. Finally, the reliability testing indicated satisfactory reliability with an alpha of 0.95.

Introduction

Background

Humans are one of the main factors causing plane crashes, apart from technical and environmental factors. According to aviation accident investigation data in 2010-2016, humans are responsible for 67.12% of accidents (NTSC, 2016). As an aircraft operator, the pilot is fully responsible for the condition of the aircraft, ensures the safety of passengers and cabin crew during the flight, is administratively responsible for everything that happens during the flight, and decides on rescue operations (Nrangwesti, 2011). According to Syabilah et al., (2015), the burden and work demands for the pilot are very high, especially for commercial airplane pilots, one of the professions with the highest stress levels in the world. According to Wilson et al. (2022), pilots have better health status than the general group because of work demands that encourage pilots to always be in top shape (fit). However, being a pilot is a profession with high risk. Their work conditions include irregular work patterns, fatigue, exposure to cosmic radiation, irregular mealtimes, and the mental stress demands associated with flight safety, noise, vibration, and cabin air quality. According to the Indonesian Aviation Health Center, the results of examinations on 20,100 cross-class pilots in September 2019 showed that 2,023 pilots should have to re-check their health because they were in an unacceptable condition. (Antara, 2019). However, the examination results were limited to a physical examination and did not include a psychological assessment. Even though the pilot's wellness guide states that well-being comprises physical, psychological, and social health, the comprehensive measurement will enable pilots to manage stress well to increase their work performance. (Flight Safety Foundation, 2020).

The German Wing 9525 plane crash is one of the events caused by the human factor. This accident happened because the pilot attempted suicide during his flying assignment due to a mental health disorder that the aviation health agency did not identify (Branch, 2021). In Indonesia, plane crashes have been repeatedly found due to human factors, for example, the crash of the Lion Air JT 904 in the Bali Sea. In that incident, the pilot claimed to have experienced hallucinations during the landing (Torrence, 2018). Another example is the crash of Air Asia QZ 8501 in the Java Sea due to miscommunication between the two pilots on the flight (Bellamy, 2015). The most recent incident was the crash of the Boeing 737-800NG aircraft owned by China Airlines in March 2022. The analysis of the Flight Data Recorder and Cockpit Voice Recorder revealed an allegation that the pilot deliberately crashed the plane. (BBC News, 2022). We assume that one reason for these incidents is the lack of regular mental health testing/evaluation conducted by aviation health agencies.

Since the German Wing 9525 incident, the world's aviation health agencies have a well-being guide to guide pilots in managing their physical, mental, and social well-being is a state of a person who is happy because he is healthy, has good education and wages, is optimistic, free from worry, is religious, confident, and has good insight. By paying attention to pilot well-being, international aviation agencies plan to improve the

performance of pilots and maintain security and safety in air transportation. Still, even though the guideline has mentioned it, there is no specific training or follow-up and evaluation of the success of the well-being guide.

Aim of the study

Considering the need for a well-being assessment, we developed an instrument measuring the pilot's well-being. This study aimed to test the psychometric elements of the instrument items comprising the item's validity and test reliability. The measurement tool is prominent because it includes the work and life context welfare aligned with the Flight Safety Foundation standard (2020).

Well-being

We specified the dimension of pilot well-being based on an analysis of several well-being perspectives and theories (Huppert & So, 2013; Page, 2005; Seligman, 2011). We define well-being as a state in which a person feels positive, healthy, prosperous, comfortable, and valued in their work, can control, and contribute to his social environment and can develop himself.

Dimensions

1. *Positive emotion*: positive emotions involve comfortable and enjoyable life, hope, happiness, and life satisfaction. Positive emotions are exemplified by feeling happy and hopeful. Positive emotions can consciously direct a person's feelings for the better, increase resilience, and reduce stress.
2. *Health*: healthy physical, mental, and spiritual state. It entails practicing a nutritious diet, moving regularly, having quality sleep, and maintaining cleanliness. Individuals with high well-being will maximize their energy to keep moving and choose healthy behaviors even though they feel busy.
3. *Social Relationship*: making the best use of time to form a balance activity between work and personal life (work-life balance). Individuals have social networks that support each other.
4. *Receive recognition*: acknowledgment for good performance. Here, the employee has positive feelings toward work and life because they receive appreciation and acknowledgment from the organization and others for their excellent performance. They feel that they are valued and accepted by their families and superiors. They gain the appropriate amount of wages, benefits, and rewards.
5. *Competence*: Individuals feel that they have good abilities to carry out their duties. A belief that they can think clearly, concentrate, and make good decisions.
6. *Job security*: Security is an employee's satisfaction with the work position and having a sense of security in the work position.

Method

Item development phases

The development of the instrument follows the stages of constructing a psychological measurement tool described by Crocker and Algina (2008):

1. Collect behavioral characteristics of well-being from literature studies and interviews with pilots.
2. Formulate the instrument specifications (item statements and responses)
3. Writing the items.

4. Content validity testing.
5. Revise the item based on content validity testing.
6. Collect data for construct validity and reliability testing.
7. Test the validity and reliability.

Statistical method for construct validity and reliability testing

In this study, we implemented content validity, construct validity, and reliability. The content validity testing involved several experts and participants. The evaluation by experts involves quantitative (content validity index) and qualitative assessments. In quantitative assessment, the experts reviewed by using a 4-point rating scale whether the statements were relevant to the indicator, domains, and construct. Whether they were clear sentences, simple and easily understood, and unambiguous. Scores 1 and 2 indicate inappropriate quality, while scores 3 and 4 describe a good-quality item. We calculated the index at the item (I-CVI average) and scale level (S-CVI/UA). S-CVI/UA (universal agreement) is the percentage of the total number of items that get a score of 3 or 4 from all experts (Grant & Davis, 1997).

Meanwhile, S-CVI/AVE is the average value of I-CVI for all items in the measuring instrument (Polit & Beck, 2006). The participants can evaluate whether the items are relevant to them, readable - easily comprehended, and unambiguous. They can tell the construct of the scale with their understanding and words. Their explanation may help us identify the scale's construct from the participants' perspective.

We applied the internal criteria analysis in construct validity testing by employing corrected item-total correlation. The total score represented the scale's construct because it is a cumulative score from each item presenting the construct's behavioural indicators. It is said "corrected" because we subtract the total score from the item score that is being analysed. This technique can make the two variables (item and total scores) independent. We employed the Pearson correlation to test the items. The decision to retain or reject the items used r-critical values that depend on the number of participants (Gravetter et al., 2018).

The reliability testing included Cronbach's Alpha method. We employed the standard of 0.70 to affirm the scale's reliability (Anastasi & Urbina, 1997). We need to test whether the items measure the construct consistently. In the data analysis, we utilized the JASP ver. 0.14.3.

Participants

We recruited 117 pilots from national commercial airlines via the convenience sampling method. All the participants have an active flight certification. Most participants are males (N=79, 69.9%) from 22 to 63 years old. They have been working as a pilot for six months to 39 years (M = 9.38, SD = 8.03). We distributed the scale to participants online.

Item specification

We applied a summated rating scale method for item writing by using a 6-point rating scale that participants should choose one response option from "Very disagree /never" (score 1) to "Very agree/almost always" (score 6).

Result

Items

We draft 80 items deriving from six dimensions. Table 1 describes examples of the scale.

Table 1. Sample of items from each dimension

Dimension	No.	Items
Positive emotion (P)	1.	I am happy with my life as a pilot (disagree-agree)
	2.	I am proud as a pilot (disagree-agree)
Health (H)	3.	I have trouble sleeping with a regular pattern (never-almost always)
	4.	My current physical condition is fit as a pilot (disagree-agree)
Competent (C)	5.	I can control an emergency during my duty (never-almost always).
Receive recognition (R)	6.	In decision-making, opinions of all levels of positions in my airline are listened to (disagree-agree)
Job security (K)	7.	I feel my future is guaranteed with my profession as a pilot disagree-agree)
Social relationship (S)	8.	I have many friends who give positive support to me (disagree-agree)

Content validity testing

We incorporate three experts to evaluate the items. Two of the experts were psychological measurement specialists, and one expert was a well-being researcher. We also involved one experienced pilot as a participant in readability and face validity testing. Table 2 describes the result of the quantitative analysis of content validity testing.

Table 2. Content Validity Index of Pilot's Well-being Scale

Aspects	I-CVI average	S-CVI/UI
Relevance	0.92	0.93
Sentence clarity	0.89	0.80
Sentence simplicity	0.87	0.75
Sentence unambiguity	0.89	0.81

Data in Table 2 showed that the relevance aspect of the items was above 0.90 (recommended ≥ 0.90 (Waltz et al., 2005), which means the items were valid, representing the indicator, domain, and construct. However, the assessment of the other three aspects did not show satisfying results. We dropped 17 items and revised the others. The revision was based on the expert's recommendation to enhance the expression of the statements.

Construct validity testing

In construct validity testing, we run the corrected item-total correlation to 63 items by correlating the items' scores to the domain and construct scores. We used the critical value of 0.156 (df = 115, alpha = 0.05) to retain or reject the items. Most coefficient correlations were above the required value, but one item, "I am reluctant to be known as a pilot," did not succeed ($r_{\text{item-domain}} = 0.139$, $r_{\text{item-total}} = 0.097$). Therefore, we retain

62 items (Table 3). Subsequently, we chose 20 items with a high coefficient as the final scale (Table 4).

Table 3. The corrected item-total correlation coefficients of 62 items

No.	Items in Bahasa Indonesia	Items in English	*r*
P1.1	Saya merasa cemas dengan kondisi penerbangan saat ini	I feel anxious about the current flight conditions	0.319
P1.2	Saya bahagia dengan hidup saya sebagai penerbang	I am happy with my life as a pilot	0.596
P1.5	Aktivitas sebagai penerbang sangat cocok untuk saya	The activity as a pilot is perfect for me	0.511
P2.1	Saya merasa nyaman dengan bekerja sebagai penerbang.	I feel comfortable working as a pilot.	0.523
P2.2	Saya merasa tenang pada saat tugas terbang	I feel calm on flying assignments	0.690
P2.3	Bekerja di pesawat membuat saya merasa gelisah	Working on an airplane makes me feel uneasy	0.437
P2.5	Rotasi jadwal terbang di maskapai saya bekerja sangat nyaman	The flight schedule rotation on my airline works very conveniently	0.668
P2.6	Saya nyaman dengan kondisi pesawat di maskapai saya bekerja	I am comfortable with the condition of the aircraft in the airline I work for	0.637
P2.8	Saya tidak nyaman dengan lingkungan di bandar udara	I am not comfortable with the environment at the airport	0.353
P3.1	Saya merasa bangga dengan kehidupan sebagai penerbang	I feel proud of life as a pilot	0.549
P3.2	Saya enggan dikenal sebagai seorang penerbang	I am reluctant to be known as a pilot	0.097
P3.5	Saya bersyukur dapat bekerja sebagai penerbang	I am grateful to be able to work as a pilot	0.421
H1.1	Kondisi tubuh saya fit sebagai penerbang	My body condition is fit as a pilot	0.444
H1.2	Saya merasa fisik saya cukup kuat untuk menghadapi jadwal terbang dengan waktu tidak menentu	I feel that I am physically strong enough to face flight schedules with uncertain times	0.533
H1.3	Saya mudah merasa lelah setelah menjalankan tugas terbang	I get tired easily after running errands	0.336
H2.1	Saya mudah merasa tertekan dalam kehidupan sebagai penerbang	I easily feel pressured in life as a pilot	0.699
H2.2	Profesi penerbang membuat saya mudah merasa kesepian	The flying profession makes me feel lonely easily	0.593
H2.4	Saya mendapatkan jatah cuti yang layak dari perusahaan untuk membuat saya lebih relaks	I get proper time off from the company to make me more relaxed	0.477
H2.6	Pekerjaan saya membuat saya lebih rajin beribadah	My job makes me more diligent in worship	0.318

H3.1	Profesi saya membuat saya memilih makan makanan yang sehat	My profession makes me choose to eat healthy food	0.467
H3.2	Saya rutin berolahraga walaupun saya sibuk dengan tugas terbang	I exercise regularly even though I am busy with flying duties	0.194
H3.3	Saya mengalami kesulitan tidur dengan pola teratur	I have trouble sleeping on a regular pattern	0.477
H4.1	Saya memiliki indikasi penyakit jantung, hipertensi,diabetes	I have indications of heart disease, hypertension, diabetes	0.198
H4.3	Saya menderita sakit maag	I have stomach ulcers	0.317
H4.4	Seberapa sering anda gagal/ recheck dalam medical check-up?	How often do you fail/recheck in medical check-ups?	0.169
S1.1	Saya mampu membagi jadwal terbang dan kehidupan pribadi dengan baik	I am able to divide flight schedule and personal life well	0.567
S1.2	Saya tidak membawa urusan pribadi pada saat melakukan tugas terbang	I do not bring personal matters when flying assignments	0.388
S1.4	Saya merasa tertekan saat harus meninggalkan rumah/keluarga pada saat bertugas	I feel pressured when I have to leave my home/family while on duty	0.535
S2.1	Saya mempunyai banyak teman yang memberikan dukungan positif pada saya	I have many friends who give me positive support	0.468
S2.3	Saya memiliki komunitas penerbang yang mendukung saya	I have a community of aviators that support me	0.510
S2.4	Saya merasa lingkungan saya mendukung saya menjadi penerbang	I feel my environment supports me to be a pilot	0.473
S3.1	Saya memiliki relasi yang baik dengan keluarga saya	I have a good relationship with my family	0.534
S3.3	Keluarga saya selalu menantikan kedatangan saya di rumah sepulang saya menjalankan tugas terbang	My family always looks forward to welcoming me home after my flying assignments	0.234
S3.4	Saya menikmati waktu libur saya bersama orang-orang yang saya cintai	I enjoy my time off with the people I love	0.398
R1.1	Saya mempunyai riwayat performa terbang yang baik	I have a history of good flying performance	0.491
R1.2	Saya merasa hasil kerja saya dalam menjalankan tugas terbang kurang diakui oleh perusahaan	I feel that my work in carrying out flying assignments is not recognized by the company	0.587
R1.4	Kinerja saya dipuji oleh atasan saya	My performance is praised by my superiors	0.433
R2.1	Saya merasa mendapatkan jadwal terbang yang adil dari perusahaan	I feel I get a fair amount of flying from the company	0.647
R2.4	Saya diterima di lingkungan kerja saya sama seperti orang lain	I am accepted in my work environment just like everyone else	0.633
R2.5	Semua level jabatan di masakapai saya didengarkan suaranya dalam pengambilan keputusan	All levels of positions in my airline are heard in decision-making	0.527

R3.1	Saya merasa upah yang diberikan perusahaan sesuai dengan resiko pekerjaan saya	I feel the wages given by the company are in accordance with the risk of my work	0.633
R3.2	Saya bersyukur dengan keuntungan dan keistimewaan dari perusahaan saya (selain upah)	I am grateful for the benefits and privileges of my company (besides wages)	0.660
R3.3	Saya merasa maskapai memberikan fasilitas kesehatan yang baik kepada saya	I feel the airline provides good health facilities to me	0.595
R3.4	Saya merasa puas dengan hotel yang dipilih maskapai saya untuk remain overnight duty.	I am satisfied with the hotel my airline chose to remain on overnight duty.	0.671
C1.2	Saya takut ketika menghadapi proficiency check	I was scared when I faced a proficiency check	0.553
C1.1	Saya dapat menyelesaikan proficiency check saya dengan hasil yang memuaskan	I was able to complete my proficiency check with satisfactory results	0.482
C1.5	Saya mampu menjalankan tugas terbang saya sesuai dengan standar operating procedure	I am able to carry out my flying duties in accordance with standard operating procedures	0.467
C2.1	Saya mampu berkonsetrasi dengan baik selama tugas terbang	I am able to concentrate well during flying assignments	0.550
C2.2	Saya mampu berfikir jernih dalam situasi apapun tidak terkecuali situasi kritis/ darurat	I am able to think clearly in any situation, including critical/emergency situations	0.449
C2.3	Saya memiliki situational awareness yang baik setiap tugas terbang.	I have good situational awareness of every flying assignment.	0.495
C3.2	Saya mampu membuat keputusan dalam situasi apapun	I am able to make decisions in any situation	0.556
C3.4	Saya mampu mengambil keputusan dengan cepat dalam tugas terbang	I am able to make quick decisions on flying assignments	0.487
C3.5	Terkadang saya merasa gugup ketika membuat keputusan dalam tekanan	Sometimes I get nervous when I make decisions under pressure	0.480
C3.6	Saya mampu mengendalikan situasi darurat dengan baik	I am able to handle emergency situations well	0.560
K1.2	Terkadang saya merasa ingin berhenti menjadi penerbang.	Sometimes I feel like I want to quit being a pilot.	0.574
K1.3	Saya puas bekerja dengan tipe pesawat saya saat ini	I am satisfied working with my current aircraft type	0.440
K1.4	Saya merasa aman bekerja di perusahaan saya saat ini	I feel safe working at my current company	0.672
K2.2	Profesi penerbang kurang terjamin masa depannya	The future of the pilot profession is less secure	0.627
K2.3	Perusahaan menjamin kemanan pekerjaan saya sebagai penerbang sampai saya pensiun	The company guarantees the security of my job as a pilot until I retire	0.571

K3.1	Saya merasa aman karena diberikan kompensasi jika saya kehilangan lisensi penerbang secara mendadak	I feel safe because I am compensated if I suddenly lose my pilot's license	0.330
K3.3	Saya merasa dapat bekerja di airlines saya hingga saya pensiun.	I think I can work in my airlines until I retire.	0.617
K3.2	Saya merasa takut akan kesejahteraan hidup saya pada saat saya pensiun sebagai penerbang.	I fear for the well-being of my life when I retire as a pilot.	0.495

Table 4. The list of the final 20 items

No	Bahasa Indonesia	English	$r_{item-total}$ (total score of 20 items)
1	Saya bahagia dengan hidup saya sebagai penerbang (F)	I am happy with my life as a pilot (F)	0.593
2	Saya merasa tenang pada saat tugas terbang (F)	I feel calm when flying (F)	0.691
3	Saya merasa bangga sebagai seorang penerbang (F)	I feel proud as a pilot (F)	0.572
4	Kondisi tubuh saya saat ini *fit* sebagai penerbang (F)	My current body condition is fit as a pilot (F)	0.410
5	Saya mudah merasa tertekan dalam kehidupan sebagai penerbang (UF)	I easily feel pressured in life as a pilot (UF)	0.688
6	Saya mengalami kesulitan tidur secara teratur (UF)	I have trouble sleeping on a regular pattern (UF) (frequency)	0.494
7	Saya merasa tertekan saat harus meninggalkan rumah/keluarga pada saat bertugas (UF)	I feel pressure when I have to leave home/family while on duty (UF)	0.458
8	Saya mempunyai banyak teman yang memberikan dukungan positif pada saya (F)	I have many friends who give me positive support (F)	0.482
9	Saya memiliki relasi yang baik dengan keluarga saya (F)	I have a good relationship with my family (F)	0.533
10	Saya merasa hasil kerja saya dalam menjalankan tugas terbang kurang diakui oleh perusahaan (UF)	I feel that my work in carrying out flying assignments is not recognized by the company (UF)	0.632
11	Semua level jabatan di masakapai saya didengarkan suaranya dalam pengambilan keputusan (F)	All levels of positions in my airline are heard in decision-making (F)	0.576
12	Saya merasa upah yang diberikan perusahaan sesuai dengan resiko pekerjaan saya (F)	I feel the wages given by the company are in accordance with the risk of my work (F)	0.629

13	Saya merasa maskapai memberikan fasilitas kesehatan yang baik kepada saya (F)	I feel the airline provides good health facilities to me (F)	0.645
14	Saya merasa puas dengan hotel yang dipilih maskapai saya untuk *remain overnight duty*. (F)	I am satisfied with the hotel my airline has chosen to remain on night duty. (F)	0.708
15	Saya dapat menyelesaikan *proficiency check* saya dengan hasil yang memuaskan (F)	I was able to complete my proficiency check with satisfactory results (F)	0.446
16	Saya memiliki *situational awareness* yang baik setiap tugas terbang. (F)	I have good situational awareness of every flying task. (F)	0.473
17	Saya mampu mengendalikan situasi darurat dengan baik(F)	I am able to handle emergency situations well(F)	0.537
18	Saya merasa aman bekerja di perusahaan saya saat ini(F)	I feel safe working in my current company(F)	0.636
19	Saya merasa masa depan saya terjamin dengan profesi sebagai penerbang (F)	I feel my future is secure with the profession as a pilot (F)	0.699
20	Profesi penerbang kurang terjamin masa depannya (UF)	The future of the flying profession is less secure (UF)	0.644

Reliability

The scale with 62 items is reliable with alpha 0.95 (95% CI *lower bound* = 0.94 dan 95% CI *upper bound* = 0.97). All dimensions are also reliable, with an alpha of 0.83 for positive emotion, 0.78 for health and social relationship, 0.89 for competence and recognition, and 0.86 for security. The reliability coefficient for the short version with 20 items was satisfactory, with an alpha of 0.92.

