PROCEEDING

INTERNATIONAL CONFERENCE ON EDUCATIONAL RESEARCH AND EVALUATION (ICERE)

“Assessment for Improving Students’ Performance”

May 29 – 31 2016
Rectorate Hall and Graduate School
Yogyakarta State University
Indonesia
Organized by:
Study Program of Educational Research and Evaluation
Graduate School, Yogyakarta State University
in Cooperation with Indonesian Educational Evaluation Association (HEPI),
and Center for Educational Assessment (PUSPENDIK) Ministry of Education and Culture

Publishing Institute
Yogyakarta State University

Director of Publication
Prof. Djemari Mardapi, Ph.D.

Board of Reviewers
Prof. Djemari Mardapi, Ph.D.
Prof. Dr. Badrun Kartowagiran
Prof. Geoff Masters, Ph.D.
Prof. Frederick Leung, Ph.D.
Bahrul Hayat, Ph.D.
Jahja Umar, Ph.D.
Prof. Burhanuddin Tola, Ph.D
Bambang Suryadi, Ph.D

Editors
Ashadi, Ed.D.
Suhaini M. Saleh, M.A.
Titik Sudartinah, M.A.

Layout
Rohmat Purwoko, S.Kom.
Syarief Fajaruddin, S.Pd.

Address
Yogyakarta State University
ISSN: 2407-1501
@ 2016 Yogyakarta State University

All right reserved. No part of this publication may be reproduced without the prior written permission of Yogyakarta State University.

All articles in the proceeding of International Conference on Educational Research and Evaluation (ICERE) 2016 are not the official opinions and standings of editors. Contents and consequences resulted from the articles are sole responsibilities of individual writers.
Foreword of the Chairman

Assalamualaikum wr. wb.

Good morning ladies and gentlemen.

Praise be to Allah who has given abundant blessings so that we can hold this international conference.

This conference is aimed at improving the quality of assessment implemented in schools and other institutions. The quality of assessment determines students' ways of learning, so that it is hoped that the quality of education improves. Besides, this conference is a means of information exchanges in the forms of seminars dealing with results of research in educational assessment and evaluation. The expectation is that there is always improvement in educational assessment and evaluation methods, including in it is the instrument – both cognitive and noncognitive instruments.

The participants of this conference are the lecturers and teachers who teach educational assessment and evaluation, practitioners of assessment and evaluation, and researchers of assessment and evaluation. This conference can be held in cooperation with the Graduate School, Yogyakarta State University, Association of Educational Evaluation of Indonesia (HEPI), and Centre for Educational Research, Ministry of Education and Culture of Indonesia, supported by the Australian Council for Educational Research (ACER), Intel, Intan Pariwara Publisher, and many other institutions. For this reason, on behalf of the Organizing Committee, I would like to thank the Rector of Yogyakarta State University, Prof. Dr. Rochmat Wahab, M.Pd., M.A., and the Director of Graduate School, Yogyakarta State University, Prof. Dr. Zuhdan Kun Prasetyo, M.Ed., and all other institutions for their assistance and contribution that have made this conference possible. I would like to thank HEPI’s Local Coordination Unit and all sponsors for supporting this conference and also all the audience for participating in this conference.

To the committee members, both in Jakarta and Yogyakarta, I would like to thank them for the hard work they have performed and for the togetherness so that this conference can be held.

Last but not least, we apologize for all the inconveniences you might encounter during this conference. Please enjoy the conference.

Wassalamu’alaikum wr. wb.

Prof. Djemari Mardapi, Ph.D.
Foreword of the Chairman of Himpunan Evaluasi Pendidikan Indonesia (HEPI)

Assalamu’alaikum Wr. Wb.