Conclusion

This study aimed to develop a well-being scale for the pilot. We successfully created 61 items with suitable validity and reliability for both long and short versions. This research is outstanding because of two aspects, and firstly, the indicator includes interviews and literature reviews. Secondly, the participants were pilots with a wide range of ages and flying experiences, guaranteeing data heterogeneity. For the following research, we suggest including confirmatory factor analysis and correlation with other well-being scales for construct validity testing and performance indicator for predictive validity.

References

Anastasi, A., & Urbina, S. (1997). *Psychological testing*. (7th ed.). Prentice Hall/Pearson Education.

Antara. (2019). Survei: Pilot paling banyak mengalami gangguan jantung. *Tempo*. https://bisnis.tempo.co/read/1260821/survei-pilot-paling-banyak-mengalami-gangguan-jantung/full&view=ok

BBC. (2022). *China Eastern plane crash likely intentional, US reports say*.

Bellamy, W. (2015). *AirAsia Flight 8501 Crash Caused by Pilot Error, Rudder Units*. AviationToday.Com. https://www.aviationtoday.com/2015/12/03/airasia-flight-8501-crash-caused-by-pilot-error-rudder-units/

Branch, A. A. I. (2021). *Deviation from expected flightpath during a go-around, Aberdeen Airport, 11 September 2021.* AAIB Special Bulletin S2/2021 on Boeing 737-8K5, G-FDZF. https://www.gov.uk/aaib-reports/aaib-special-bulletin-s2-slash-2021-on-boeing-737-8k5-g-fdzf

Butler, J., & Kern, M. L. (2016). The PERMA-Profiler: A brief multidimensional measure of flourishing. *International Journal of Wellbeing*, *6*(3), 1–48. https://doi.org/10.5502/ijw.v6i3.526

Crocker, L. M., & Algina, J. (2008). Introduction to classical and modern test theory. New York: Holt, Rinehart, and Winston.

Flight Safety Foundation. (2020, April). *An aviation professional's guide to well being*

Huppert, F. A., & So, T. T. C. (2013). Flourishing Across Europe: Application of a New Conceptual Framework for Defining Well-Being. *Social Indicators Research*, *110*(3), 837–861. https://doi.org/10.1007/s11205-011-9966-7.

Grant, J. S., & Davis, L. L. (1997). Selection and Use of Content Experts for Instrument Development. Research in Nursing & Health, 20, 269-274. https://doi.org/10.1002/(SICI)1098-240X(199706)20:3<269::AID-NUR9>3.0.CO;2-G

Gravetter, F. J., Wallnau, L. B., & Forzano, L. A. B. (2018). *Essentials of Statistics for the Behavioral Sciences* (9th ed.). Cengage Learning.

Nrangwesti, A. (2011). Aspek yuridis normatif tentang pilot pesawat udara. *Syiar Hukum: Jurnal Ilmu Hukum, 13* (1). https://media.neliti.com/media/publications/25266-ID-aspek-yuridis-normatif-tentang-pilot-pesawat-udara.pdf

Page, K. M., & Dianne, V. B. (2005). The Working for Wellness Program: RCT of an Employee Well-Being Intervention. Journal of Happiness Studies. 14. 10.1007/s10902-012-9366-y.

Polit, D.E. and Beck, C.T. (2006) Essentials of Nursing Research. 6th Edition, Lippincott Williams & Wilkins, Philadelphia.

Torrence, J. (2018). *Pilot Error : The Lion Air Flight 904 Crash*. Embry Riddle Aeronautical University.

Saputra, A. D. (2016). Studi Analisis Penyebab Runway Excursion di Indonesia Berdasarkan Data Komite Nasional Keselamatan Transportasi (KNKT) Tahun 2007-2016. Warta Ardhia. DOI:10.25104/wa.v43i2.305.93-104.

Seligman, M. E. P. (2011). *Flourish: A visionary new understanding of happiness and wellbeing.* Free Press.

Syabilah, C. A., Wahyuni, I., & Widjasena, B. (2015). Faktor-faktor yang berhubungan dengan stres kerja pada penerbang sipil di balai kesehatan penerbangan jakarta. *Jurnal Kesehatan Masyarakat, 3*(1), 328-338. https://doi.org/10.14710/jkm.v3i1.11390

Waltz, Carolyn & Strickland, Ora & Lenz, Elizabeth. (2016). Measurement in Nursing and Health Research. 10.1891/9780826170620.

Wilson, W., Driller, M., Johnston, B., & Gill, N. (2022). The prevalence and distribution of health risk factors in airline pilots: A cross-sectional comparison with the general population. *Australian and New Zealand Journal of Public Health,* 1-9. doi: 10.1111/1753-6405.13231

Educational Quality Mapping Instrument (IP-SNP)

Rina Mutaqinah[1]
Bambang Sumintono[2]
Yanti Triatna[1]

[1]BBPMP Provinsi Jawa Barat
[2]Universitas Islam Internasional Indonesia (UIII)

https://doi.org/10.37517/978-1-74286-697-0-10

Abstract

Decision-making must be based on valid data, as well as decision-making in the field of education. It is very important to have a valid instrument, therefore it is necessary to develop an educational quality mapping instrument that can measure the achievement of the quality of education. This instrument was developed based on educational quality standards, namely the competency quality of graduates, content, process, assessment, educators and education personnel, infrastructure, management, and financing. The purpose of this study was to obtain an instrument that has high validity. The educational quality mapping instrument (IP-SNP) uses a rating scale, collecting data through a survey of principals in 19,460 elementary schools (SD) in West Java Province. Construct validity was carried out by 3 experts in educational measurement and management, several items were revised based on expert input. The data processing and analysis method in this study uses the IRT (Item Response Theory) method with the Rasch Model Winstep application version 3.7. The results of the Rasch Model analysis show that the SD level IP-SNP instrument has a Cronbach Alpha score of 0.90 which means that the reliability of persons and items is high, while the reliability of 0.93 means that the instrument is very good. Item validity of 0.97 means that the item can measure the achievement of the quality of education, strengthened by the separation of 2.85 items. This question has a good distribution of responses. Educational quality mapping instruments can be used.

Background

Policymaking requires data that is valid, easy to read and up to date. The data is used as a material consideration in every decision-making. If a policy is formulated without data, it is certain that the policy will not be able to become a problem solver.

The monitoring dashboard for RKPD inspection on data-based planning in all local governments in Indonesia shows that education development planning for the 2023 fiscal year is 93.22% not in accordance with the performance indicators for affairs (IKU) that have been determined by the Ministry of Education and Culture based on the Education report card (Bagren-Datin Kemdikbudristek, 2022)[1]. The discrepancy between education quality data and education planning in the regions was mostly caused by a lack of coordination and cooperation between the Education Development Planning section and the Education Report Card Analysis Team, besides that there were difficulties for the Analyst Team in translating data on education report cards (BBPMP Provinsi Jawa Barat, 2022).

Anticipating education report card instruments originating from data on the results of AN (National Assessment), Dapodik and others, which can only be processed and analyzed by the Central Government, an instrument is needed that makes it easier for the Regional Government and/or Education units to measure the achievement of program implementation results for internal purposes and as comparative data on educational reports. The instrument developed is an instrument based on national education standards PP Number 4 of 2022 concerning Amendments to PP Number 57 of 2021 concerning National Education Standards in 2021. The standards used are Management Standards; Educators and Education Personnel; and Financing. Instruments for other standards based on the latest PP SNP (Competencies of Graduates, Content, Process, Assessment, Infrastructure) have not been developed.

Education Quality Standards

There is no universally accepted definition of quality. Quality of Education is defined in the context of the education system as a whole (including schools and related bodies, teaching-learning environment, policies, etc.) and the quality of what the system offers to students/learners (i.e., quality of teaching and learning processes, curriculum etc.). Terms such as efficiency, effectiveness, equity, and quality are often used synonymously. To improve the quality of education it is necessary to improve the quality of its components. According to our definition, all educational space is the component itself: standards and curricula; academic literature; teaching staff (their professional skills); monitoring of education; moral and patriotic education; scientific research in the sphere of education (Kousainov, 2016)[2]

According to Hoy et al., 2005, quality is often in terms of outcomes to match a customer's satisfaction. This gives rise to the definition of quality as the extent to which the outcomes meet the customer's requirements. Competences base quality results need to be supported by a capability to deliver the service on the part of everyone involved. The inevitable outcome of this system is to reward those who can deliver a quality product, and train those who aren't skilled.

(Goddard & Leask, 1992) state the definition of quality from different perceptions, as

[1] Bagren-Datin Kemdikbudristek, *Dashboard PBD .*(Retrieved from Dashboar Monitoring Pemeriksaan RKPD, 2022, September 27).: https://metabase.pauddasmen.id/dashboard/5374-iku-kabupaten-kota

[2] Kousainov, A. K*., The ways of improving the quality of the secondary education in the Republic of Kazakhstan*. EEIA-2016 (p. SHS Web of Conferences). Astana, Kazakhstan: EDP Sciences.

meeting customer needs. They have included different customers for education—parents, governments, students, teachers, employers, and institutions—who seek different quality characteristics. Education is a service and not a product, its quality cannot lie exclusively in the final output. Its quality should also be manifested in the delivery process. Quality of education should also take into account determinants such as the provision of teachers, building, curriculum, equipment, textbooks and teaching processes (Grisay & Mahlck, 1991)[3].

For them, the quality of education has a three-dimensional approach consisting of the quality of human resources and materials available for teaching (input), teaching practice (process), and results (outcomes). Furthermore, according to them, there are several indicators—repetition, dropout, promotion, and transition rates—that planners frequently visit to arrive at approximate quality measures. UNICEF also strongly emphasized the desirable dimensions of quality, as identified in the Dakar Framework. Its paper 'Defining Quality in Education' recognizes five dimensions of quality: learners, environment, content, processes, and outcomes, founded on 'the rights of the whole child, and all children, to survival, protection, development, and participation (UNICEF Programme Division, United Nations Children's Fund (UNICEF, 2000). The Communiqué of the World Conference on Higher Education 2009 states that 'Quality criteria must reflect the aim of cultivating in students critical and independent thought and the capacity of learning throughout life. They should encourage innovation and diversity (UNESCO 2009).

Mapping the quality of education is one of the education evaluation processes by measuring the achievement of national education standards (SNP) listed in the Government Regulation of the Republic of Indonesia Number 32 of 2013 concerning Amendments to Government Regulation Number 19 of 2005 concerning National Education Standards. The Education Quality Assurance Institute (LPMP) of West Java Province is the Ministry of Education's Technical Implementation Unit (UPT) located in West Java Province according to Permendikbud 28 of 2016 concerning SPMP for primary and secondary education, has the main task of carrying out quality assurance of education in West Java Province including mapping quality of education, facilitation of quality improvement, assistance to the regions and monitoring of educational evaluations.

The results of quality mapping will serve as the basis and reference in determining program planning in the regions or in educational units because mapping the quality of education or measuring the achievement of the eight SNPs is an initial part of improving the quality of education. Measuring the quality of education must of course be supported by measurement instruments that can represent all the indicators contained in the eight SNP.

So far, the Ministry of Education and Culture through the Directorate of Primary and Secondary Education has measured the achievement of the SNP through e-EDS by completing the PMP instrument every year. The PMP instrument, which accommodates all educational units in Indonesia, certainly experiences many problems both in substance and technically, so the results of this national quality mapping cannot be quickly accessed and

[3] Grisay, A., & Mahlck, L. *A Preview of Some Research Studies and Policy Documents: The quality of education in developing countries* . (Paris: IIEP, 1991)

utilized by stakeholders in the regions or in educational units. Therefore, it is necessary to develop a similar alternative quality mapping instrument that can be quickly accessed by regional governments and education units.

LPMP West Java has developed this alternative quality mapping instrument called the SNP Measurement Instrument (IP SNP), which is expected to meet regional and educational unit interests in West Java Province. The IP SNP was developed based on the key indicators contained in the eight SNPs. It is hoped that the results of measuring SNP achievements through IP SNP will produce accurate data in the form of profiles, quality maps that show the achievements of the strengths and weaknesses of educational units and regions in achieving education quality indicators in the eight SNPs, which are then reported to relevant stakeholders as a basis for further policy making. Measuring the achievement of educational quality can be done using various data collection techniques such as rating scales, questionnaires, observations and interviews.

Attitude measurement data could be processed and analyzed using a model of: 1) classic, 2) modern or IRT (Item Response Theory). IRT had various models including a model of 1 parameter, 2 parameters and 3 parameters and a model which was similar to the 1 parameter namely Rasch models. Rasch Model was developed by Georg Rasch in the 1950s (Naga, 2013)[4]

Rasch Model was one of IRT (Item Response Theory) models. It was the general framework of a specific mathematical function that described the interaction between the person (persons) and items (item test). IRT was not dependent on the sample or respondents chose in a test (free item and free person free). Therefore, it caused this pattern to be more precise measurement and calibration on items was carried out (Sumintono & Widhiarso, Aplikasi Rasch Model untuk Ilmu-ilmu Sosial, 2014)[5].

Research Method

The research method was carried out using the instrument development method from Robert K. Gable, with the following development steps: (1) developing a conceptual definition, (2) developing an operational definition, (3) choosing a scaling technique, (4) reviewing/justifying items, (5) selecting the response format, (6) compiling instructions for the response, (7) preparing a draft instrument and conducting an initial trial (8) preparing the final instrument, (9) collecting final data, (10) analyzing trial data using factor analysis techniques, item analysis and reliability, (11) revise the instrument, (12) conduct the final trial, (13) produce the instrument, (14) perform additional validity and reliability analysis, and (15) prepare a test manual (Gable, 1986)

Data analysis was performed using the Rasch model approach through the Winsteps program. In the Rasch model approach, in addition to paying attention to item items, it also

[4] Naga, D. S. *Teori Sekor pada Pengukuran Mental* (Jakarta: PT Nagarani Citarasa, 2013)
[5] Sumintono, B., & Widhiarso, W, *Aplikasi Rasch Model untuk Ilmu-ilmu Sosial* (Cimahi: Trim Komunikata Publishing House, 2014)

pays attention to the respondent's aspects and calculates the magnitude of the correlation. The results of the analysis shown are statistical summary, item accuracy index, respondent accuracy index, scalogram, unidimensionality, respondent item map, and rating scale analysis.

Based on the stages of development that have been described, the first trial was conducted on 3 education management experts and instrument development experts consisting of measurement lecturers and education quality management lecturers. Expert experts review the constructs, dimensions, indicators and items of educational quality instruments. The instrument was then tested on … 30 teachers and school principals to test the legibility of the instrument, which was then given to 16 thousand elementary schools in West Java. Then analyze the validity of the instrument using the RASCH model with the Winstep type 3.1 application.

This research applied Rasch Model analysis to the instrument with of the 38 items using Winsteps version 3.73 software (Linacre, 2013)This was conducted to investigate overall respondents' agreement to the items based on each school, and regarding the item difficulty estimates in the instrument. Rasch modelling is built on conjoint measurement which is a formulation that stipulates the relationship between a person (e.g., respondent) and an item based on a mutual latent trait ((Andrich, 1988); (Bond & Fox, 2007))Rasch modelling produces a measurement scale with equal interval units called logit (logarithm odd unit) that shows the level of difficulty of each item and the level of each person agrees to the items (Alagumalai & Curtis, 2005; Sumintono & Widhiarso, 2014). Rasch measurement models then can inform overall trends of items and respondents based on the logit distribution of item and persons.

Several instrument analysis indexes from Winsteps are used to know the quality of the instrument and its item. Fit statistics from the Rasch model for example, such as outfit of mean-square and z-standard and point measure correlation indicate how good is the item really measured based on respondents' responses (Bond & Fox, 2007). Other important issues regarding measurement are unidimensionality and different item functioning (DIF) which can detect if is there anything wrong at instrument and item levels (Boone, Staver, &Yale, 2014). Another important piece of information to know the quality of instruments and items is the Wright map or item-person map which can illustrate a comprehensive pattern of response ((Bond & Fox, 2007); Boone, Staver, & Yale, 2014)[6].

Result and Discussion

The dimensions or aspects and indicators that have been developed are validated by experts in instrument development, management, and quality assurance; the quality aspects of graduates are obtained; contents/ content; learning process; learning assessment; educators and education personnel; infrastructure; management and financing of Education. From this aspect, it is further developed into indicators and instrument items. The results of

[6] Boone, W. S., *Rasch Analysis in the Human Sciences*. (Dordrecht: Springer, 2014)

the empirical trials show: Out of all the items/indicators in the 'IP SNP' instrument, three items/indicators from the three constructs must be repaired. This is because the items/indicators are not in accordance with the model (misfit), namely in the honesty construct (4th item/indicator) and one item/indicator in the integrity construct (2nd item/indicator); whereas the 19th item/indicator (honesty-1) gives the same response pattern as the other items/indicators in the honesty construct (20th item/indicator), so it also needs to be rearranged.

Table 1. Value of Item Suitability Criteria

Criteria	**Value**
Outfit Mean Square (MNSQ) accepted	0.5<MNSQ<1.5
Outfit Z-Standard (ZSTD) accepted	-2.0<ZSTD<+2.0
Point Measure Correlation (Pt Measure Corr)	0.4<Pt Measure Corr<0.85

Based on the results of the analysis using the Rasch model, item information and participant responses were obtained in the scale trial (person). In this study, data analysis was performed several times to obtain a number of items that met the item-model accuracy index. The stages of analysis are summarized in Table 2 below. In the first stage of analysis, 63 respondents were identified as outliers or not quite right with the model. According to Boone, Staver, & Yale (2014), the parameters used to determine the accuracy or suitability of respondents include: first, the received outfit mean square (MNSQ) value: 0.5 < MNSQ < 1.5. Second, the value of the Z-standard outfit (ZSTD) received: -2.0 < ZSTD < +2.0. Third, value of Point Measure Correlation 0.4<Pt Measure Corr<0.85. The following table presents a summary of the data analysis results.

Table 2 Summary of Final Analysis Results

Summary of Final Analysis Results

	Output	*Results*
Item	Reliability Item	1,00
	Index Separation (H)	64,77
	Strata (H)	86,69
	Highest Logit Value	2,72 Logit
	Lowest Logit Value	-1,70 Logit
Respondents	Respondent Average Value	1,56 Logit
	Respondent Reliability	0,89
	Index Separation	2,85
	Index Separation (H)	4,13
	Highest Logit Value	6,49 Logit
	Lowest Logit Value	-5,33 Logit
Instrument	Alpha Cronbach	0,93
	Raw Variance Explained by Measures	44,6%
	Unexplained Variance in 1st Contrast	9,1%
	Unexplained Variance in 2nd Contrast	5,1%

The results showed that the construct validity was performed on a panel of experts who had

knowledge of the attitude scale of instrument development and or the environment. The items considered not 'good' were revised, particularly on item number 7, 8, 15, 26, 30, 37, and 38. The items processing used the Rasch model to get the item information and valid respondents, in accordance with the characteristics and the paradigm of the Rasch model. The function of the Rasch model was to provide directions and detect a problem with the instrument.

According to Boone, Staver, & Yale (2014), the values of outfit mean-square, outfit z-standard, and point measure correlation are the criteria used to see the level of item suitability. If the item does not meet the criteria, it is better if the item is repaired or replaced. Guidelines for assessing item suitability criteria according to Boone, 2014.

Table 3. Reliability Test Processing Results

Alpha Cronbach	*Interpretation*	*Item Reliability*	*Interpretation*	*Person Reliability*	*Interpretation*	*Conclusion*
0,93	**Very Good**	**1.00**	**Excelent**	**0.89**	**Very Good**	**Reliable**

Table 4. Validity Test Processing Results

Raw Variance Explained by Measure	*Interpretation*	*Unexplained Variance 1st Contrast*		*Interpretation*
		Eigenvalue	*Observe*	
44.6%	**Good**	**8.0**	**9.1%**	**Excellent**

The item reliability value of 1.00 indicates that the quality of the items in this instrument is very high. In other words, the 47 items identified as having accuracy with the model are indeed quality items. Furthermore, the respondent's reliability value of 0.89 indicates that the consistency of the respondents' answers is high.

In other words, respondents answered all items seriously (not carelessly).

Unidimensionality Item to answer the question of whether the developed instrument is able to measure what it should measure. In the context of this study, the educational quality mapping instrument, from Table 4 it is known that the raw variance data measurement results are 44.6%. According to Sumintono and Widhiarso (2014), the minimum requirement for unidimensionality is 20%, and if the value is more than 40%, then even better, and the variance that cannot be explained by the instrument should ideally not exceed 15%. Based on this explanation, the raw data variance of 44.6% indicates that the minimum unidimensionality requirement of 20% can be met and is even classified as good because it is more than 40%.The results of the analysis of variance that cannot be explained by the instrument of 13.5% also meet the criteria, namely, not exceeding 15%.

Based on Table 4. the results of the raw variance values explained by measures indicate that the test items for mapping the quality of education are in the "good" category. Furthermore, based on the values observed in the unexplained variance 1 contrast, it shows that there is no trend of the discrepancy between the items so that they can be used, but

the eigenvalues greater than 3 indicate that there are problematic items so that further analysis can be done with item fit order analysis to determine whether an item can be maintained or must be replaced.

Item fit referred to as item suitability can explain whether the items function normally to take measurements.

Outfit means-square, outfit z-standard, and point measure correlation are the criteria used to see the level of item fit (Item fit) (Boone, 2014). The criteria used to check the suitability of the items can be seen in Table 1.

Conclusions and Policy Proposal or Options

The results of the alpha reliability coefficient of 0.91 indicate that the self-efficacy scale in career decision-making has a high-reliability coefficient. That is, this scale produces a measurement score that is consistent and reliable. The reliability coefficients of the items and the respondents are also quite good, namely 0.91 and 0.91. This shows that these twenty items are quality items, and the group of respondents answered them seriously. These two results further strengthen and confirm that the self-efficacy scale in career decision-making is indeed a quality measurement tool because not only are the measurement results reliable but also the twenty items are quality items.

Instrument development begins with developing a blueprint and items, then validating the construct to the experts. Revise blueprints and items according to input from experts, then the results were analyzed using the Rasch Model. From 86 items, it was eliminated to 47 items. Item reliability in the very good category, Reliable. Item validity in the good category. based on the values observed in the unexplained variance 1 contrast, it shows that there is no trend of the discrepancy between the items so that they can be used, but the eigenvalues greater than 3 indicate that there are problematic items.

Based on the results of this study, the items for this quality mapping instrument can be used by all schools with the standard requirements and the dimensions measured are the same as this instrument.

References

Alagumalai, S., & Curtis, D. (2005). *Applied Rasch Measurement: Book of Exemplars Papers in Honour of John P Keeves.* Dordrecht: Springer.

Bagren-Datin Kemdikbudristek. (2022, September 27). *Dashboard PBD* . Retrieved from Dashboar Monitoring Pemeriksaan RKPD: https://metabase.pauddasmen.id/dashboard/5374-iku-kabupaten-kota

BBPMP Provinsi Jawa Barat. (2022). *Laporan Pendampingan Penyusunan Policy Brief dan Penyesuaian IKU Pemda Kab./ Kota di Provinsi Jawa Barat.* Bandung : BBPMP Provinsi Jawa Barat.

Boone, W. S. (2014). *Rasch Analysis in the Human Sciences.* Dordrecht: Springer.

Grisay, A., & Mahlck, L. (1991). *A Preview of Some Research Studies and Policy Documents: The quality of education in developing countries .* Paris: IIEP.

Hoy, C., Baryne-Jardine, C., & Wood, M. (2005). *Improving Quality of Education.* New York : Falmer Press.

Kousainov, A. K. (2016). The ways of improving the quality of the secondary education in the Republic of Kazakhstan. *EEIA-2016* (p. SHS Web of Conferences). Astana, Kazakhstan: EDP Sciences. Retrieved from https://www.bing.com/ck/a?!&&p=e33c4dc19faef96dJmltdHM9MTY2NDIzNjgwMCZpZ3VpZD0wYTI5MTYwYy0xNzZiLTZlNjktMzNkYi0wNDJiMTYzZDZmMDMmaW5zaWQ9NTU1Mw&ptn=3&hsh=3&fclid=0a29160c-176b-6e69-33db-042b163d6f03&u=a1aHR0cHM6Ly93d3cucmVzZWFyY2hnYXRlLm5ldC9wdWJsaWNhdGlvbi

Naga, D. S. (2013). *Teori Sekor pada Pengukuran Mental.* Jakarta: PT Nagarani Citarasa.

Peraturan Pemerintah. (2013). Peraturan Pemerintah tentang Perubahan Atas Peraturan Pemerintah Nomor 19 Tahun 2005 Tentang Standar Nasional Pendidikan.

Permendikbud. (2016). Peraturan Menteri Pendidikan dan Kebudayaan Nomor 28 Tahun 2016 tentang Sistem Penjaminan Mutu Pendidikan Dasar dan Menengah.

Peraturan Pemerintah. (2022). Peraturan Pemerintah tentang Perubahan atas Peraturan Pemerintah Nomor 57 Tahun 2021 tentang Standar Nasional Pendidikan. Retrieved from https://peraturan.bpk.go.id/Home/Details/196151/pp-no-4-tahun-2022

Sumintono, B., & Widhiarso, W. (2014). *Aplikasi Rasch Model untuk Ilmu-ilmu Sosial.* Cimahi: Trim Komunikata Publishing House.

UNICEF Programme Division, United Nations Children's Fund (UNICEF). (2000). *Defining quality in education. A paper presented by UNICEF at the meeting of the international working Education Section.* USA: Programme Division, United Nations Children's Fund (UNICEF).

Are women more resilient in literacy and numeracy than men?

Sri Irene Astuti Dwiningrum[1]
Rahmawati[2]
Aditya Ramadhan[2]
Dian Rahdiani[2]
Haryo Susetiyo[2]
Mega Riyanti Bayu Putri[2]

[1]Faculty of Education, Yogyakarta State University
[2]Centre for Education Assessment, Standards Agency Education Curriculum and Assessment

https://doi.org/10.37517/978-1-74286-697-0-11

Abstract

Based on data from the Indonesian Child Protection Commission (KPAI) from 2011 to 2016, there were 7,698 cases of children in conflict with the law, 2,435 cases of children in the education sector (perpetrators of student brawls, perpetrators of violence at school, and so on), and 1,709 cases of children related to pornography and cybercrime. The data shows the importance of student resilience in order to fortify themselves from various behaviors that can harm their future. The formation of student resilience has not been designed comprehensively in the learning process at school. Therefore, there are still many students who have not been able to overcome problems, one of which has an impact on achievement that is not optimal. In fact, resilient students will have better adaptability in overcoming various problems. In general, student resilience is defined as the ability of students to bounce back or rebound after experiencing pressure or stress. Resilience is needed to achieve literacy and numeracy achievements for students in Indonesia. The existence of inequality in the learning process in schools has an impact on differences in student achievement in Indonesia. The level of resilience of students' literacy and numeracy at all levels of education based on data shows that there is a tendency for female students to dominate. Descriptive analysis was carried out in this study to test the hypothesis. The results showed that the level of literacy and numeracy resilience of students nationally from the elementary school level to the same level was highest for female students with a percentage of 11.28% for the level of literacy endurance and numeracy endurance level. the field is at a percentage of 5.97%. for male students. Based on the analysis of the data above, it can be interpreted that from a gender perspective, women's literacy and numeracy resilience is rated higher than that of men. This finding still needs to be strengthened by analyzing the data by calculating in more detail the effect of education level. The literacy and numeracy resilience profile describes the social dynamics that are determined by the level of education. This proves that resilience is not limited by gender but is also related to the learning process experienced by different students.

Introduction

Literacy is always related tightly with development society – no except in the marked modern era with development technology information. In fact, standard education in the modern era also seems is at in influence literacy. In period time two decades final for example, surveys conducted institution international for knowing level success education in a country is through measurement ability literacy students (OECD, 2016; Viennet & Pont, 2017) and literacy population age adults (Martin, 2018; OECD, 2000, 2011). If linked with fact social prove that literacy becomes issue related matters with educational reform efforts no only at the global level ; as response on PISA report *(Program for International Students Assessment)* and TIMSS *(Trends in International Mathematics and Science Study)* published by the OECD *(the Organization for Economic Co-operation and Development),* but have also Becomes policies in the local area - national , for example only in England (Arthur et al., 2014), Canada (Bellous & Clinton, 2016), and Indonesia (Effendy, 2019; Ministry of Education and Culture, 2016a).

Measurement about literacy science, literacy mathematics (numeracy), literacy reading and literacy financial still becomes interesting study for researched. This occur because results the PISA *(Program for International Students Assessment)* study, it turns out no always linear with success participant educate after mature or after graduating from level education particular (OECD, 2015; So & Kang, 2014) because various factors that influence it. Because it, for answer challenge the orientation of the educational process that facilitates individual in order to Becomes learner throughout life *(lifelong learning)* and make it happen education for life sustainable *(education for sustainable living)* also started done. In the report issued by the OECD *(the Organization for Economic Co-operation and Development)* in 2018, stated that *lifelong learning* delivers contribution for non-economic goals, e.g., fulfillment individual needs *(personal fulfillment)*, improvement health *(improved health)*, participation citizens (civic *participation)*, and inclusion social *(social inclusion)* (OECD, 2018).

Effort for build community that has high literacy be a program by every country. This is done by every country because gap quality education is not related with level literacy society. Literacy is the expression of culture on what 's going on around everyday individuals, fine that related with system always social cover it, class society that it inhabits, the power and domination that governs it, to relation social life (Freire & Macedo, 1987). Issuesbe called become theory discussion among educator with participant educate for stimulate emerge and develop awareness critical individual, up to awareness social collective, which integrates activity reading, writing, and doing action social or could called *'emancipatory literacy'* (Freire & Macedo, 1987). Various issues and phenomena gap social events that occur in Indonesian society, for example can becomes substance learning literacy in schools (Pratiwi, 2021) which is described portrait poverty and inequality in Indonesia which has going on more from half century (Hill, 2021; Rahman, 2017), gap economic and social between provinces in Indonesia (Kataoka, 2018; Marantika, 2018).

For Indonesia policy enhancement literacy conducted with the Literacy Movement School (GLS) implemented in schools. However so, because result not optimal, then still needed with studies academic latest, as well policies at the global level. In explore, describe, and make sense of practice literacy in school, necessary for think things new about " what ", "where", and " how " literacy practiced daily in life schooling (Halse, 2015a; Jones,

2018; Nyachae, 2019; Wilder & Msseemmaa, 2019; Pratiwi, 2021). Because that, Halse (2015a) proposes necessity exploration and elaboration (new) conception of relevant literacy with practice daily students in schools that have score pedagogical; for explain variety existing complexities and offers alternative/potential solution.
Student school face extraordinary challenge commonplace in society in the digital age. Most students will face serious challenges and threats problem social. Based on Commission data Indonesian Child Protection (KPAI) from 2011 to 2016, showing as many as 7,698 cases opposite child with law, 2,435 cases child in the field education (actors brawl student, actor violence at school, and so on), and 1,709 cases child related pornography and cybercrime. The data show the importance of resilience students so they can fortify self from various behaviors that can jeopardize their future. Problem resilience student analyzed with involve all dimensions the environment around students, fine environment family, school nor society, as findings study (Dwiningrum, 2017), that environment home, community, and friends peer have influence direct to formation resilience student.

Formation resilience student not yet designed in a manner comprehensive in the learning process at school. Because it, still many students who haven't can resolve one of the problems have an impact on performance that is not optimal. In fact, students who are resilient will have power more adaptation good in resolve various problem. Kindly general resilience student defined as ability student for could rise back (*bounce back or rebound*) after experience pressure or stress (Moore, 2013). *The Center for the Study of Social Policy* (2005) stated that student toughness means capable call strength inner for in a manner positive face challenge, manage trouble, heal effects of trauma, and then develop becomes characteristics unique, which then capable adapt with objectives and conditions.

Resilience formed by the process of socialization experienced by students, both in the environment of family, school, and community. Because that's study about resilience becomes aspect important in support quality school and achievement student. Issue this interesting for discussed in a manner special because condition social economy student will form resilience in various supporting aspects success student in reach achievement. Paper this will discuss linkages resilience literacy and numeracy from various aspects. This is needed to give description comprehensive about linkages among gap social with resilience literacy and numeracy. Kindly special will study about: a) How profile resilience students in the field literacy and numeracy in a manner national; b) How profile resilience literacy and numeracy based on type sex /gender.

Resilience and Education

Growing students in poverty tend to be maladaptive. However, some students have the ability to survive and develop potency themselves with maximum. Education has role strategic in building resilience in schools. Success school in operate role social determined by strength resilience possessed by each inhabitant school. Because of that, quality education is determined by level resilience school (Dwiningrum, 2011). Teacher has role important in increasing quality school, like help student dig potency himself. The role of the teacher in the process of developing resilience school is very significant. Resilience school is condition dynamic something involving organizations determination

and power stand dig potency school for face threats, challenges, and obstacles from in nor outside environment school. Henderson (2003) mentions resilience school as an involving process six phase, such as: tighten connection with school, clarity rule, teach ability live, awareness and support, communicate, and embody good wishes with school. Teacher discipline is distinguished with clarity rules and punishments from head school. Teacher involvement in project school believed will have more results good because the teacher knows vision and mission school. However, reality During this teacher's understanding is limited to development cognitive students, as well not enough involved in the process of determining vision and mission school, resilience '*construct education*' no looked at as attribute permanent but as something that can promoted with focuses on factors that "can changed" which can affect success individual at school. Approach this no focus on attributes like ability because ability not yet of course found as characteristics tough students (Bernard, 1991; Masten, 2014). However, there are a number of possible factors that have changed that have been found for influence resilience in children. McMillan and Reed (1994), for example, found that there are four related factors with resilience like motivation and orientation purpose, use time in a manner positive (eg, behavior work task, completion profession home, participation in experience extracurricular), life family (eg, support and hope family) and the environment study school and classroom (i.e., facilities, exposure technology, leadership, and climate in a manner whole).

Student possible caught experience education that is not proper through family them, school, or community (Pallas et al., 1989). Temporary educator no could control demographics society and conditions family, they could change policies and practices education for ensure that they fulfill needs special at risk of students failing academics (Comer, 1987). Maker policies, administrators, teachers, and parents need understand why a number of student tough and doing well in school, while others from background behind social the same economy and school and neighborhood similar house no achievers good in a manner academic.

Teachers have an important role in building resilience in schools. The role of teacher and head school have important role in the development process resilience school because have strategic role for understand development student optimally and effectively keep going continuously and intensively. Besides that's school will be able to create conditions for good learning and developing management school in a manner creative and contextual. For successful teachers must have deep abilities for maintain effective communication, be empathic, listening aspirations effective student, and keep communication personal with student moment needed (Dwiningrum, 2008). In short, effective communication between teachers and students will occur if the teacher understands what is wanted students know what is felt students, appreciate what happened student during the learning process, and act objective with use easy language understood by students (Dwiningrum, 2008).

For build resilient students determined by strength elements shaper resilient personality. As for resilience individual as described by Reivich K. and Shatte A (1999), states that resilience covers seven abilities, namely: *regulation emotions, control impulse, empathy, optimism, analysis causal, self-efficacy, and reaching out*. Following is exposure to each of the factors covered in resilience.

Table 1. Seven aspect resilience according to Reivich K. and Shatte A., (1999)

Aspect	Information Aspect
a. Emotional Regulation	Regulation emotion is the ability for permanent calm in full condition pressure. Resilient individual use series skills that have developed for help control emotions, attention, and behavior.
b. Control impulse	Control impulse is closely related with ability regulation emotion. Individual with control low impulse tend accept belief in a manner impulsive, that is something situation as truth and action on base thing the condition this often raises consequence negative can hinder resilience.
c. Optimism	A resilient individual is an optimistic individual, confident that various things could be changed for the better. Have hope towards the future & believe that could control the direction of life. Compared to pessimistic people, optimistic individuals more health in a manner physical, inclined no experience depression, achievement more goods at school, more productive in work and more achievers in sports.
d. Causal Analysis	Analysis causal is terms used for refers to ability individual for in a manner accurate identify causes from problem. If someone no capable estimate reason from the problem in a manner accurate, then the individual will make same error.
e. Empathy	Empathy describes as good what somebody could read instruction from other related people with condition psychological and emotional of the person. A number of individuals could interpret behavior *non*-verbal others, such as expression face, tone of voice and language body as well as determine what the person thinks and worries about.
f. Self-Efficacy	*Self-efficacy* describe belief somebody for solve problems and beliefs _ _ somebody to ability for reach success.
g. Reaching Out	*Reaching out* describes the ability somebody for reach success. Resilience is source for reach *reaching out* because resilience possible for increased aspects positive in life.

Literacy and Numeracy

Literacy and numeracy are cross- domain studies, this is based on variety results stated study that resilience math at school. Becomes more predictor good from development adaptive in the domain than resilience literacy. This result is consistent with study previously shown ability mathematics beginning is predictor strongest mathematics (Claessens et al., 2009). If linked with resilience literacy Becomes learned skills moment enter school give benefit period short, so needed strengthening continuing skills in school needed for results ongoing positive. On the side that is, development Skills literacy far more specialized as described by Lonigan et al. (2013) found that assigned children with give intervention for focus on digging meaning in reading and assigned children for code - focused interventions show enhancement in awareness phonological or knowledge letters, but no there is indicated intervention improvement all over skills literacy. Future research must keep going explore how academic domain component certain, like awareness phonological or knowledge letter must next in a manner thorough so that all aspect literacy so that could develop more optimal.

The results of other studies also found consistent evidence about promotion skills cross-domain be ongoing study conducted. As results related research with resilience in math at the time enter school base predicted becomes base for develop skills more literacy height and resistance beginning in literacy push achievement performance in math. Because that is, the skills required for studying math and literacy depend more on each other depend on than predicted before, and skills this each other strengthen. Explanation this in line with the concept of "developmental cascades", which refers to the process of interaction and transactions Among moderate system developing and suggesting that adaptation in one domain will be spread to other domains (Masten & Cicchetti, 2010). Sarama et al. (2012) made argument similar, observing that set skills literacy and mathematics studied at the same time and that one suite skills strengthen the other. Study previously has focuses on investigating the academic domain, mathematics or which literacy is likely to be most useful for development academic and this has cause evidence to the contrary (Duncan et al., 2009; Clements et al., 2011; Morgan et al., 2015 Purpura et al., 2011; Sarama et al., 2012). Based on results study could interpreted that resilience in mathematics or literacy in between children in poverty could leads to growth positive in results academic within domains and across domains. Important things that need to be studied in the future must Keep going explore how resilience varied according to the domain of development and how resilience in one or several related domains with development next.

Method Study—Identification Variable Study

Variable main studied in study this is resilience literacy and numeracy. The resilience in question here is individuals age school that has Low Socioeconomic Status (SES), are in the bottom 20% category, however, show results study literacy numeration above criteria minimum competency. Variable resilience in study this Act as variable depends on or *dependent variable*. Temporary that variable free (*independent variable*) in study this is factors demographic. Following this is variable proposed demographics as variable free in study this:

- Education Level
- Type Sex
- School location by region of Indonesia (west central east)
- School location based on village-city (rural -urban)

Research Data

A total of 6,685,471 students comprised on various level education of 242,049 schools be the initial data study this. Temporary it's the data being analyzed is data from lower school coordination Ministry of Education and Culture, so the schools involved in analysis is formal school, that is schools that use identification such as elementary, middle, high school, and vocational school. School besides that no involved in research. Temporarily, the retrieved data through activity observation and interview school will be made as primary data for confirm, verify, or elaborate results analysis performed before.

Instrument

Instruments used in study this is survey about resilience literacy numeration. resilience literacy numeration defined in a manner operational through condition social economy student is at a percentage of 20 to lower however Fulfill minimum criterion of ability literacy and numeracy in accordance level his education.

Level	Literacy	Numeration
Elementary School	46.66	39.99
Junior High School	57.49	55.83
Senior High School	60.00	53.32
Vocational High School	60.00	53.32

Instrument second is guide observation and interview are not standardized (not analyzed in a manner psychometric). Instrument this arranged based on destination more research broad and for get more detailed information on the field. A number of observed aspects and interviews including the understanding of the teacher and the head school in thing resilience, honing activities resilience, constraints in enhancement resilience literacy and numeracy, as well action strategic in enhancement resilience literacy and numeracy. Temporary example he asked is as following:

- “Activities what just do in hone resilience literacy and numeracy community school (principal, teacher, staff, and student)? Mention!”,
- “How is the process done in increase resilience literacy and numeracy for community school? Mention!”,
- “Constraints what just what you are facing in enhancement resilience literacy and numeracy in schools? Say and how finish!”

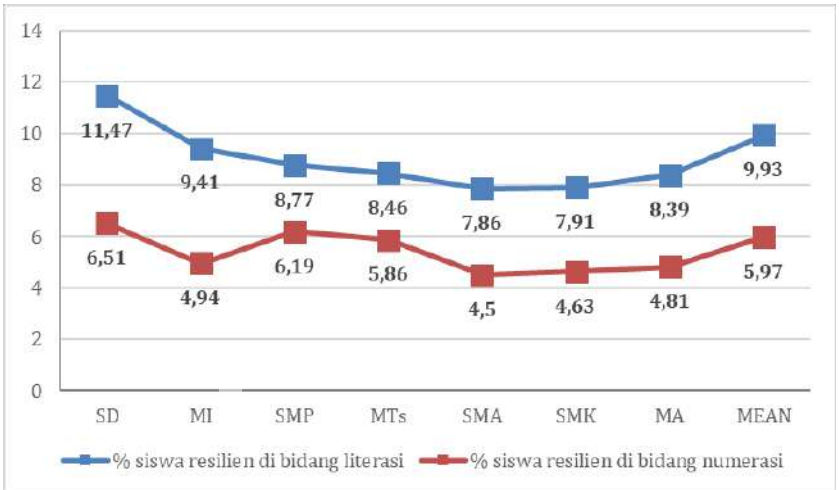
Procedure Data Analysis

Analysis on research this conducted in a manner gradually. Stage first is data screening and cleaning. At stage this is the data being analyzed reviewed based on a number of criteria so that all data is correct with a specified range. Stage second is extraction and filtering of data accordingly with need. Assessment national 2021 has data with amount very large variables and sizes very large sample. Not all variables and samples were worn in research. Stage this conducted with choose relevant variables for worn in study this, such as score literacy and numeracy, as well variables demographics (level education condition social economy and type gender). At stage third, data that has filtered then analyzed with statistics descriptive with compare average between group. Average comparisons are presented in tables and graphs to make it easy to draw a conclusion.