Indonesian Association for Educational Evaluation (HEPI) is a professional organization in education holding in high esteem the principles of professionalism and knowledge development in the field of educational and psychological measurement, assessment, and evaluation. HEPI was established in November 19, 2000 in Yogyakarta, with a vision to become a professional organization that excels in the field of evaluation and measurement in education and psychology in Indonesia. Its mission is to develop up-to-date methodologies of evaluation, assessment, measurement, and data analysis in education and psychology, as well as studies of policies and technical implementation of the field for improving Indonesian education quality.

As a professional organization, HEPI brings together experts, practitioners and interested persons in the field of evaluation, assessment, and measurement of education, psychology and other social sciences. HEPI is open to anyone who has the interest the field with no restriction in terms of educational background and working experiences. Hopefully, through HEPI, members of the association can sustainably develop themselves as professionals. The existence of HEPI is also expected to contribute to the improvement of the quality of national education through research, consultancy, seminar, conference, publication, and training for members of the organization and for public audiences.

HEPI organizes annual workshop and conference in cooperation with the Regional Chapter of HEPI and universities. In 2016, for the first time HEPI organized International Conference on Educational Research and Evaluation: Assessment for Improving Student’s Performance in May 29-30 2016 in Yogyakarta. This conference is jointly organized by HEPI and Yogyakarta State University and supported by the Center for Educational Assessment the Ministry of Education and Culture, Australian Council for Educational Research (ACER), INTEL Indonesia, and Intan Pariwara Publisher.

It is important to note that the choice of the HEPI 2016 conference theme is driven by the fact that the quality of our national education is still under expectation as shown by the results from School National Exam and international surveys conducted by some international agencies. HEPI believes that a number of factors contribute to the low quality of national education, including low teacher’s knowledge and skills in classroom and school assessment. Therefore, improving the competence of teachers in classroom and school assessment is urgently required. In this context HEPI as a professional organization and individual members of the organization have to play an active role in improving teachers’ competence in quality learning assessment.

In line with 2016 conference theme, HEPI invited two respected guest speakers, namely, Professor Geoffrey Masters, Ph.D., Director of the Australian Council for Educational Research (ACER), who presented a paper on Assessment to Improve Student Competency and Professor Frederick Leung, Ph.D., from the University of Hong Kong, who delivered a paper on the International Assessment for Improving Classroom Assessment.

As a tradition, in 2016 conference HEPI organized two pre-conference workshops. The first workshop is on the conceptual introduction of Rasch model by Jahja Umar, Ph.D., senior lecturer at the Faculty of Psychology, State Islamic University Jakarta and the second workshop was delivered by Heru Widiatmo, Ph.D., researcher at American College Testing (ACT) Iowa, United States on Measuring Higher Order Thinking Skills (HOTS).

On behalf of HEPI, I would like to express my heartfelt gratitude to Rector of the Yogyakarta State University, invited speakers, resource persons, HEPI regional chapters, sponsors, speakers, participants, invited guests, and organizing committee who have worked hard in making this international conference a success. Thank you very much for your participation and support and we are looking forward to seeing you in the next conference.

Last but not least, we hope that all of us get much benefit from this conference for enhancing Indonesian quality education through quality assessment.

Wassalamualaikum wr. wb.

Chairman,

BAHRUL HAYAT, Ph.D.
# Table of Contents

## Foreword of the Chairman

i

## Foreword of the Chairman of Himpunan Evaluasi Pendidikan Indonesia (HEPI)

ii

## Table of Contents

iii

## Invited Speakers

- Assessment for Improving Student Performance  
  *Prof. Geoff Master, Ph.D.*,  
  International Assessment for Improving Classroom Assessment  
  *Prof. Frederick Leung, Ph.D.*  
  Educational Quality assurance For Improving Quality of Education  
  *Bahrul Hayat, Ph.D.*

## Parallel Session Speakers

### I. Sub Themes:

- **Assessment Methods for Improving Student’s Performance**
  
  - Assessment Model for Critical Thinking in Learning Global Warming Scientific Approach  
    *Agus Suyatna, Undang Rosidin*  
  
  - The Nationalism Attitude Assessment of Students of State Senior High School 1 Pakem Sleman  
    *Aman*  
  