Results and Discussion—Resilience Field Students Literacy and Numeracy Nationally

Following this will explain the level of resilience students in all levels from elementary to high school in bottom 20% SES category. The national difference in literacy and numeracy resilience profiles in terms of educational level proves that building resilience is a dynamic social process experienced by students that is influenced by various internal and external factors while students are learning in their lives. This agree with results study Grotberg (1995) mentions that ability resilience no same among individual with individual others, this occur because resilience individual is largely determined by level age, grade development, intensity individual in face situation that is not fun , and how much big support in the formation.

Figure 1. Resilience Literacy and Numeracy Student nationally



Based on picture above could explained that *mean* level resilience student for field literacy in a manner national for all levels (SD, MI, SMP, MTs, SMA, SMK, MA) are in percentages as big as 9.93% and *mean* level resilience student for field numeration in a manner national for all level education (SD, MI, SMP, MTs, SMA, SMK, MA) is in the percentage as big as 5.97% this explain that level resilience student for field literacy more high / good compared with numeration , meaning that all level from level school base until school medium on has literate with literacy. This also proves that students who have level resilience high / good literacy, have considered dominate knowledge and skills for reading, writing, and literacy knowledge (science) and technology (digital), finance (financial), culture and citizenship. In thing this student not just has ability read, however students in all level education have ability analyze something reading, understanding the concept behind the article.

If it is described based on the level of student resilience in the field of literacy and numeracy nationally the highest is at the Elementary School (SD) level, namely for the numeric field of 6.51% and literacy of 11.47%, the MI level/level in the picture is explained by students sitting in MI bench has a literacy resilience rate of 9.41% and a numeracy resilience level below that is equal to 4.94%, the next level is junior high school, students who are in junior high school have a literacy resilience rate of 8.77% while the numeracy resilience level SMP students only have a percentage of 6.19%, as well as students who are in MTs, stating that the level of literacy resilience is better than

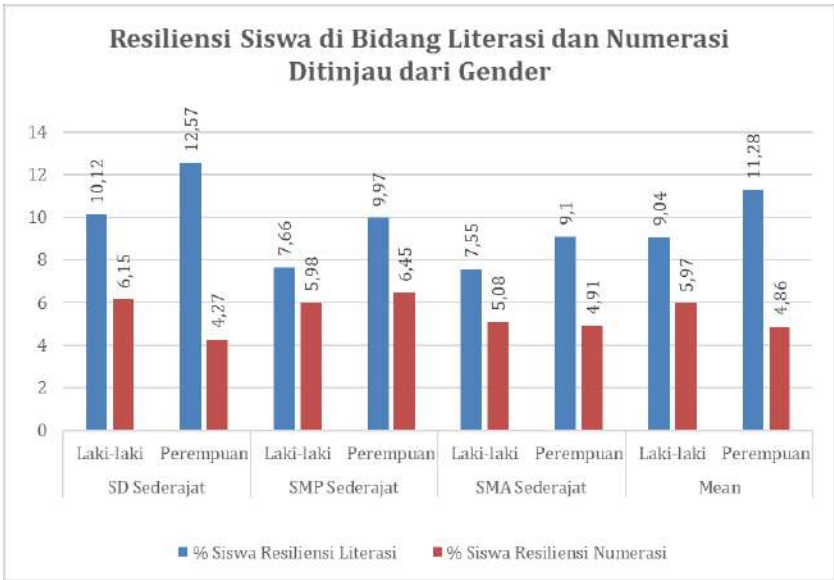
the level of numeracy resilience, namely 8.46% for the level of literacy resilience and 5.86% for numeracy resilience level. Meanwhile, students who are at the high school level (SMA) have a literacy resilience level of 7.86% and the numeracy resilience level is far below, with a percentage of 4.5%. This is also the case for students at the MA level, with a literacy resilience rate of 8.39% and for a numeracy resilience rate of 4.81%. At the SMK level, SMK students have a better level of literacy resilience than SMA, with a percentage of 7.91% and for the numeracy resilience level is a percentage of 4.63%. Based on the data and figures above, it is stated that the level of literacy resilience is higher/better than the level of numeracy resilience for all levels of education.

Further analysis, the level of literacy resilience is higher/better than numeracy, indicating that students at all these levels have the ability to understand, reflect, and evaluate the form and content of texts in everyday life and the ability to solve problems around them better. Of resilience to numeracy. Students with high levels of literacy resilience can be explained as students who have an understanding of texts and are capable of using, evaluating, and reflecting on the contents of texts to meet the demands of life. Indirectly students with high/good levels of literacy resilience could understand and integrate texts, assess the quality and credibility of texts, reflect on the form and content of texts, and find and solve problems in real everyday situations. In short, students who have high/good levels of literacy resilience can be said to be students with good abilities in terms of analyzing the form and content of a text. So, it can be concluded that students from elementary to high school who are in the lowest 20% SES category but show literacy or numeracy learning outcomes above the minimum competency criteria, are influenced by several factors. The factor referred to in this case is student resilience; this group can survive, rise, and adapt from limitations and powerlessness so that students can grow strong. In addition, social institutions become social capital for students to be able to adapt from difficulties. Technology that modifies the social structure in teaching/educating students to interact directly involves emotional skills, empathy, tolerance, and respect for others (Ruiz, 2017) making students more resilient to survive. The individual's ability to bounce back from adversity that occurs in its development is called resilience (Siebert, 2005). Resilience is the ability to be resilient, grow and adapt well in difficult situations (Helton & Smith, 2004), be able to recover from traumatic exposure, be able to overcome negative situations and adapt in very stressful conditions (Holaday & McPhearson, 1997).

Resilience Literacy and Numeracy Student Boy and Girl

Profile resilience literacy and numeracy describe dynamics social class determined by level education. This proves that resilience is not limited by type gender, will but also related with the learning process experienced by students different. For more he explained following this will explain level resilience student all levels (SD, MI, SMP, MTs, SMA, SMK, MA) for field literacy and numeracy based on type sex /gender.

Figure 2. Resilience Literacy and Numeracy Student reviewed of Gender



When viewed from gender, the level of literacy and numeracy resilience of students nationally from elementary to high school level is the highest for female students with a percentage of 11.28% for the level of literacy resilience and the level of resilience in the field of numeracy is at a percentage of 5, 97% for male students. If broken down by education level, for students at the elementary school level it is explained that the highest level of literacy and numeracy resilience is for female students, which is 12.57% literacy resilience level and 4, 27% for the numeracy resilience level, while elementary school students of the same gender have a literacy resilience rate of 10.12% and a literacy resilience rate of 6.15%.

At the junior high school (SMP) level, based on the data it can be explained that the highest level of resilience in the field of literacy and numeracy is for female junior high school students, namely 9.97% for the level of literacy resilience and 6.45% for the level of numeracy resilience while for junior high school students of the same sex, the percentage is only 7.66% for the literacy resilience level and 5.98% for the numeracy resilience level. Senior High School (SMA) level and equivalent, the data shows that students who sit at the same high school level have the highest level of literacy and numeracy resilience in female students, namely 9.1% for the level of literacy resilience and 4.91% for the level of resilience numeracy while male students only have a literacy level of 7.55% and a numeracy resilience rate of 5.058%.

Resilience level literacy and numeracy students in all level education based on existing data tendency domination of students manifold sex girl. this _ occur caused factor as revealed by Brougham et al. (2009) that in a manner general woman behave more put forward aspect affective in take risk, meanwhile man more put forward consideration cognitive in looked risk and danger as part live. Same thing with the opinion of Hamilton and Fagot (1988) mentions woman tend focuses on *coping* emotion, which means there is more focus big for resolve pressure emotional, meanwhile man more focus on *problem focused coping* it means tend use logic in a manner rational and final problem at the source. Rizkiana (2020) explains that resilience woman is higher than men's resilience,

because woman when sick, got pressure and situation difficult tend capable survive and adapt self with circumstances (Marlia, 2010). Pohl et al. (2005) found that woman tend show empathy more from men, and men tend more assertive than woman. In addition, women also have many roles and multitasking in life. Women have more control well and higher compared male, have regulation good emotions, optimism, ability identify problem as well as looked problem as a necessary life process live and have to accept with sincere, have a good sense of empathy, so personality woman formed in a manner natural with pattern enduring resilience in carrying out the adaptation process in a manner positive to situation threatening, challenging or sources other stress. Based on data analysis above could interpreted that in resilience gender perspective literacy and numeracy woman rated higher than compared to kind man. Findings this still need strengthened again with to do data analysis with count in a manner more detail on the effect of educational level.

Conclusion

Resilience has role important in build life society. Resilience is needed for repair quality life Indonesian society. Resilience is aspect important for form student more achievement. Resilience study multidimensional, so need comprehensive understanding for designing effort for build resilience in student school.

Resilience literacy and numeracy needed by students in face challenges in the digital age. The ability of students to build resilience literacy and numeracy are largely determined by the learning process that is adaptive and sustainable in accordance with needs students at their educational level. Stagnation in the learning process will take effect to resilience possessed by students.

Resilience characteristic dynamic so that level resilience student can decrease if not designed by the teacher in the learning process at school. Strengthening resilience needs cross- domain resilience in math at the time enter school base predicted Becomes base for develop skills more literacy height and resistance beginning in literacy push achievement performance in math. Required skills for study math and literacy more each other depend on than expected before, and skills this each other strengthen as a form of ‘developmental cascade’, requires interaction and transaction processes among system that requires power adaptation currently in one domain will be spread to another domain.

Resilience school is very important for support resilience literacy and numeracy students, proved that in a manner national there is tendency level education influence dynamics resilience. There is a trend that the higher level of education, the more decrease resilience literacy and numeracy caused because of the learning process not continuous by teachers at school. Dynamics resilience that becomes point important for teachers to keep going strengthen resilience literacy and numeracy with a proactive and creative way. From a gender perspective, yes tendency woman is higher thing, this prove that reinforcement process resilience in men need studied by approach priority culture importance values character that makes man more resilient in face life. In a manner general woman behave more put forward aspect affective in take risk, meanwhile man more put forward consideration cognitive in looked risk and danger as part live, so

woman tend focuses on *coping* emotion, which means there is more focus big for resolve pressure emotional, meanwhile man more focus on *problem focused coping* it means tend use logic in a manner rational and final problem at the source. Approaching culture for strengthening resilience from gender aspect is necessary design in the learning process.

References

Arthur, J., Davison, J., & Stow, W. (2014). *Social Literacy, Citizenship Education and National Curriculum.* London & New York: Routledge.

Bellous, J. E., & Clinton, J. M. (2016). *Learning Social Literacy*. Canada: High Pine Press.

Bernard, B. (1991). *Building Resilience in Children: Protective Factors in Families, Schools, and Communities: Portland*, OR: Northwest Regional Educational Laboratory. (ERIC Document Reproduction Service No. ED335781)

Brougham, R. R., Zail, C. M., Mendoza, C. M., & Miller, J. R. (2009). Stress, gender differences, and coping strategies among college students. Psychology Today: A Journal for Different Perspectives on Different Psychological Problems, 28(2), 85–97. https://doi.org/10.1007/s12144-009-9047-0

Claessens, A., Duncan, G., & Engel, M. (2009). Kindergarten skills and fifth-grade achievement: Evidence from the ECLS-K. An Overview of Educational Economics, 28, 415–427. http://dx.doi.org/10.1016/j.econedurev.2008.09.003

Clements, D. H., Sarama, J., Spitler, M. E., Lange, A. A., & Wolfe, C. B. (2011). Mathematics learned by young children in interventions by learning trajectory: Large-scale cluster randomized trial. Journal of Mathematics Education Research, 42, 127–166. http://dx.doi.org/10.5951/jresematheduc. 42.2.0127

Comer, J. P., Haynes, N. M., Joyner, E. T., & Ben-Avie, M. (1996). Rallying the whole village: Comer's process for reforming education. New York: Teachers' College Press.

Dwiningrum. (2008). Recovery psychology-social post-earthquake by teachers in Bantul Regency DIY. Educational Horizon, 2(2), 201-213.

Dwiningrum. (2011). Decentralization and participation Public in education. _ Yogyakarta: Student Library.

Dwiningrum. (2017). Social Capital and Resilience School for Disaster Mitigation Education in Yogyakarta Schools. *Journal Education Vol 1 No. 1*

Effendy, M. (2019). Literacy not only free blind characters. Retrieved from http://gln.kemdikbud.go.id/glnsite/literasi-tak-only-gratis-buta-aksara/

Freire, P., & Macedo, D. (1987). Literacy: Reading the word and the world. London: Routledge.

Goldstein, J. (2000). Causality and Emergence in Chaos and Complexity Theory. *W. Sulis and A. Combs (Eds.), Nonlinear Dynamics in Human Behavior (Studies of Nonlinear Phenomena in Life Sciences Volume 5)*

Grotberg, E. (1995). *A Guide to Promoting Resilience in Children: Strengthening the Human Spirit. Early Childhood Development: Practice and Reflection, v. 8.* The Hague-NO: Bernard Van Leer Foundation.

Halse, C. (2015a). (Re)think Asian literacy. In C. Halse (Ed.), Asia literate schooling in the Asian century (pp. 1–10). London: Taylor and Francis.

Hamilton, S., & Fagot, B. I. (1988). Chronic stress and coping styles: A comparison of male and female undergraduate students. *Journal of Personality and Social Psychology, 55(5), 819–823. https://doi.org/10.1037/0022-3514.55.5.819*

Helton, L. R., & Smith, M. K. (2004). *Mental health practice with Children and Adolescents.* Oxfordshire: Routledge

Henderson, A. T., & Mapp, K. L. (Eds.). (2002). M new wave evidence: The impact of school, family, and community connections on student achievement [Electronic version]. Austin, TX: National Center for Family and Community Connections with Schools, Southwest Educational Development Laboratory

Henderson, D. (2015). Globalization and national curriculum reform in Australia: The push for Asian literacy. In J. Zajda (Ed.), The second international handbook on globalization, education, and policy research (pp. 633–647). Dordrecht: Jumper.

Hill, H. (2021). What Happened to Poverty and Inequality in Indonesia in Half a Century? Asia Development Review, 38(1), 68–97. https://doi.org/10.1162/adev_a_00158

Holaday, M., & McPhearson. (1997). Endurance and Severe Burns. *Journal of Counseling &Amp; Development*

Culture. (2016a). Master design movement literacy school. Jakarta: Ministry of Education and Culture, Director General Basic Education

Lonigan, C. J., Purpura, D. J., Wilson, S. B., Walker, P. M., & Clancy-Menchetti, J. (2013). Evaluating the components of an emergent literacy intervention for preschool children at risk for reading difficulties. Journal of Experimental Child Psychology, 114, 111–130. http://dx.doi.org/10.1016/j.jecp.2012.08.010

Martin, J. P. (2018). Skills for the 21st century: Findings and policy lessons from the OECD survey of adult skills. In OECD Education Working Papers. https://doi.org/10.1787/96e69229-en

Marantika, S. (2018). The Effectiveness of Bibliotherapy Techniques for Increase Character Respect Students _ Class XI at SMK GPRI 4 Kediri. *Journal Guidance and counseling.*

Masten, USA. (2014). A global perspective on resilience in children and youth. Child Development, 85(1), 6–20. http://dx.doi.org/10.1111/cdev.12205

Masten, A. S., & Cicchetti, D. (2010). Developmental cascade. Developmental and Psychopathology, 22, 491–495.

McMillan, J. H., & Reed, D. F. (1994). Students at risk and resilience: Factors contributing to academic success. Clearing House, 67, 137-40.

Moore, J. (2013). *Resilience and Children and Adolescents at Risk.* United States of America: National Center for Homeless Education

Morgan, P. L., Farkas, G., Hillemeier, M. M., Hammer, C. S., & Maczuga, S. (2015). A 24-month-old child with a larger spoken vocabulary displayed better academic

and behavioral functioning at kindergarten. Child Development, 86, 1351–1370. http://dx.doi.org/10.1111/cdev.12398

Nyachae, T. M. (2019). Social justice literacy workshop for critical dialogue. Journal of Youth & Adult Literacy, 63(1),106–110.doi: 10.1002/jaal.977

OECD. (2000). Literacy in the information age: The final report of an international adult literacy survey. https://doi.org/10.1177/003172170008200118

OECD. (2009). Education at a Glance 2009: OECD Indicators. Paris.

OECD. (2011). Literacy for life: Further results from a survey of adult literacy and life skills. Retrieved from http://www.statcan.gc.ca

OECD. (2015). Country Notes: Key findings from PISA 2015 for the United States. At the OECD better policies for a better life. Retrieved from https://www.oecd.org/pisa/PISA-2015-United-States.pdf

OECD. (2016). PISA Results 2015 (Vol.II): Policies and practices for successful schools. https://doi.org/10.1787/9789264285521-en

OECD. (2017). Education at a Glance 2017: OECD Indicators. Retrieved from the OECD Publishing website: http://dx.doi.org/10.1787/eag-2017-en .

Pallas, A. M., Natriello, G., & McDill, E. L. (1989). The Changing Nature of Disadvantaged People: Current Dimensions and Future Trends. Educational Researcher, 18, 16-22.

Pohl, R. F., Bender, M., & Lachmann, G. (2005). Autobiographical Memory and Social Skills of Men and Women. *Applied Cognitive Psychology, 19(6), 745–759.* https://doi.org/10.1002/acp.1104

Pratiwi, P. H. (2021). Application Method Group Investigation (GI) Learning for Increase Competence Investigation Group On Students Class XI IPS 1 SMA Negeri 11 Yogyakarta. *Dimensia: Journal study Sociology DOI:10.21831/ Dimensia.v 10i1.41049*

Purpura, D. J., Hume, L. E., Sims, D. M., & Lonigan, C. J. (2011). Early literacy and early numeracy: Scores include early literacy skills in predicting numeracy development. Journal of Experimental Child Psychology, 110, 647–658. http://dx.doi.org/10.1016/j.jecp.2011.07.004

Rahman, H. (2017). Portrait growth economy, inequality and poverty in Indonesia in review economy political development. Journal Science and Culture, 40(55), 6305-6328. http://journal.unas.ac.id/ilmu-daya/article/view/409

Reivich, K., & Shatte, A. (1999). *The Resilience Factor: 7 Essential Skills for Overcoming Life's Inevitable Obstacles* Krishna S.Personal .

Rizkiana, D. (2020). resilience Academic Student Major Guidance Islamic Counseling (BKI) IAIN Tulungagung currently through Script in the middle Covid-19 Outbreak Review from type Gender. *Repository of UIN ONE Tulungagung*

Ruiz. (2017). Psychometric properties of the Cognitive Fusion Questionnaire in Colombia. *Revista Latinoamericana de Psicolagia*

Sarama, J., Lange, A. A., Clements, D. H., & Wolfe, C. B. (2012). The impact of the early mathematics curriculum on spoken language and literacy. Early Childhood Research Quarterly, 27, 489–502. http://dx.doi.org/10.1016/j.ecresq.2011.12 . 002 StataCorp. (2015)

Siebert, A. (2005). Endurance Gain. Portland: Practical Psychology Press.Center for Social Policy Studies (2005)

Wina, R., & Pont, B. (2017). Education policy implementation: Literature review and OECD education working paper framework No. 162. OECD Education Working Paper No. 162, December. https://doi.org/10.1787/fc467a64-en

Wilder, P., & Msseemmaa, D. (2019). Focusing disciplinary literacy on student awareness: A Tanzanian case study. Journal of Youth & Adult Literacy, 62(5),479–484, doi: 10.1002/jaal.880.

Yudiana, Eka Marlia. (2010). Penerapan Multi Metode dalam meningkatkan Motivasi dan prestasi belajar Mata Pelajaran Ekonomi siswa kelas VII SMPN 4 Malang. Undergraduate thesis, Universitas Islam Negeri Maulana Malik Ibrahim.

Assessment of the results of early childhood aqidah development at TASKA QF Bukit Kemiling Permai Bandar Lampung

Sovia Mas Ayu
Asmara Dewi
Okta Rijaya

[1]Universitas Islam Negeri (UIN) Raden Intan Lampung

https://doi.org/10.37517/978-1-74286-697-0-12

Dr. Sovia Mas Ayu, M.A. is a lecturer at UIN Raden Intan Lampung. This alumnus of the UNY PEP Doctoral Program have an interest in the field of research related to educational evaluation, including those related to the field of measurement and assessment. Apart from being a lecturer, he is also active in the management of Lampung HEPI (Indonesian Education Evaluation Association) and is part of the management of the Lampung Indonesian National Education Commission.

Abstract

TASKA QF (Quran First Kindergarten) at Bukit Kemiling Permai Bandar Lampung is one of the out-of-school educational institutions combining science and technology that is guided by the Qur'an and the sunnah of Rasulullah SAW. Fostering the foundation of *aqidah* from an early age is important because at that time the child's nature is still clean. If the faith is strong, he will become a figure of strong faith, both in his attitude and actions. Because they always feel under God's supervision, they can prevent children from doing bad deeds. Moreover, in the process of early childhood development, one phase is known, namely the "golden years," namely the ages between 0-5 years. This phase is important because, during this period, physical, mental, and personality growth occurs. This phase determines the growth of intelligence in all aspects, such as intellectual, personality, moral, and social. To determine the level of success in fostering early childhood faith, an assessment of the faith development that has occurred is required. The main issues discussed in this thesis include how to evaluate the fostering of God's love, how to assess the fostering of love for the Prophet and his family, and how to evaluate the fostering of love for the Qur'an at TASKA QF (Quran First Kindergarten) at Bukit Kemiling Permai Bandar Lampung. This type of research is classified as "field research," which is carried out using descriptive methods with a qualitative approach. Data collection techniques were carried out through interviews, observation, and documentation. The data were analyzed using Miles and Huberman's model analysis, then the credibility of the research was tested by triangulation. Aqidah development the love of Allah development at TASKA QF (Quran First Kindergarten) is carried out in 3 stages (learning *toyyibah* sentences, learning the names of Allah and their characteristics, and practicing prayer). The form of fostering the love of the prophet in TASKA QF (Quran First Kindergarten) through 4 stages (introduction to the prophet and his family, learning the characteristics of the prophet and his family, and learning and

practicing the prophet's daily manners, studying, and practicing the prophet's hadiths). Methods of fostering faith in the form of love for Allah and love for the Prophet and his family include exemplary methods, habituation, advice, supervision/attention, stories, and singing. Fostering a passion for the Qur'an at TASKA QF (Quran First Kindergarten) is carried out through several stages (recognizing the letters of the Quran/*hijaiyyah* letters, getting to know the basics of simple tajwid and *mahroj* knowledge, memorizing short surahs in the Qur'an along with names and their meanings). Methods of fostering faith in the form of love for the Qur'an include exemplary methods, habituation, stories, *sima'i, muroja'ah*, supervision/attention, advice, and singing. The results of this study show that the assessment of faith development in schools is based on attitudes, knowledge, and skills. Assessment is carried out by means of observation, oral tests, assignments, and portfolios.

Introduction

The education sector is facing a really serious test in this millennial era, namely the era from the 1980s to the 2000s. the era where all life processes go very swiftly and instantaneously. This convenience, as well as a trap that is consciously carried out in obedience, coupled with the lack of attention from the government in terms of regulation of public education, has caused an imbalance in the entire educational process in formal, non-formal, and informal institutions (Nuryana, 2019). Educational goals are very general and can only be achieved over a long period. It is not possible to achieve this goal in one or two years or only through one or two levels of school, but through lifelong education both inside and outside school.(Nurkholis, 2021) In principle, the educational goals of a community or nation are usually rooted in the philosophy of life and beliefs held by a nation. Likewise, the goals of Muslim life are influenced by the beliefs of Muslims themselves, which are based on the Al-Quran and Sunnah.(Nurhidayah, 2016) Islamic education should be built on the concept of Islam, to be able to form human beings who have superior intellect, rich in charity, and are elegant in morals and policies.

The achievement of educational goals has not been achieved properly. This is marked by the increasing moral degradation of the nation's children. Crimes involving children have increased every year. Based on data from the Indonesian Child Protection Commission (KPAI), there has been an increase in child crime cases in Indonesia. The predominant crimes are theft, motorcycle gangs, robberies, and street crimes. KPAI noted that the significant increase committed by children as perpetrators of crimes increased drastically. The data shows that in 2011, 695 cases were recorded, and in 2018, the number of children who became perpetrators of crimes reached 1434. (Mansyur et al., 2022)

In Lampung itself, many crimes are committed by children under the age of 13 and still in elementary school. Even this child acted with a sharp weapon and stabbed the victim to death. According to the police statement, the reason for killing the victim was just because he wanted to have money by selling the motorcycle he had stolen.(Gusnita, 2019) Another case that still occurs in 2019, is regarding the abuse of 6th-grade elementary school students in Bandar Lampung by their friends. The video of the abuse, which went viral on social media, shows elementary and junior high school students ganging up on a 13-year-old child on a plantation. Due to the beating, the victim experienced shortness of breath and bruises on his head and hands. The case of

persecution by the 13-year-old child also ended up in the police (Auliya, 2018). Cases that are no less exciting have also occurred recently, namely elementary school students impregnating high school students and giving birth to premature babies in Probolinggo, on April 15, 2019. (Hanafi, 1970) This could have happened because the formation of the faith was not strong at an early age. As stated by psychology figures, Suyadi and Maulidya Ulfah "Juvenile delinquency is not a new phenomenon from adolescence but is a continuation of the pattern of associational behavior that started from childhood. Since children aged 2-6 years, there is a possibility of being able to recognize those who later become naughty teenagers.(Azis, 1386)

True belief is the foundation of religion and the key to acceptance of the practice. This is as stipulated by Allah Ta'ala in His word:

It means: *"So whoever hopes to meet his Lord should do righteous deeds and not associate anything with Him in worshiping Him." (QS. Al Kahfi: 110)*

Also, the word of God, which means: *"And Abraham bequeathed that speech to his children", so did Ya'qub. (Ibrahim said): "O my children! Indeed, Allah has chosen this religion for you, so you should not die except in embracing the religion of Islam". (QS. Al-Baqarah : 132)*

Fostering the foundation of aqidah from an early age is important because at that time the child's nature is still clean. If the faith is strong, he will become a figure of strong faith, both in his attitude and actions. Because they feel under God's supervision, so they can minimize children from doing bad deeds. Moreover, in the process of early childhood development, one phase is known, namely the golden age, namely the age between 0-5 years. (Yani, 2013) This phase is important because during this period physical, mental, and personality growth occurs. This phase determines the growth of intelligence in all aspects such as intellectual, personality, moral and social. If this growth period goes well then the next phase of development will be good too.(Jannah & Maemonah, 2022) It is the responsibility of parents and teachers to raise their children morally. Faith is the key to happiness and security in this world and the hereafter. Prophets and apostles have also called on children to have a straight faith by instilling an understanding of faith from an early age. (Ulfa et al., 2018) So, the Islamic faith is something that must be taught first. Therefore, the development of faith at an early age is very necessary. The development of faith, noble character, and intelligence needs to be done as early as possible because it determines the faith, character, and intelligence of children in the future. TASKA QF (Quran First Kindergarten) is a pre-kindergarten and kindergarten education institution combined with science and technology which is guided by the Al-quran and sunnah of Rasulullah SAW. At one time when the researcher visited TASKA QF the researcher was impressed by the interaction between his students. During the lunch break in the yard of TASKA QF, all students in each class gather to enjoy their lunch. One of the students drank while standing, simultaneously the other students who saw said:

The hadith comes from Abu Hurairah Ra, and Rasulullah SAW recommends drinking while sitting, except for an excuse. because the Prophet Muhammad, Sallallahu 'alaihi wa Sallam, forbade drinking while standing. After hearing his friends, the student who was drinking while standing immediately sat down. Seeing this, the teacher then advises on eating manners. After observing the students, the researcher concluded that they were practicing a form of the creed of love for Rasulullah SAW by studying the Prophet's hadiths.(Kodina et al., 2016) The creed of a true Muslim always takes an example from

the deeds of the Prophet Muhammad, starting with small things like eating and drinking and not taking other figures as role models. Based on this background, researchers are interested in researching the Method of Guiding Faith in Early Childhood in the Quran First Kindergarten (TASKA QF).

Any learning process, of course, has achievement goals that must be met. Likewise with the development of early childhood beliefs at TASKA QF. To see how successful the coaching process is, of course, an assessment is needed. Assessment is the process of collecting and processing information to measure the achievement of student learning outcomes.(Peraturan Pemerintah, 2013) Following the assessment, a meaningful evaluation is performed. While evaluation is an assessment that has arrived at a decision based on government regulation number 19 of 2005 concerning the the National Education Standards, articles 1 no. 24 are assessed, which in turn will become input for information gathering (assessment). (Munthe, 2015)

Following the assessment, a meaningful evaluation is performed. While evaluation is an assessment that has arrived at a decision about something based on government regulation number 19 of 2005 concerning National Education Standards, articles 1 no. 24 are assessed, which in turn will become input for information gathering (assessment). (Fatah et al., 2018)

Following the assessment, a meaningful evaluation is performed. While evaluation is an assessment that has arrived at a decision about something based on government regulation number 19 of 2005 concerning National Education Standards, articles 1 no. 24 are assessed, which in turn will become input for information gathering (assessment). (Mustopa et al., 2021)

Assessment of the learning process uses an authentic assessment approach that assesses the readiness of students, processes, and learning outcomes as a whole. The integration of the assessment of the three components will describe the capacity, style, and learning gain of students, or even be able to produce an instructional effect and an accompaniment effect (nurturant effect) from learning. The results of authentic assessments can be used by educators to plan remedial, enrichment, or counseling services. In addition, the results of authentic assessments can be used as material to improve the learning process in accordance with improvement assessment standards. Evaluation of the learning process is carried out during the learning process using the tools: questionnaires, observations, anecdotal records and reflection. (Trisnamansyah, 2015)

Broadly speaking, the evaluation techniques used can be divided into two categories: non-test techniques and test techniques.

1. Non-test techniques. There are several non-test techniques, namely: (Magdalena et al., 2021)
    a. Multilevel scale (rating scale): This scale describes a numerical value on a consideration result. Usually, the numbers are laid out in stages, from low to high.
    b. Questionnaire (questionnaire): A questionnaire is often known as a questionnaire. A questionnaire is a list of questions that must be filled in by the person to be measured (respondents).

2. Test. A test is a tool or procedure that is systematic and objective and that obtains the desired data or information about a person in a way that can be said to be precise and fast. Disclosed by Amir Daien Indrakusuma (Riinawati, 2021). Muchtar Bukhori agreed, stating that a test is an experiment conducted to determine whether or not certain learning outcomes exist for a student or group of students (Lisnasari, 2022). From the opinion above, it can be seen that the test technique is a systematic assessment technique to find out the results of the learning process that has taken place. Thus, it is possible to see how far the students' understanding of the learning material presented has progressed. To measure students, the test is divided into three parts, namely: (Zamzania & Aristia, 2018)
    a. Diagnostic test: A good educator, of course, will feel happy if he can help his students so that they can achieve maximum progress according to their abilities. To find out whether the assistance provided was sufficient, an assessment was held. However, the information on the results of this assessment will be of no use if it is not used for consideration of further action.
    b. Formative test: Formative comes from the word "form," which is the basis of the term "formative evaluation," so formative evaluation is intended to determine the extent to which students have been formed after following a particular program. The benefits of formative tests for students include;(Khafid & Nurhayati, 2014)
        - Used to determine whether students have mastered the program material thoroughly.
        - Is reinforcement (reinforcement) beneficial to students? By knowing that the test has produced a high score as expected, students will feel that they are getting a "head nod" from the educator, and this is a sign that what they already have is good knowledge. This is correct, thus motivating students to study harder.
        - Repair effort.With the feedback (feedback) obtained after taking the test, students know their weaknesses, so they can improve on them.
        - As a diagnosis. The subject matter that is being studied by students is a series of knowledge, skills, or concepts, for that students, can find out which subjects are considered difficult.

    3. Summative Tests: Summative evaluations or summative tests are carried out after the end of the administration of a group of programs or a larger program. In my experience at school, formative tests can be equated with daily tests, while summative tests can be equated with general tests, which are usually carried out at the end of each quarter or end of the semester.

Several previous studies also discussed the same problem, including research conducted by Khairunnisa, with the title "Embedding Aqidah Values in Early Childhood in Raudhatul Athfal Banjarmasin Timur." Postgraduate Islamic Religious Education Study Program Thesis, IAIN Antasari Banjarmasin, 2016.(Khairunnisa, 2016) Research by Dyah Hesti Kayuntami, Istania Widayati Hidayati in their journal entitled "Development of Religious Values in Early Childhood in Kindergarten." Islamic Religious Education Study Program, Faculty of Islamic Religion, University of Muhammadiyah Magelang Journal of University Research Colloquium 2018, Muhammadiyah University, Purwokerto. The results showed that: 1) The development of religious values in early childhood at Pertiwi Kindergarten in

Magelang City includes 4 activities, namely, opening activities, core activities, recalling activities, and closing activities. The material presented for fostering religious values includes education in aqidah, worship, and morals. 2) The methods used in the process of cultivating religious values at an early age are the playing method, the storytelling method, the singing method, the field trip method, the demonstration method, the habituation method, the dialogue method, and the exemplary method. 3) Fostering religious values in early childhood has an impact on developing children's religion, changing children's behavior, and channeling children's talents.(Khan, 2021) Research by Hermanto, with the title "Fostering Islamic Values in Forming Politeness at the Rhaudatul Athfal Nurul Bahra Foundation, Bone Regency."

The research method used by researchers is a qualitative method with a phenomenological approach, which involves seeking data through interviews, observation, and documentation of several data sources, after which data analysis from Miles and Huberman is performed along with data triangulation. The findings indicate that the best way to instill Islamic values in children's behavior is to conceptualize playing while learning, telling exemplary stories, watching Islamic films while studying, and providing examples regularly so that it becomes a habit in children. While the teacher faces obstacles, the parents pay less attention to their children due to their various activities.(Syaepul Manan, 2017) Titin Nurhidayah conducted research titled "Efforts to Foster 5-6-Year-Old Group Moral Values" at PAUD Karima Salama Potrojayan. This research uses a qualitative approach with a descriptive research type. The research data was collected through interviews, observations, and documents. The collected data were analyzed descriptively and qualitatively using an interactive analysis model. The validity of the research data was re-tested by using extended participation, observation, persistence, and triangulation. The results of the study show that there are three moral values instilled in PAUD Karima Salama: morals towards God, morals towards fellow human beings, and morals towards the environment. the process of implementing the development of moral values through habituation and storytelling. While the supporting factors of the environment include parents, schools, the community, professional educators, learning media, and school infrastructure, there are also inhibiting factors, namely the problem of the timing of the implementation of learning, the lack of awareness among parents about repeating learning at home.(Nurhidayah, 2016)

The difference between the previous studies presented and this research lies in the location and field of study. The research that will be carried out by researchers is located at the Qur'an First Kindergarten (TASKA QF) in Bandar Lampung. Another difference that appears from the subject matter Whereas research conducted by researchers will only focus on methods of fostering faith, such as the method of fostering the love of Allah, the method of loving the Prophet and his family, and the method of loving the Qur'an. In terms of the place of research, this includes field research, namely research that is directly carried out in the field or on respondents.(Fadli, 2021) The data collected from the field was directly related to the object in question, namely TASKA QF (Qur'an First Kindergarten). This study uses qualitative methods, namely scientific research that is not obtained from the process of statistical calculations.(Nurdiani, 2014)

The data collection techniques used are interviews, observations, and the utilization of documents. This research is used to answer questions about how things are (event phenomena) and report as they are. The researcher describes the results of the research in the form of describing facts that occur in the field regarding various matters relating to methods of developing faith in early childhood Qur'an First Kindergarten (TASKA QF).

The main data collected is related to the assessment of the results of fostering aqidah in early childhood, which includes the assessment of the development of monotheism, the assessment of fostering the love of Rasulullah and his family, as well as the assessment of fostering the love of the Qur'an. Sources of data in this study were school principals, teachers, and TASKA QF documents related to the assessment of early childhood faith development at the school.

The data analysis procedure was carried out in several stages, namely data reduction, data presentation, and conclusion. While checking the validity of the data is done by source triangulation and technique triangulation.

Content

Aqidah development in early childhood at TASKA QF (Qur'an First Kindergarten), three main things are important to discuss and are the focus of research in this study, namely: the method of fostering monotheism, the method of fostering the love of the Messenger of Allah and his family, and the method of teaching the Qur'an.

The method of cultivating the love of Allah

Fostering the pillars of aqidah, in the form of cultivating Allah's love, is accomplished in three stages. The first stage, by studying the sentences toyyibah, (laa illa ha illallah, masyaa Allah, subhannallah, alhamdulilah, Allahu Akbar, astagfirullahaladzim). The second stage is learning the names of Allah and their attributes (Allah is all-seeing, Allah is all-hearing, Allah loves beauty, and so on). In stages one and two, this is learned by children in grades K2, who are 2 years old, and K3, who are 3 years old. Because children have different developments, the material must be chosen according to the abilities of the students.

This is by Umm Ihsan's words that talqin in the toyyibah sentence can be started from the age of 2 years.(Julianti et al., 2014) The third stage is practicing prayer in the form of readings and prayer movements. At this stage, children have learned about Thoyyibah sentences, and know Allah and His attributes; for this reason, they are taught to practice prayer. The prayer performed at school is the Duha prayer. Children are asked to pray by amplifying their reading. Methods of fostering faith in the form of love for Allah include exemplary methods, habituation methods, advice methods, supervision/attention methods, dialogue methods, stories/stories, and singing methods.

The introduction of Allah and the attributes of Allah can be obtained through the singing method. The teaching of monotheism is also carried out in teaching and learning activities using the question-and-answer method. This dialogue method is used by asking simple things related to God and his characteristics. like associating everything that happens with God. For example, when it rains, the teacher asks, "Who created the rain?" Then explain the wisdom behind the rain, etc. In applying this method, the teacher can adjust according to the reasoning of the students. Because early childhood religious development is influenced by age. By the characteristics of early childhood, they like things that are imaginative, fantastic, and curious in many ways, so naturally, the teacher facilitates stories according to the characteristics of early childhood.(Juwairiah, 2017)

The story/story method is also used to cultivate God's love by using 2-dimensional media, namely books, and full-color pictures so that children are more interested in listening. For example, the teacher shows a picture of a fire, then asks who asked about the fire. Then he talks about the fires of hell, all the torments, animals, and food in it, and that all hell was created by Allah for sinners. Likewise with stories of heaven, earth, etc. Teachers are also sometimes able to tell self-taught stories and then relate these stories to God. An introduction to the nature of God is also obtained through the material taught, as in a beautiful hadith "innallaha jamilun wa yuhibbul jamal" which means that God is beautiful and loves beauty. The hadith explains the nature of God, who loves beauty and applied so that children like cleanliness, tidiness, and beauty. An-Tawhid coaching among children aims to teach simply. It is also done by getting used to doing good, setting a good example, teaching discipline, motivating, and rewarding psychologically. This is considered sufficient to support the development of monotheism.(Ulfa et al., 2018)

The method of loving the Prophet and his family

The pillar of faith development is the love of the prophet. Loving the prophet means following his teachings and obeying them. The more you love something, the more enthusiastic you are about obeying it. In Taska QF, there are four stages to cultivating prophetic love. The first stage of the introduction of the prophet and his family. Second, study the characters of the prophet and his family. Third, learn and practice the Prophet's daily manners. Fourth, study and practice the hadiths of the prophet. These 4 stages can be applied to all classes: K2 (2 years), K3 (3 years), K4 (4 years), and K5 (5 years). Methods of fostering faith in the form of love for the Prophet and his family include exemplary methods, habituation methods, advice methods, supervision/attention methods, dialogue methods, story methods, and singing methods.

The Prophet's love development is carried out by studying the hadiths of the Prophet, following his sunnahs, and applying them in daily life with the method of habituation. This statement was reinforced by Sayyid Sabiq who stated that knowledge is obtained by learning, while manners and main character are obtained through practice and habituation.(Ashiong Parhehean Munthe & Halim, 2019) In its application, it requires a role model, and this is where the teacher plays an important role in providing examples of good morals. While the introduction of the Prophet, his family, nature, and character are done by telling stories using dialogue and singing methods.

The method of cultivating a love of the Qur'an

The pillars of fostering the faith of the Qur'an are symbols to strengthen faith, which should be instilled from an early age. A sense of love for the Qur’an must be developed before children are taught to memorize it. Without a sense of love, the Qur’an will not be able to produce goodness for its readers. The process of fostering a love for the Qur’an in TASKA QF is carried out in several stages. In the first stage, get to know the letters of the Quran/hijaiyyah alphabet. The second stage, knowing the basic science of recitation and mahroj. For stages one and two this can be applied to grades K2 (2 years) and K3 (3 years) by adjusting the children's different language abilities. The letters he memorized include Al-Fatihah, An-nas, Al-Ikhlas, Al-kautsar, An-nashr with the sima'i method where the teacher reads repeatedly both using audio and orally. Of course, the sima'i method cannot be forced on children aged 2-3 years because educators must consider children's different language skills. The third stage is memorizing short letters in the Qur’an along with their names and meanings. This stage is implemented in grades

K4 and K5, which are 4-5 years old. Where they already have good language skills, so they can memorize short letters and the meaning of the letters. Methods of fostering faith in the form of love for the Quran include exemplary methods, habituation methods, story methods, sima'i methods, muroja'ah methods, supervision/attention methods, advice methods, and singing methods.