  - The Design of Formative Assessment by Inquiry Based Learning in Improving Students’ Self-Regulation  
    *Asih Sulistia Ningrum, Chandra Ertikanto*  
  
  - Exploring the Use of One Meeting Theme-Based Extended Response A Practical Critical Thinking Assessment Tool for Classroom Practices  
    *Ayu Alif Nur Maharani Akbar, Rahmad Adi Wijaya*  
  
  - Application of Instructional Model of Daily Assessment for Improvement of Processes Quality and Instructional Outcomes  
    *Benidiktus Tanujaya*  
  
  - Assessing Student’s Pragmatics’ Knowledge at Islamic University of Riau  
    *Betty Sailun*  
  
  - The Teacher’s Performance in Learning Process Management And Chemistry Learning Difficulties Identification  
    *Budi Utami, Sulistyos Saputro, Ashadi, Mohammad Masykuri, Nonoh Siti Aminah*
Components of Scientific Attitude for Teacher Observation in Physics Learning in Senior High School
Elvin Yusliana Ekawati

The Development of Psychomotor Competency Assessment on Physics Education Student of Palangka Raya University
Enny Wijayanti

Implementation of Authentic Assessment in Bahasa Indonesia Subject for Senior High School in West Sumbawa
Eny Rusmaini

Summative Assessment Design through the PjBL to Improve Students’ Higher-Order Thinking Skills
Erlida Amnie

Assessment Model Multiple Intelligences Learning Approach in Primary School Mathematics Subjects
Helmiah Suryani, Badrun Kartowagiran

Indicator Development of Learning Model Evaluation Instrument
Herpratiwi, Tien Yulianti, Adil Fadillah H, Bajawati

Performance Assessment in Model of Learning Superflex®
Huriah Rachmah

The Identification of Teachers Difficulties in Implementing of 2013 Curriculum at Elementary Schools
Ika Maryani, Sri Tutur Martaningsih

Aerobic Gymnastics, Fitness, and Academic Grade of Health Diploma Students from Remote Areas In Indonesia
Lucky Herawati, Maryana, Suharyono

Analyzing the Authenticity of Authentic Assessment
Luki Yunita, Salamah Agung, Eka Novi

Design of Performance Assessment Based on Problem Based Learning in Improving Students’ Self Regulation
Luthfi Riadina, Agus Suyatna, Undang Rosidin

Implementation of Performance Assessment to Increase Biology Learning Achievement by Using Inquiry Model
Murni Saptara Sari

Teachers’ Belief in Implementing Feedback for Students’ Writing in ESP Classroom
Nisrin Adelyna Darayani, Rini Amelia

Comparison of Character Value Between Lower Class and Upper Class at Salman Al Farisi 2 Elementary Integrated School
Rosaria Irjanti, Farida Agus Setiawati

Authentic Assessment in the Learning of Social Studies
Rudy Gunawan
The Implementation of Assessment Model Based on Character Building to Improve Discipline and Student’s Achievement
Rusijono

The Design of Performance Assessment Based Guided Inquiry for Empowering Students’ Argumentation Skills
Saiful Imam Ali Nuradin, Viyanti

The Influence of Class Climate and Self Concept towards Achievement Motivation and Physics Learning Result of Student at XI IPA Grade SMA Negeri 1 Kahu
Satriani, Kaharuddin Arafah, Muris

Assessment Cognitive for Physic: Development of Misconception Physic Test for Junior High School in Bangka Barat with Polotomous Model (PCM)
Sikto Widi Asta, Dedek Andrian

Identifying of Undergraduate’s Analytical Ability about Electric Current in Transistor Using Isomorphic Assessment
Sri Hartini, Dewi Dewantara, Misbah, Syubhan Annur

A Performance-Based Assessment as a Current Trend in ELT: Investigating Its Washback Effects on Secondary-School Student Learning
Sumardi