Fostering a love for the Qur'an is carried out in stages, starting with the introduction of hijaiyyah letters and introducing its makhraj by singing and playing through the media of cards/flash cards. Meanwhile, to find out the recitation, the length, and a short reading of the Al-Quran is done by using the medium of the IQRA' book. The method of memorizing the Qur'an for early childhood at TASKA, uses repeating the reading from verse to verse repeatedly, which is called the sima'i method. Then, memorization is maintained with the muraja'ah method with the help of teachers/parents and media in the form of headphones. The spirit of memorizing is also supported by the advice in the form of learned hadiths such as "tholabul 'ilmi fariidhotun 'ala kulli Muslim" which means studying is mandatory for every Muslim. And "khoirukum man ta'alamal Qur'ana wa 'alamahu" means the best of you are those who learn the Qur'an and teach it. The story/story in the Qur'an is also very fond of children, such as the story of Abu Lahab in Surah Al Lahab, stories of prophets such as Prophet Noah, the story of Musa and Pharaoh, Prophet Yusuf, Prophet Yunus, and other prophets. In addition to the stories of the Prophet, teachers also tell Islamic stories to teach Islamic characters and noble morals. The exemplary method (Al-ahzab: 21), the habituation method (Ruum: 30), the advice method (Luqman: 13), the supervision/attention method (At-tahrim: 6), the dialogue method, the story method (Yusuf: 111) are methods which are based on the Qur'an and exemplified by the prophet Muhammad SAW. For this reason, as educators, we should use a method that has a clear basis and was taught by the prophet Muhammad. These methods are certainly interrelated and mutually support each other. If educators can apply these methods in fostering faith. Then a generation of strong faith will be formed so that moral degradation, which is a big problem in this era will be easily overcome. Apart from that, faith can provide other positive aspects, such as an attitude of optimism in life, a disciplined attitude due to obedience to Allah and even natural law (sunnatullah) with full awareness and responsibility, as well as a high work ethic.(Adam & Hakim, 2017)

To form children who have strong beliefs, noble morals, and straight shari'at we need two important factors, namely, a noble education and a conducive environment. This is stated by Nashih Ulwan regarding noble Islamic education. Rasulullah stated in a hadith that "there is no gift given by a father to his child better than a good education" (HR. Tirmidhi), and the conducive environmental factor is also stated in a hadith that "every baby born with a natural state, its parents who make it Jewish, Christian, or Magian." In another hadith, it is also explained that "a person depends on the religion of his friend, so one of you should pay attention to who that person is friends with" (HR. Tirmidhi) From the hadiths, it can be understood that if a child is educated by a pious and pious person, then they will teach the principles of faith and Islam so that children will grow up with a strong belief in faith and Islam. In addition, friends also influence children's beliefs. If a child makes friends with a pious friend, then it can be certain that he will also be pious, because the child will imitate the habits of the friend.

Taska QF, as a place for children to socialize outside of the family, is also a means for children to interact with friends of the same age, not wasting the opportunity during the golden age of children to instill faith, because the religious soul of children is primarily

influenced by adults. They tend to accept the teachings of adults even if they are not yet fully aware of the benefits of these teachings.(Nandya, 2010) For that, adults/educators should use methods that are based on the instructions of the Qur'an and the guidance of the Prophet. TASKA QF combines the method with the contemporary singing method. The singing method is very popular with early-age children, especially if it is followed by simple body movements. By singing, children are free to express themselves, develop aspects of their language, and eliminate boredom in and out of class. Especially if the lyrics of the song that is sung contain a deep meaning about faith, love of God, love of the Prophet, and love of the Qur'an, it can be confirmed that it will be very influential on the religious soul of the child. By singing unconsciously, children can memorize material about faith. In this regard, Imam Ghazali said that in instilling faith, the first step is to give memorization, then understanding, then belief (I'tiqad) confidence, and justification.(Kholiq, 2021)

Another method used in TASKA QF for fostering a love of the Qur'an is the sima'i and muroja'ah methods. These two methods, of course, had been taught by Rasulullah SAW to his friends by Rasulullah reading, the friends listening, and then Rasulullah SAW listening to the friends' memorization. Jibril also used the same method in teaching the Qur'an to Rasulullah, even though Rasulullah was an Ummi, who could not read or write. As a result, this method is excellent for children. Because children who memorize when they are young are much more attached than adults who memorize when they are older. Busying oneself with reading the Qur'an and its interpretations, reading hadiths and their meanings, and engaging oneself with worship is part of the way of developing aqidah.(Nasution & Casmini, 2020)

The teacher/educator plays a very important role in fostering faith in schools because the teacher is a role model who imparts knowledge of the faith, the best figure in the eyes of children, whose behavior, and manners, whether they realize it or not, will be imitated by them. Even the form of words, deeds,, and actions will always be embedded in the child's personality.(Jannah, 2015) It is educators who introduce good and bad values (indoctrination) to children to form habits so that faith is firmly ingrained. The teacher also selects and delivers materials and methods about faith, which must also be adjusted to the stages of child development so that what is conveyed to children is well received. Educators should not ignore faith development methods such as the exemplary method, the habituation method, the advice method, the supervision/attention method, the advice method, the story method because in addition to these methods sourced from the Qur'an and exemplified by the Prophet these methods are also used as a transfer of value to students so that one day they will become a generation of Muslims who have strong beliefs and good morals. In addition, early childhood education can improve performance and increase work productivity in adulthood.(Azis, 2016)

Conclusion

After reviewing and analyzing in depth the method of fostering faith in early childhood. The teacher/educator plays a very important role in fostering faith in schools because the teacher is a role model who imparts knowledge of the faith, the best figure in the eyes of children, whose behavior and manners, whether they realize it or not, will be imitated by them. It is educators who repeatedly introduce good and bad values (indoctrination) to children to form habits so that faith is firmly ingrained.

1. Fostering the pillars of faith, in the form of fostering a love for Allah, is carried out in 3 stages. The first stage, by studying toyyibah sentences, (laa illa ha illAllah, masyaa Allah, subhannAllah, alhamdulilah, Allah u akbar, astagfirullahaladzim). The second stage is by studying the names of Allah and their attributes (Allah is all-seeing, Allah is all-hearing, Allah loves beauty, etc.). The third stage is by practicing prayer in the form of readings and prayer movements. Methods of fostering faith in the form of love for Allah include exemplary methods, habituation, advice, supervision/attention, dialogue, stories/stories, and singing. Assessment is done by observing and oral tests.
2. The pillars of faith development are the love of the prophet and his family. Loving the prophet means following his teachings and obeying them. The more you love something, the more enthusiastic you are in obeying it. In Taska QF, there are four stages to cultivating prophetic love. The first stage of the introduction of the prophet and his family. Second, study the characters of the prophet and his family. Third, learn and practice the Prophet's daily manners. Fourth, study and practice the hadiths of the prophet. Methods of fostering faith in the form of love for the Prophet and his family include exemplary methods, habituation, advice, supervision/attention, dialogue, stories, and singing. Assessment is done by observing and giving oral tests.
3. The pillars of fostering the faith of love in the Al-Quran are symbols to strengthen faith, which should be instilled from an early age. The process of fostering a love for the Qur'an in TASKA QF is carried out in several stages. In the first stage, get to know the letters of the Qur'an / hijaiyyah letters. The second stage, knowing the basic science of recitation and mahroj. The third stage is memorizing short letters in the Qur'an along with their names and meanings. Methods of fostering faith in the form of love for the Qur'an include exemplary methods, habituation, stories, sima'i, muroja'ah, supervision/attention, advice, and singing. Assessment is done by observing and giving oral tests.

Based on this research, it is suggested to

1. Researchers are expected to be able to continue to develop their research related to the assessment of early childhood faith development, especially related to the instrument for carrying out this assessment.
2. Principals are expected to improve facilities to support the learning process of faith development activities and strive for teacher abilities by increasing broader knowledge with workshops or training.
3. Teachers are expected to improve their abilities and skills in teaching so that the methods of fostering faith are more varied and effective for early childhood at TASKA QF (Quran First Kindergarten) Bukit Kemiling Permai Bandar Lampung.

References

Adam, M. R., & Hakim, S. (2017). MENELUSURI JALAN SUFI (Kajian Kitab 'Umdat al-Muhtajîn ilâ Suluk Maslak al- Mufradîn Karya 'Abd al-Râuf al-Sinkîlî). *Kordinat: Jurnal Komunikasi Antar Perguruan Tinggi Agama Islam*, *16*(2), 367–388. https://doi.org/10.15408/kordinat.v16i2.6447

Auliya, R. U. (2018). *Kenakalan* Orangtua Penyebab Kenakalan Remaja. *Jurnal Al-Taujih : Bingkai Bimbingan Dan Konseling Islami*, *4*(2), 92–103.

https://www.ejournal.uinib.ac.id/jurnal/index.php/attaujih/article/view/505
Azis, S. (*2016*). *Pentingnya Pendidikan Agama Islam bagi Orang Tua dalam* Mendidik *Anak*. 12–28.
Fadli, M. R. (2021). Memahami desain metode penelitian kualitatif. *Humanika, Kajian Ilmiah Mata Kuliah Umum*, *21*(1), 33–54. https://doi.org/10.21831/hum.v21i1.
Fatah, P. A., Kusumah, I. H., & Komaro, M. (2018). Kesesuaian Antara Penilaian Belajar Oleh Guru Produktif Dengan Standar Penilaian Pendidikan Pada Mata Pelajaran Teknologi Dasar Otomotif. *Journal of Mechanical Engineering* Education, *5*(1), 71. https://doi.org/10.17509/jmee.v5i1.12622
Gusnita, C. (2019). Fenomena Anak Sebagai Pelaku Persekusi Di Media Sosial. Jurnal *Muara Ilmu Sosial, Humaniora, Dan Seni*, *3*(1), 11. https://doi.org/10.24912/jmishumsen.v3i1.3455
Hanafi, I. (1970). Perkembangan Manusia Dalam Tinjauan Psikologi dan Al-Qur'an. *IQ (Ilmu Al-Qur'an): Jurnal Pendidikan Islam*, *1*(01), 84–99. https://doi.org/10.37542/iq.v1i01.7
Jannah, M. (2015). Tugas-Tugas Perkembangan Pada Usia Kanak-Kanak. *Gender Equality: International Journal of Child and Gender Studies*, *1*(2), 89–91.
Jannah, M., & Maemonah, M. (2022). Implementasi Bimbingan Spiritual pada Anak Usia Dini Melalui Kisah Uwais Al-Qarni di TPA Safinatussafa, Aceh Selatan, Indonesia. *DAYAH: Journal of Islamic Education*, *5*(1), 134. https://doi.org/10.22373/jie.v5i1.10139
Julianti, K., Syukri, M., & Astuti, I. (2014). *Peningkatan pengucapan kalimat* thayyibah *melalui trategi modeling the way pada anak usia 4-5 tahun*. 2.
Juwairiah. (2017). Membentuk Karakter Anak Usia Dini dengan Mengenalkan Cerita Rakyat Dari Aceh. *Bunayya: Jurnal Pendidikan Anak*, *3*(1), 1–18.
Khafid, M., & Nurhayati, F. (2014). Pengaruh pemberian tes formatif terhadap peningkatan hasil belajar pendidikan kesehatan pada siswa kelas VII SMP Negeri 2 Krian. *Jurnal Pendidikan Olahraga Dan Kesehatan*, *2*(1), 119–124. http://ejournal.unesa.ac.id/index.php/jurnal-pendidikan-jasmani/issue/archive
Khairunnisa. (2016). *Penanaman Nilai-nilai Aqidah pada Anak Usia Dini di Raudhatul Athfal Banjarmasin Timur*. IAIN Antasari Banjarmasin.
Khan, R. I. (2021). Urgensi Pendidikan Untuk Anak Usia Dini Di Indonesia. *Jurnal Al–Hikmah*, *9*(1), 42–48.
Kholiq, A. (2021). Pendidikan Agama Islam Perspektif Al-Ghazali dalam Kitab Ihya'Ulumuddin. *Intizam, Jurnal Manajemen Pendidikan …*, 42–62. http://ejournal.staida-krempyang.ac.id/index.php/intizam/article/view/299
Kodina, Y. E., Rama, B., Getteng, A. R., & Said, N. (2016). Hakikat Materi Akidah Perspektif Pendidikan Agama Islam Dalam Kurikulum Sekolah Dasar Kelas V. *Jurnal Diskursus Islam*, *04*(03), 523–529.
Lisnasari, S. F. (2022). Pengembangan Instrumen Tes Berbasis Buku Penilaian Autentik (Bupena). *Jurnal Penelitian Pendidikan Dan Sains*, *1*(1).
Magdalena, I., Ismawati, A., & Amelia, S. A. (2021). Penggunaan Evaluasi Nontes dan Kesulitannya di SD N Gempol Sari. *Jurnal Pendidikan Dan Ilmu Sosial*, *3*(2), 187–199. https://ejournal.stitpn.ac.id/index.php/pensa/article/view/1299/907
Mansyur, M., Mas, M., & Renggong, R. (2022). Analisis Penerapan Sanksi Pidana Terhadap Pelaku Pencurian Handphone Di Wilayah Polres Pinrang. Indonesian *Journal of Legality of Law*, *4*(2), 153–166. https://doi.org/10.35965/ijlf.v4i2.1365
Munthe, A. P. (2015). Pentingya Evaluasi Program Di Institusi Pendidikan: Sebuah Pengantar, Pengertian, Tujuan dan Manfaat. *Scholaria : Jurnal Pendidikan Dan*

Kebudayaan, *5*(2), 1. https://doi.org/10.24246/j.scholaria.2015.v5.i2.p1-14
Munthe, A. P., & Halim, D. (2019). Pendidikan Karakter Bagi Anak Usia Dini Melalui Buku Cerita Bergambar. *Satya Widya*, *35*(2), 98–111. https://doi.org/10.24246/j.sw.2019.v35.i2.p98-111
Mustopa, A., Jasim, J., Basri, H., & Barlian, U. C. (2021). Analisis Standar Penilaian Pendidikan. *Jurnal Manajemen Pendidikan*, *9*(1), 24–29. https://doi.org/10.33751/jmp.v9i1.3364
Nandya, A. (2010). Etika Murid Terhadap Guru (Analisis Kitab Ta'lim Muta'allim Karangan Syaikh Az-Zarnuji). *Mudarissa*, *2*(1), 163–328.
Nasution, U., & Casmini, C. (2020). Integrasi Pemikiran Imam Al-Ghazali & Ivan Pavlov Dalam Membentuk Prilaku Peserta Didik. *INSANIA : Jurnal Pemikiran Alternatif* Kependidikan, *25*(1), 103–113. https://doi.org/10.24090/insania.v25i1.3651
Nurdiani, N. (2014). Teknik Sampling Snowball dalam Penelitian Lapangan. ComTech*: Computer, Mathematics and Engineering Applications*, *5*(2), 1110. https://doi.org/10.21512/comtech.v5i2.2427
Nurhidayah, T. (2016). Pelaksanaan penanaman nilai-nilai akhlak kelompok 5-6 tahun di paud karima salama potrojayan. *Jurnal Pendidikan Anak Usia Dini*, *5*(2), 182–193.
Nurkholis. (2021). Idealitas Pendidikan Anak dalam Islam. *Jurnal Kependidikan*, *7*(2), 395–407.
Nuryana, Z. (2019). *Kurikulum 2013 dan Masa Depan Pendidikan Agama Islam di* Indonesia. Komojoyo Press. https://doi.org/10.31219/osf.io/4j6ur
Peraturan Pemerintah. (2005). Peraturan Pemerintah Republik Indonesia Nomor 19 Tahun 2005 Tentang Standar Nasional Pendidikan.
Peraturan Pemerintah. (2013). Peraturan Pemerintah Republik Indonesia Nomor 32 Tahun 2013 Tentang Standar Nasional Pendidikan. *Sekretariat Negara*, *2*(32), 148–164.
Riinawati. (2021). Pengantar *Evaluasi Pendidikan*.
Syaepul Manan. (2017). Pembinaan Akhlak Mulia Melalui Keteladanan dan Pembiasaan. *Jurnal Pendidikan Agama Islam-Ta'lim*, *XV*(2), 1.
Trisnamansyah, S. (2015). Evaluasi Pembelajaran. *Syria Studies*, *7*(1), 37–72. https://www.researchgate.net/publication/269107473_What_is_governance/link/548173090cf22525dcb61443/download%0Ahttp://www.econ.upf.edu/~reynal/Civil wars_12December2010.pdf%0Ahttps://think-asia.org/handle/11540/8282%0Ahttps://www.jstor.org/stable/41857625
Ulfa, H. F., Abdussalam, A., & Surahman, C. (2018). Metode Pendidikan Tauhid Dalam Kisah Ibrāhīm As. Dan Implikasinya Terhadap Pembelajaran Pai Di Sekolah. *TARBAWY : Indonesian Journal of Islamic Education*, *4*(2), 80. https://doi.org/10.17509/t.v4i2.8599
Yani, A. (2013). Pendidikan Agama Pada Anak Oleh Orang Tua : Tinjauan Psikologi Islam. *Jurnal Ilmu Agama UIN Raden Fatah*, *14*(1), 33–44.
Zamzania, A. W. H., & Aristia, R. (2018). Jenis - Jenis Instrumen dalam Evaluasi Pembelajaran. *Universitas Muhammadiyah Sidoarjo*, 1–13.

Is it necessary to use two-way items?: Testing the effects of using favourable and unfavourbale Items

Sukaesi Marianti
Dera Kariema
Puri Intan Larasati
Dhia Ulfah Purwati
Nur Hasanah

Universitas Brawijaya

https://doi.org/10.37517/978-1-74286-697-0-13

SUKAESI MARIANTI is a researcher at the Department of Psychology, University of Brawijaya. Her research interests are in Response Time Modeling, and Item response Theory, especially in educational and psychological research.

Abstract

The topic of ethnicity is a topic that is often discussed in research, especially psychology research. Considering that Indonesia is a country with many ethnicities, therefore, ethnicity is a very important and relevant topic to be discussed in many studies. In this regard, robust tests measuring ethnicity are needed to collect data and further the data can be validly interpreted. The purpose of this study was to psychometrically evaluate the effects of the method of using two-way items (favourable and unfavourable) on two scales, Ethnic Experience and Ethnic Identity. This evaluation will lead to an evaluation of possible new factors that may arise due to the use of two-way items. The approach used is the Multritrait Multimethod (MTMM), with Confirmatory Factor Analysis (CFA), to test the four alternative models used. The results of this study indicate that two (CTCM, and CTUM) of the four proposed models have a satisfactory fit. This proves that there is indeed an effect from the method of using two-way items (favourable and unfavourable).

Introduction

The use of favourable and unfavourable items is often found in the test development, which are useful for the research data collection. Some argue that the use of favourable and unfavourable items is useful to avoid possible bias in responses, based on the assumption that if the items measure the same construct, then even using favourable items and unfavourable (bidirectional) items, it will still produce an equivalent response (Marsh, 1996; Nunnally, 1978).

However, not all conditions indicate that the use of unfavourable items is the best option, according to Van Roy et al. (2008) there are several problems that can arise from the use of two-way items (favourable and unfavourable), where favourable and unfavourable items in groups form their own factors. According to the analysis of the grouping factors, it shows that there are latent factors underlying the group of items. This indirectly shows that there is a bias from the use of favourable and unfavourable items, or what is often referred to as the method effect.

Several studies, such as those conducted by Van de Looij-Jansen et al. (2011) and the research of Rodebaugh et al. (2007) found the effect of methods from the use of favorabble item and unfavourable item so that the items measuring constructs differs from what should have been measured. This grouping will certainly bring its own problems, where the problem not only affects the validity of the construct, but also affects the interpretation of the test results. Moreover, it will also produce a biased correlation between the scores produced from biased instruments (DiStefano & Motl, 2006).

Because ideally the items should be grouped based on the underlying construct not because of the method being used. Research conducted by Lindwall et al. (2012) shows that the method effect is related to the use of favourable and unfavourable sentences. According to Marsh, the use of items with different measuring directions should still measure the same concept and meaning, except that the sentence form is different, such as positive or negative sentences (DiStefano & Motl, 2006). Research conducted by Conroy (2002); DiStefano and Motl (2002) show that if a method effect occurs, it will threaten validity.

Analysis of the effect of the method can be carried out using Confirmatory Factor Analysis (CFA) with the Multritrait Multimethod (MTMM) approach. There are two concepts of the MTMM model used for method effect analysis, such as the correlated traits-correlated unique (CTCU) model, correlated traits-correlated methods (CTCM) or by combining the two models (DiStefano & Motl, 2006). The two concepts of the approach are a development of the concept of MTMM analysis in CFA. According to Brown (2015), the use of CFA in examining the effect of methods on measurements has its own advantages, including being able to determine a more conceptual measurement model and being able to determine the number of variances in each dimension. In addition, the use of CFA can better know the relationship between one dimension and another and its relationship with the measurement construct.

The purpose of this study is to determine the existence of the method effect on the ethnic experience scale (EE) and ethnic identity scale (EI), where both scales are widely used as an instrument in psychological research. The EE and EI scales use favourable and unfavourable items, but there has never been an investigation on the use of two response directions (favourable and unfavourable) of the two scales. So far, the existing studies on the EI scale are Umana-Taylor et al. (2004), and Paramita (2014). They investigated the internal consistency using Cronbach's Alpha, and the result exhibited satisfactory alpha coefficient values that are above .80. Research on EE conducted by Malcarne et al. (2006) found that the scale has an internal consistency value of .83 to .91 and test-retest reliability of .71 to .86, with an estimated time of 6 weeks.

As can be seen from the constructs measured by the two scales, it is very clear that these two scales are very potential scales to be used in many studies in Indonesia, considering that Indonesia is a country with many ethnicities, so ethnic topics are very

important and relevant topics to be studied. Given the potential of both scales in research on ethnicity, it must be ensured that the use of both measuring instruments is as free from biases as possible. One of possible biases is the bias comes from the use of favourable and unfavourable items. This prompted studies in psychometrics that were more in-depth than just Cronbach's Alpha reliability test. The more evidence about the adequate psychometric quality of a test, the more convincing it is to be used for data collection and the score can then be interpreted accurately.

This study not only tested the constructs of the two scales, but also tested the effect of using favourable and unfavourable items on the EE and EI scales. In addition, this study used several alternative models to be tested using CFA. Testing was carried out to determine the existence of method effects from the EE and EI scales through testing the alternative models that had been determined. In addition, the use of the EE and EI scales in this study also aims to test the internal structure of the scales in the Indonesian version. So, the scales can be used for research on ethnicity in Indonesia. For this consideration, this research is deemed necessary.

Research in Indonesia, especially in the field of Psychology, has raised many topics regarding ethnicity, considering that Indonesia is a multi-ethnic country, so the topik of ethnicity will really need instruments that can measure ethnic experience and ethnic identity that have high accuracy and are free from bias. The purpose of this study is to evaluate the possibility of bias in the use of favourable and unfavourable items in two scales, EE, and EI. In addition, this study also aims to evaluate the factor structure of both the EE and EI scales, which can provide evidence of the construct validity.

Furthermore, the urgency of this study is to obtain a scientific basis in deciding whether it is necessary to use two-way items (favourable and unfavourable). Another urgency of this study is to obtain results that can be used as a basis for deciding whether the EE and/or EI scale can be used to provide scores and interpret scores related to the constructs of ethnicity.

Theory and Objectives

Confirmatory Factor Analysis

Confirmatory Factor Analysis (CFA) is a type of Structural Equation Modeling (SEM) which is useful for analyzing the underlying model of a test. CFA analysis specifically deals with measurement models, such as the relationship between indicators and latent variables (Brown, 2015). CFA is used to evaluate the number of underlying dimensions of the test, the relationship between dimensions, and the relationship between items and dimensions (factor loadings). Four major purposes of CFA: (1) psychometric evaluation of measures; (2) construct validation; (3) testing method effects; and (4) testing measurement invariance (Harrington, 2009).

There are three main aspects in CFA analysis, factor loadings, unique variances, and factor variances. Factor loading is estimation of the relationship between items and common factors. Unique variance shows variations that are not shared with other items, while factor variance expresses the sample variability or dispersion of the factor that is, the extent to which sample participants' relative standing on the latent dimension is similar or different (Brown, 2015).

One of the most discussed topics within the CFA is testing model fit. Testing of the model fit was carried out to determine the index of the model fit test (goodness-of-fit) obtained from the analysis on a predetermined model (Tomas & Oliver, 1999). Model fit testing is carried out into three parts, absolute fit, parsimony correction and comparative fit (Brown, 2015).

Alternative Model

The CFA models developed for MTMM analysis are the Correlated Trait-Correlated Method (CTCM) and the Correlated Trait-Correlated Uniquenesses (CTCU) (Brown, 2015). Among the CFA models for MTMM analysis, CFA with CTCM or block-diagonal approach is an alternative model that is widely used (Tomas & Oliver, 1999). Through the concept of the CTCM approach, it can be known specifically the possibility of the method effect on test items, such as the use of positive and negative sentences on items (Lindwall et al., 2012). In addition, the use of CTCM can predict the method effect that forms other constructs other than the measurement objective itself (Lindwall et al., 2012). The four alternative models can be seen in Figure 1.

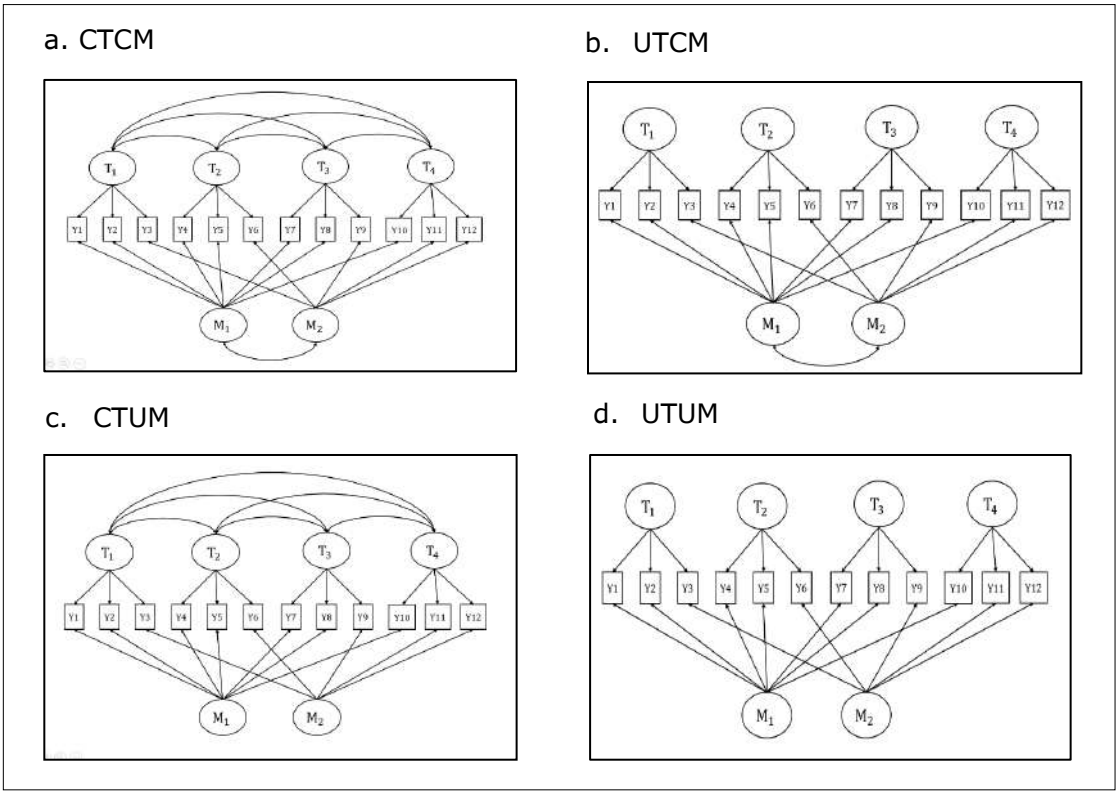


Figure 1. Conceptual framework of MTMM approach

Methodology

This study is a psychometric quantitative study that aims to evaluate the possible biases that come from the use of two-way items (favourable and unfavourable). This research was carried out in four stages, that are, review the scales, data collection, data analysis, and interpretation and decision making, as depicted in Figure 2.

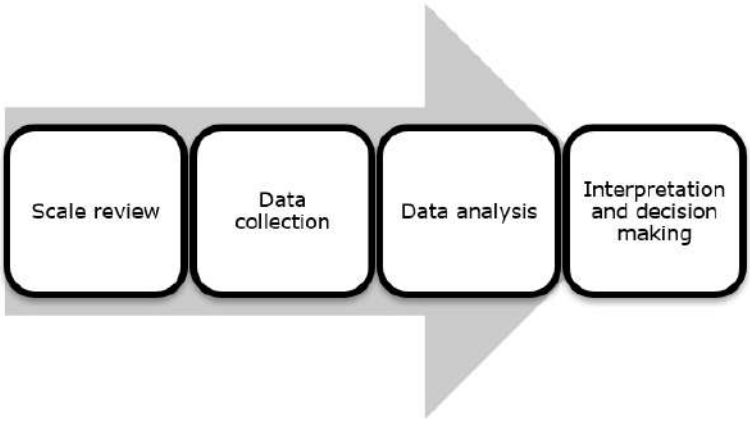


Figure 2. *Flowchart* of four stages of research in psychometric analysis

Scale review

This stage involves the test review, that are the Ethnic Identity and Ethnic Experience scales, which include the investigation on the underlying construct, the number of dimensions, and the psychometric properties that have been carried out.

Data collection

Data collection was conducted by administering the scale offline, and then the data was entered without including the identity of the partisipants. The data is only in the form of responses for all items in the test, and the responses have been coded into quantitative data, with the score ranging from 1 to 4.

Data analysis

After collecting quantitative data, we next tested four alternative models. Interpretation and decision making are carried out based on the results of data analysis, if one or more of the alternative models have a satisfying fit index, it can be interpreted that there is a bias from the use of two-way items (Favourable and Unfavourable).

Sample

All identities of the sample are not used as research data, so the identity of the participants is not listed in the research report. Data collection was carried out in two studies, namely study 1 and study 2.

1. Study 1: The research location was Brawijaya University Malang. EES was administered to students who are in the range of semester 1 to 7, with an age range of 18 – 20 years. Data collection was carried out on 1013 students using non-random sampling techniques.
2. Studi 2: The research location was the junior high school laboratory of State University of Malang, and Brawijaya University. EIS was administered to adolescents and early adults. Data collection was carried out on 650 participants using non-random sampling techniques.

Scale

This study used two scales, EES and EIS, which were administered in two studies (study 1 and study 2), as follows.

1. Ethnic Experience Scale: The Ethnic Experience scale (EES) (Malcarne et al., 2006) has an internal consistency value ranging from .83 to .91 and test-retest reliability ranging from .71 to .86, with an interval time of 6 weeks. Since EES is a

new scale in Indonesia, so in this study, transadaptation was conducted so that the EES can be adjusted to research needs. The EES consists of four dimensions, with 32 items divided into 20 favourable items and 12 unfavourable items. The detail of item distribution on each dimension can be seen in Table 1.

Table 1. Blueprint of Ethnic Experience Scale (EES)

Dimension	F	UF	Total	Item number	Percentage
Ethnic Identity	6	6	12	1* ,4* , 7*, 8*, 12* , 27*, 14, 16 , 20 , 23 , 25 , 30	37.5%
Perceived Discrimination	5	4	9	2* , 3* , 21* , 31*, 13 , 19 , 24 , 26 , 29	28.1%
Mainstream Comfort	4	2	6	11*, 32*, 6 , 9 , 17 , 28	18.8%
Social Affiliation	5	0	5	5 , 10 , 15 , 18 , 22	15.6%
Total			32	100%	

Note.F= Favourable, UF= Unfavourable

The EES contains statements that describe self-experience. The response format used is Likert with five alternative responses, strongly agree (SS), agree (S), Neutral (N), disagree (TS), and strongly disagree (STS).

2. *Ethnic Identity Scale:* This study used the Ethnic Identity Scale (EIS) developed by Umana-Taylor et al. (2004) and has been translated and adapted into Indonesian version by Paramita (2014). EIS consists of 17 items that measure three dimensions, exploration, commitment and affirmation. The scale consists of both favourable and unfavourable items, with a likert scaling model consisting of four response options. Research conducted by Paramita (2014) examined ethnic identity in adolescents of Javanese ethnicity, and ethnic Chinese. Result showed that the EIS scale has Cronbach's alpha .84 for exploration, .76 for commitment and .88 for affirmation. Furthermore, for validity testing, Paramita (2014) used content validity, specifically the face validity method which showed 99.44% of subjects agree that scale cover is clear, 98.15% said the layout and letters used were clear, 92.60% said that the sentence conveyed on the scale was clear. The EIS Blueprint used in this study is a blueprint of the Indonesian version of EIS Paramita (2014), which is presented in Table 2.

Table 2. EIS Blueprint

Dimension	F	UF	Total	Item number	Percentage
Exploration	6	1	7	2*,4,5,6,8,11,15	41,18 %
Commitment	4	0	4	3,12,14,17	23,53 %
Affimation	0	6	6	1*,7*,9*,10*,13*,16*	35,29 %
Total			17	17	100%

**):unfavourable, F=Favourable, UF=Unfavourable*

Result

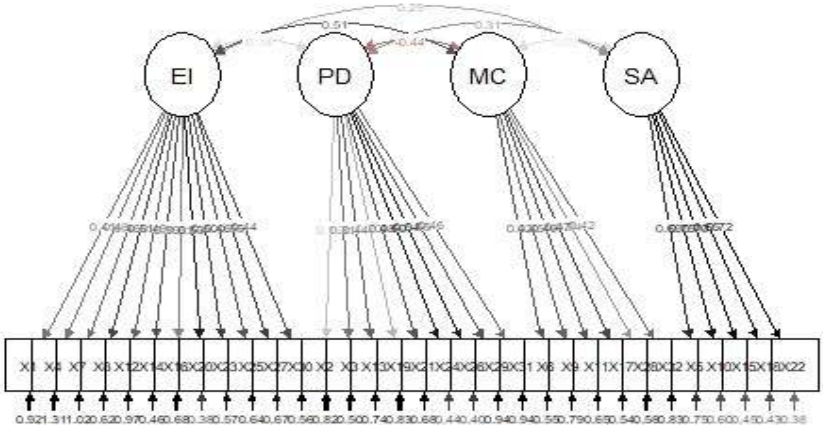
Study Result

Study 1. Confirmatory factor analysis of EES

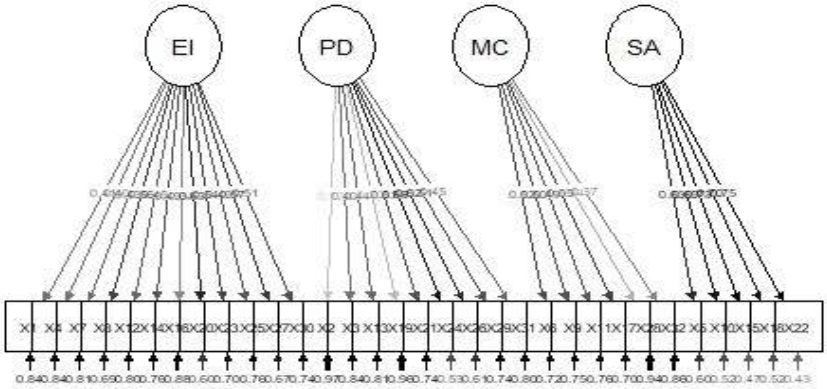
Based on the collected data of 1013 subjects, the data was then analyzed using the CFA technique with four alternative models. The development of model alternatives aims to find the best model fit.

Figure 3. (A-F) EES alternative models

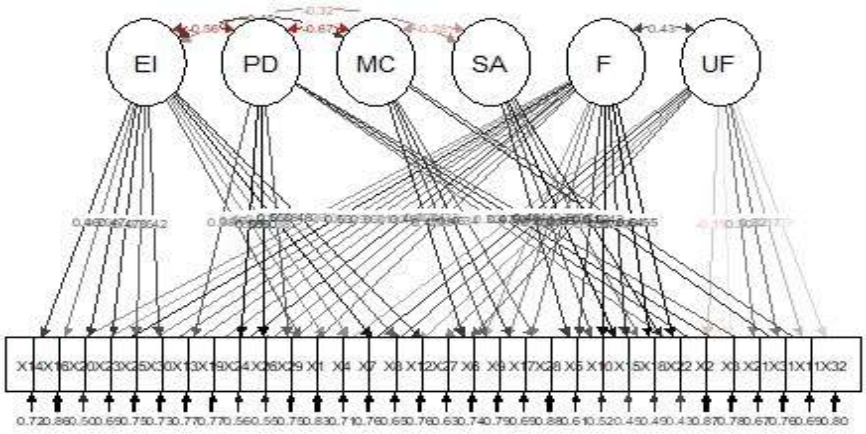
a. Baseline (Correlated Trait)



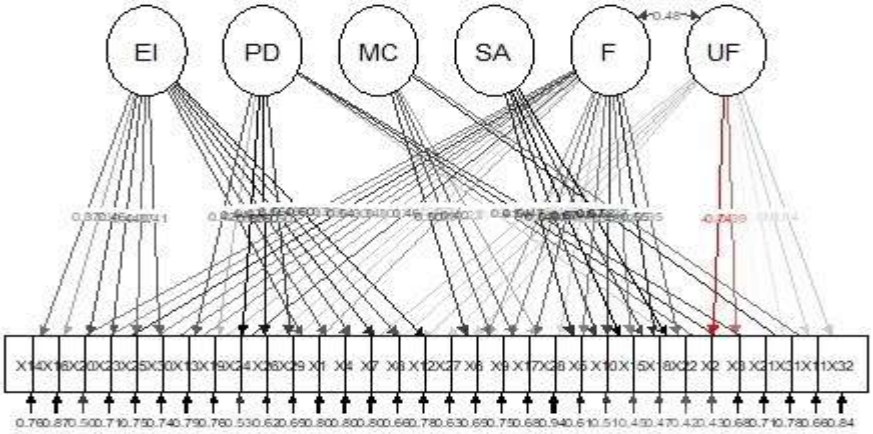
b. Baseline (Uncorrelated Trait)



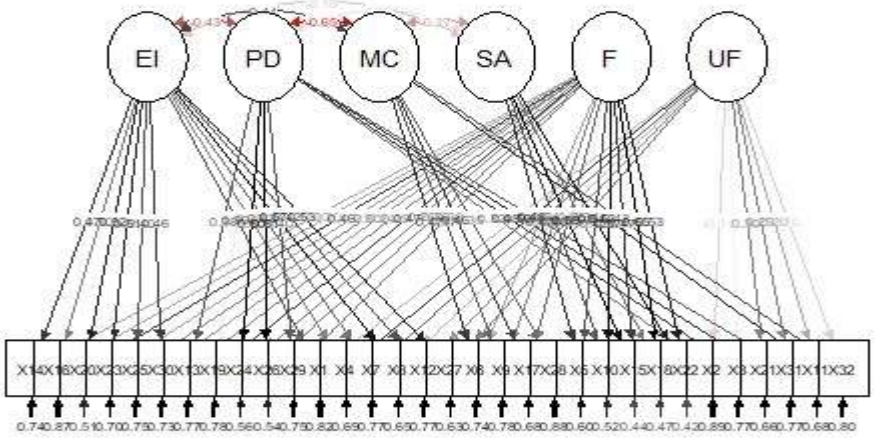
c. CTCM



d. UTCM



e. CTUM



,

f. UTUM

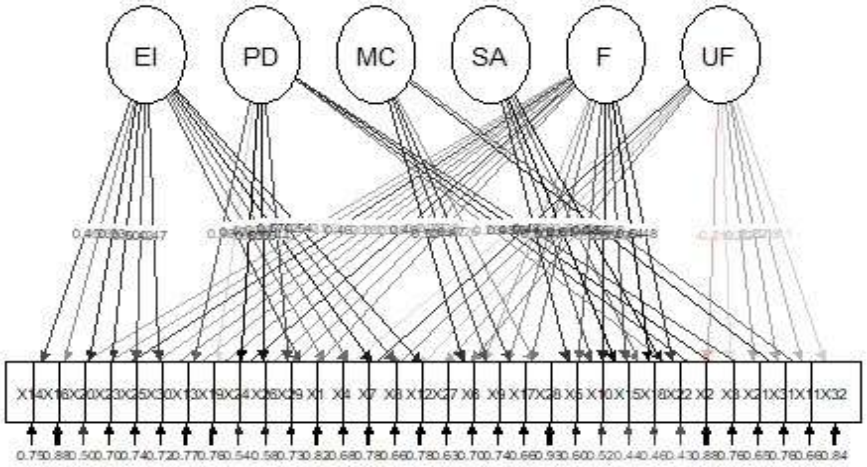


Table 3 presents the values of the goodness of fit test of four alternative models. Before discussing models involving Favourable and Unfavourable, it is necessary to analyze the baseline models, the Correlated Trait (CT) and Uncorrelated Trait (UT). Based on Table 3, the fit index for both models is slightly lower than the benchmark. Even so, when compared between the two, the CT model in general has a higher fit index value than the UT model. In addition, if you look at the correlation between factors in Table 4. The CT model has correlated factors that range from weak to strong.

Table 3. Goodness of fit test results for six alternative models

Index	CFI	TLI	RMSEA	SRMR
	>.95	>.95	≤.05	<.08
CT	.758	.738	.060	.075
UT	.709	.689	.066	.101
CTCM	.859	.836	.048	.046
UTCM	.821	.795	.053	.074
CTUM	.851	.827	.049	.055
UTUM	.808	.779	.055	.081

Note. Cut off point for CFI and TLI >.95, RMSEA <.06, and SRMR <.08 (Hu & Bentler, 1999).

Four models involving the Favourable and Unfavourable methods have fit index presented in Table 3, which indicates that All alternative models that use favourable and unfavourable methods have subpar fit indexes. Among the four models, the CTCM model showed the highest value of fit index.

Table 4. Correlations between dimensions on EES

		EI	PD	MC	SA	F	UF
CT	EI	1.000					
	PD	-.140	1.000				
	MC	.508	-.441	1.000			
	SA	.247	.312	.092	1.000		
UT	EI	1.000					
	PD	0	1.000				
	MC	0	0	1.000			
	SA	0	0	0	1.000		
CTCM	EI	1.000					
	PD	-.559	1.000				
	MC	.449	-.668	1.000			
	SA	-.323	-.016	-.277	1.000		
	F	.000	.000	.000	.000	1.000	
	UF	.000	.000	.000	.000	.432	1.000
UTCM	EI	1.000					
	PD	.000	1.000				
	MC	.000	.000	1.000			
	SA	.000	.000	.000	1.000		
	F	.000	.000	.000	.000	1.000	

	UF	.000	.000	.000	.000	.483	1.000
CTUM	EI	1.000					
	PD	-.432	1.000				
	MC	.439	-.646	1.000			
	SA	-.149	.063	-.266	1.000		
	F	.000	.000	.000	.000	1.000	
	UF	.000	.000	.000	.000	.000	1.000
UTUM	EI	1.000					
	PD	.000	1.000				
	MC	.000	.000	1.000			
	SA	.000	.000	.000	1.000		
	F	.000	.000	.000	.000	1.000	
	UF	.000	.000	.000	.000	.000	1.000

Note. N = 1013. p <.05. EI = Ethnic Identity, PD = Perceived Discrimination, MC = Mainstream Comfort, SA = Social Affiliation , F = Favourable, UF = Unfavourable

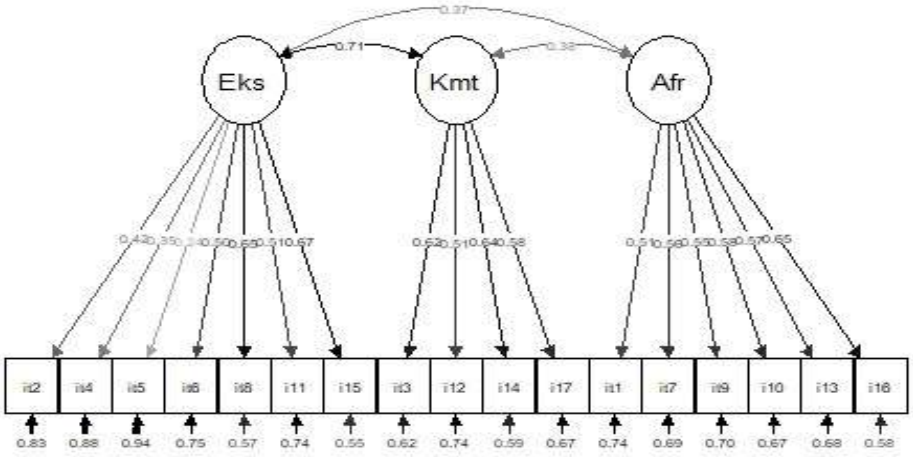
Table 4 shows the correlation values between the factors. In the CTCM and CTUM models, the two models have something in common in terms of correlated factors. The two models are the models with the highest fit index. In both models, the four factors were negatively correlated, except for the correlation between Ethnic Identity and Mainstream Comfort which was positively correlated.

Tables 3 and 4 also show that the Favourable and Unfavourable methods developed their own factors, both correlated and uncorrelated factors. In the CTCM model, both the Favourable and Unfavourable methods have a moderate positive correlation of 0.432. Followed by a UTCM model that shows that the correlation between the methods is a moderate positive of 0.483.

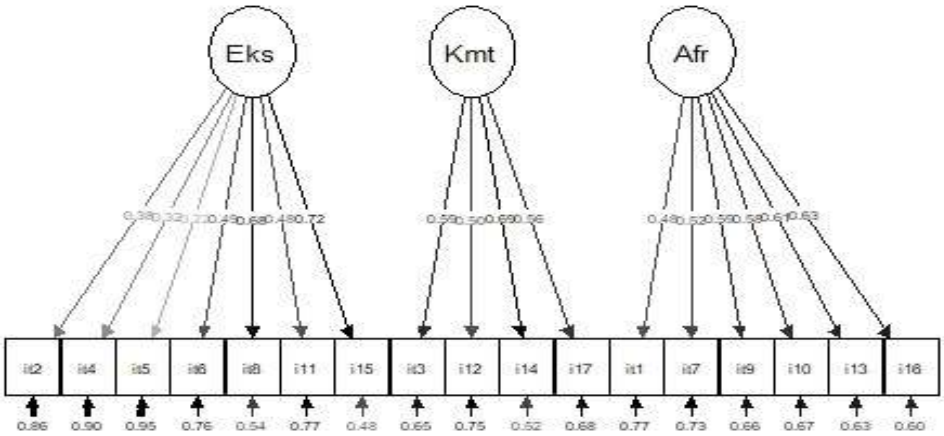
Studi 2 Confirmatory factor analysis of Ethnic Identity Scale (EIS)
Study 2 was conducted based on EIS administered to 650 subjects aged 12-40 years. Model testing was carried out on two baseline models and four MTMM models involving two methods, Favourable and Unfavourable.

Figure 4. (A-F) Six alternative models for EIS

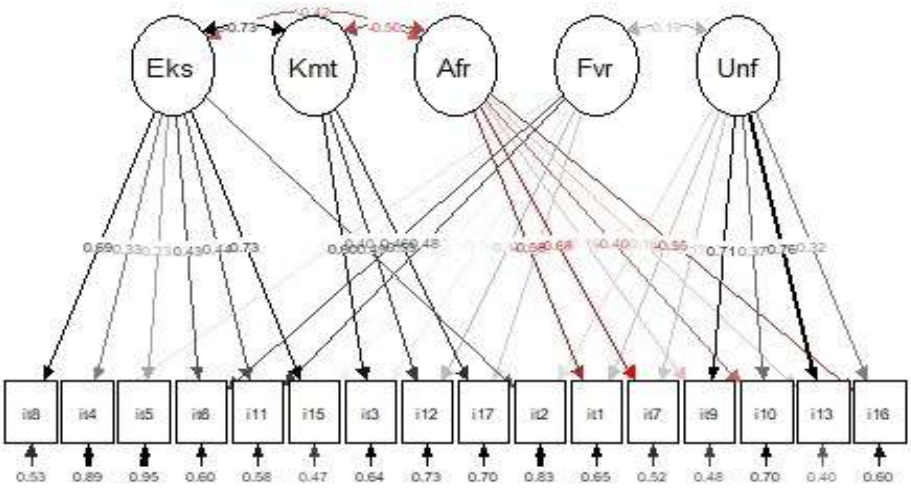
a. Baseline model (Correlated Trait)



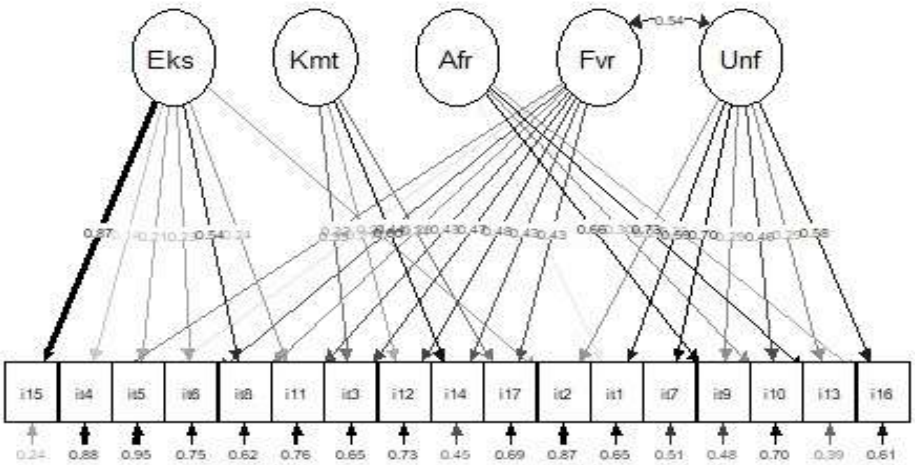
b. Baseline model (Uncorrelated Trait)



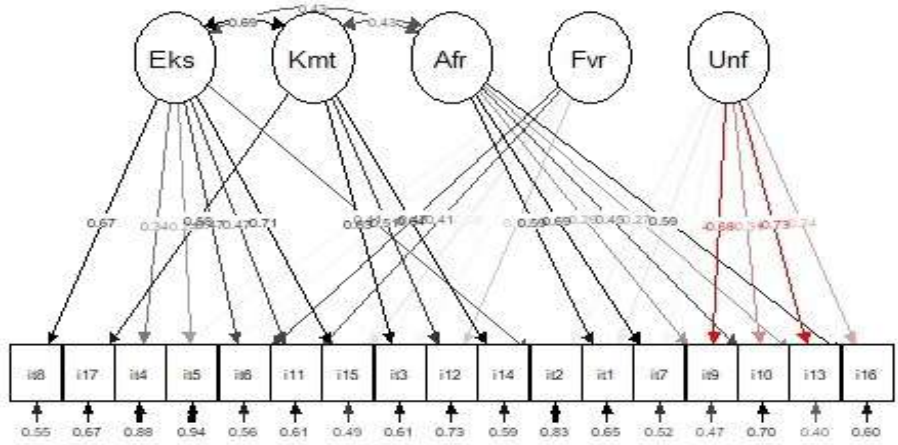
c. CTCM



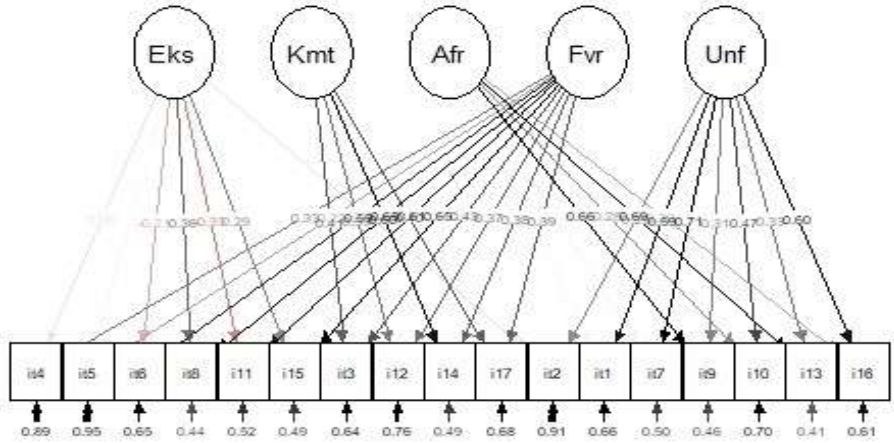
d. UTCM



e. CTUM



f. UTUM



The models in Figure 4 were tested for model fit using the CFA, and the results are presented in Table 5. In the two baseline models, CT and UT, showed unsatisfactory fit indexes. Even so, when compared to the UT model, the CT model showed a better fit index. Therefore, for the CT baseline model it is more focused on the correlation between dimensions, which is presented in Table 6. In the CT model, the correlation between the three dimensions is positive, and overall, the correlation ranges from moderate to very strong.

Table 5. Goodness of fit test results for six alternative models of EIS

Index	CFI	TLI	RMSEA	SRMR
	>.95	>.95	<.06	<.08
CT	.853	.827	.064	.055
UT	.737	.699	.085	.131
CTCM	.958	.941	.038	.037
UTCM	.898	.865	.057	.075
CTUM	.948	.930	.041	.040
UTUM	.893	.857	.058	.098

Note. Cut off point for CFI dan TLI >.95, RMSEA <.06, dan SRMR <.08 (Hu & Bentler, 1999).

For the other four models, which use favourable and unfavourable methods, the CTCM and CTUM models have the best fit indexes. Based on the cut-off point for the fit index proposed by Hu and Bentler (1999), both models are acceptable because they have a fit index that meets the requirements. This shows that the results are in line with the baseline model, where the CT model is a more powerful model when compared to UT. Both models show positive correlations between dimensions and range from moderate to strong.

Table 6. Correlations between dimensions on EIS

		Exploration	Commitment	Affirmations	F	UF
CT	Exploration	1.000				
	Commitment	.709	1.000			
	Affirmations	.375	.378	1.000		
UT	Exploration	1.000				
	Commitment	.000	1.000			
	Affirmations	.000	.000	1.000		
CTCM	Exploration	1.000				
	Commitment	.726	1.000			
	Affirmations	.423	.502	1.000		
	F	.000	.000	.000	1.000	
	UF	.000	.000	.000	.186	1.000
UTCM	Exploration	1.000				
	Commitment	.000	1.000			
	Affirmations	.000	.000	1.000		
	F	.000	.000	.000	1.000	
	UF	.000	.000	.000	.542	1.000
CTUM	Exploration	1.000				
	Commitment	.691	1.000			
	Affirmations	.426	.434	1.000		
	F	.000	.000	.000	1.000	
	UF	.000	.000	.000	.000	1.000
UTUM	Exploration	1.000				
	Commitment	.000	1.000			
	Affirmations	.000	.000	1.000		
	F	.000	.000	.000	1.000	
	UF	.000	.000	.000	.000	1.000

Discussion

This study aims to determine the method effect caused by favourable and unfavourable items in an instrument. Podsakoff defines the effect method as a response that is based on constructs other than the construct being measured (Lindwall et al., 2012). One method that can cause the method effect is the use of positively directed items (favourable) and negatively directed (unfavourable). The instrument used in this study is EES, the scale has four correlated dimensions (Malcarne et al., 2006). This study also

uses an Ethnic Identity Scale (EIS) (Paramita, 2014), consisting of three dimensions, Exploration, commitment, and affirmation.

In this study, EES has been transadapted into Indonesian with 32 items consisting of 20 favourable items and 12 unfavourable items. Similarly, EIS has been transadapted into indonesian and has 10 favourable items and 7 unfavourable items. Favourable and unfavourable are said to be two opposite directions, so the method produces a diversity of data that can form new constructs (Widhiarso, 2016a, 2016b).

Based on the two studies, in study 1, the baseline model showed, EES had a fit index that was not very satisfactory. While study 2 shows, EIS has a high fit index for the CT model, although it is not yet fully qualified. The results of the EIS showed that the instrument was satisfactory enough based on the fit model and showed that the EIS had correlated factors. These results suggest that EES and EIS still need improvement.

As a check on method effect, this study tested four alternative models. The alternative models proposed illustrate the use of favourable and unfavourable items in instruments that form new factors (Rodebaugh et al., 2007). This study used several alternative models, CTCM, UTCM, CTUM, and UTUM. These four alternative models were analyzed using CFA techniques to find the best fit model. The CFA technique serves to ensure that the methods used in measuring instruments are in accordance with the purpose of measuring (Brown, 2015).

Fit model test uses more than one fit index i.e. CFI, TLI, RMSEA, and SRMR. According to Brown (2015), the fit model shows a better fit index when the CFI and TLI values are close to 1, while the RMSEA and SRMR indices get better when they are close to 0. However, Hu and Bentler (1999) specified that each model fit index has standards to meet model fit criteria. Therefore, there is no perfect fit index to use, so it is necessary to use more than one fit index when evaluating the model (Brown, 2015).

Based on the results of the analysis, the CTCM model and the CTUM model can be said to be the most fit models in study 1 and study 2, because both models obtained better fit index values compared to UTCM, and UTUM model. This shows that there is indeed an effect of the method on the two scales, EES and EIS.

These results show that favourable, and unfavourable involvement in both EES and EIS leads to response groupings and forms new factors. In the CTCM model, the use of favourable and unfavourable items produces moderate correlation, while in the CTUM model favourable items and unfavourable items do not correlate with each other. Thus, the CTCM and CTUM models can be said to be the models that best demonstrate the method effect, due to the use of favourable and unfavourable items in EES, and EIS.

The finding showing the CTCM, and CTUM models are the most superior, indicates there exists correlation between factors within EES and EIS. These results are in line with the baseline model, which shows the CT model is the most superior, so CTCM and CTUM also have the best fit index. This further convinces that the factors in the EES and EIS constructs are indeed correlated with each other.

In accordance with the research of Malcarne et al. (2006) that EES consists of several correlated dimensions. Similarly, for EIS, the correlation between dimensions shows similar results as finding of research conducted by Yoon (2011). This finding itself is not

in line with the initial assumption in this study, that the dimension in the EIS is multidimensional with uncorrelated dimensions as stated by Umana-Taylor et al. (2004). Some of the assumptions that to avoid response bias is to apply Favourable and Unfavourable items in an instrument (Nunnally, 1978). However, both studies found that the use of opposite items can form new factors, so the use of favourable and unfavourable needs to be done with caution because of the potential of method effects that can threaten the construct validity.

Conclusion

Based on the results of two studies, it shows that the basic models of both scales still need improvement to get the best measurement results in the future. The most important result is about the method effect, which shows the method effect as a result of using favourable items, and unfavourable items on EES and EIS. So, the effect of using favourable and unfavourable items should be considered as a potential problem in worsening the construct validity.

References

Brown, T. A. (2015). *Confirmatory Factor Analysis for Applied Research Second Edition.* New York: The Guilford Press.

Conroy, D. E. (2002). Progress in the development of a multidimensional measure of fear of failure: The Performance Fear Appraisal Inventory (PFAI). *Anxiety, Stress, and Coping, 14,* 431–452.

DiStefano, C., & Motl, R. W. (2002). Personality correlates of method effects due to negatively worded items on the Rosenberg Self-Esteem Scale. *Personality and Individual Differences*, *46,* 259–388.

DiStefano, C., & Motl, R. W. (2006). Further Investigating Method Effects Associated With Negatively Worded Items on Self-Report Surveys. *Structural Equation Modeling, 13(3)* , 440-464.

Harrington, D. (2009). *Confirmatory factor analysis*. Oxford university press.

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, DOI: 10.1080/10705519909540118

Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., & Liukkonen, J. (2012). Method Effects: The Problem With Negatively Versus Positively Keyed Items. *Journal of Personality Assessment* , 196-204.

Malcarne, V. L., Chavira, D. A., Fernandez, S., & Liu, P. (2006). The Skala ethnic experience: Development and Psychimetric Properties. *Journal of Personality Assessment*, 86(2), 150-161

Marsh, H. W. (1996). Positive and negative global self-esteem: A substantively meaningful distinction or artifactors? *Journal of Personality and Social Psychology, 70,* 810–819.

Nunnally, J. C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Paramita, W. M. (2014). *Perbedaan Identitas Etnis pada Remaja Etnis Jawa dan Etnis Identitas.*Unpublished manuscript*,* Universitas Brawijaya.

Rodebaugh, T. L., Woods, C. M., & Heimberg, R. G. (2007). The Reverse of Social

Anxiety Is Not Always the Opposite: The Reverse-Scored Items of the Social Interaction Anxiety Scale Do Not Belong. *Behavior Therapy*, 38(2), 192-206. DOI: 10.1016/j.beth.2006.08.001.

Susetyo, D. P. B. (2002). Krisis Indentitas Etnis Cina di Indonesia. *Kajian Ilmu Psikologi*, Vol 6, 61-71

Tomas, J. M., & Oliver, A. (1999). Rosenberg's Self Esteem Scale: Two Factors or Method Effects. *Structural Equation Modeling: A multidisciplinary Journal* , 84-98.

Umana-Taylor, A., Yazedjian, A., & Bamaca-Gomez, M. (2004). Developing the Ethnic Identity Scale.*Journal of Theory And Research*, 9-38.

Van de Looij-Jansen, P. M., Goedhart, A. W., de Wilde, E. J., & Treffers, P. D. A. (2011). CFA And Factorial Invariance Analysis of The Adolescent Selfreport, Strengths and Difficulties Questionnaire: How Important are Method Effects and Minor Factors. *British Journal of Clinical Psychology*, 50(2), 127-144. DOI: 10.1348/014466510x498174

Van Roy, B., Veenstra, M., & Clench-Aas, J. (2008). Construct validity of the five-factor Strengths and Difficulties Questionnaire (SDQ) in pre-, early, and late adolescence. *Journal of Child Psychology and Psychiatry*, *49*(12), 1304–1312. doi:10.1111/j.1469-7610.2008.01942.x

Widhiarso, W. (2016a). *Mengakomodasi Efek Metode dalam Pengujian Validitas Konstruk Melalui Analisis Faktor Konfirmatori*. *Psikologia (Jurnal Psikologi),* 1 (1), 37-51. DOI: 10.21070/psikologia.v1i1.478

Widhiarso, W. (2016b). *Peranan Butir Unfavourable dalam Menghasilkan Dimensi Baru dalam Pengukuran Psikologi*. Fakultas Psikologi Universitas Gadjah Mada.

Yoon, E. (2011). Measuring Ethnic Identity in the Ethnic Identity Scale and the Multigroup Ethnic Identity Measure-Revised. Cultur Divers Ethnic Minor Psychol. 144-55. doi: 10.1037/a0023361. PMID: 21604838.


ACER


HIMPUNAN EVALUASI
HEPI
PENDIDIKAN INDONESIA


ICAL
International Conference
on Assessment and Learning