Developing an Authentic Assessment Science Process Skills, Creative Thinking Skills and Manipulative Skills
Supahar, Dadan Rosana, Zamzam F A, Ryani Andryani, Neviana Wijayanti

Using of Self Assessment to Determine Science Process Skill and Concept Attainment Through Inquiri Learning of 8th Grade Student on 21th Junior High School in Ambon
Wa Nurlina, K. Esomar, I. H. Wenno

Winarno

The Development of Vocational Interest Instrument for Career Exploration of Junior High School Students
Yudhi Satria Restu Artosandi, Sudji Munadi

Self-Assessment of Teachers of Mathematics Vocational High School in Yogyakarta City on the Performance Post-Certification
Zuli Nuraeni

II. Sub Themes:
- The Use Of Psychometric Method for Majoring Student’s Competence

  The measurement Model of Historical Consciousness
Aisiah

  Anbuso: Practical Software to Perform Item Analysis
Ali Muhson, Barkah Lestari, Supriyanto, Kiromim Baroroh

  Estimating of Students Capability Growth in Vertical Equating with Rasch Model Test
Anak Agung Purwa Antara

  Estimating of Students Capability Growth in Vertical Equating with Rasch Model Test
Anak Agung Purwa Antara
<table>
<thead>
<tr>
<th>Title</th>
<th>Author(s)</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diagnostic Test Characteristics of Learning Difficulties in Mathematics for Science Class 12th Grader</td>
<td>Apri Triana, Heri Retnawati</td>
<td>225</td>
</tr>
<tr>
<td>Assessing Science Process Skills using Testlet Instrument</td>
<td>Ari Syahidul Shidiq, Sri Yamtinah, Mohammad Masykuri</td>
<td>231</td>
</tr>
<tr>
<td>The Effect of Multiple Choice Scoring Methods and Risk Taking Attitude toward Chemistry Learning Outcomes (An Experiment at SMA Negeri 13 Kota Bekasi, West Java)</td>
<td>Awaluddin Tjalla, Sari Fitriani</td>
<td>235</td>
</tr>
<tr>
<td>Development of Personal Integrity Scale: Construct Validity</td>
<td>Bambang Suryadi, Yunita Faela Nisa, Nenang Tati Sumiati</td>
<td>242</td>
</tr>
<tr>
<td>Argument-based Validity of Situational Judgment Test for Assessing Teaching Aptitude</td>
<td>Budi Manfaat</td>
<td>248</td>
</tr>
<tr>
<td>Horizontal Equating in Accounting Vocational Theory Test Based on Mean/Mean Method of Item Response Theory</td>
<td>Dian Normalitasari Purnama, Sigit Santoso</td>
<td>253</td>
</tr>
<tr>
<td>The Effect of Number of Common Items on the Accuracy of Item Parameter Estimates with Fixed Parameter Calibration Method</td>
<td>Dina Huriaty</td>
<td>259</td>
</tr>
<tr>
<td>Analysis of Inter-Rater Consistency in Assessment Final Project Fashion Study Program</td>
<td>Emy Budiasasti</td>
<td>265</td>
</tr>
<tr>
<td>Using Fuzzy Logic to Select Item Test in Computerized Base Testing</td>
<td>Haryanto</td>
<td>269</td>
</tr>
<tr>
<td>An Application of the Generalized Logistic Regression Method in Identifying DIF (Analysis of School Examination in Soppeng)</td>
<td>Herwin</td>
<td>276</td>
</tr>
<tr>
<td>Effects of Complexity Matter and Grouping Students of the Statistics Analysis Capabilities</td>
<td>Ismanto</td>
<td>284</td>
</tr>
<tr>
<td>Construct Validity of the TGMD-2 in 7–10-Year-Old Surakarta Children with Mild Mental Disorder</td>
<td>Ismaryati</td>
<td>289</td>
</tr>
<tr>
<td>Measurement of the Quality of Mathematics Conceptual Understanding through Analysis of Cognitive Conflict with Intervention</td>
<td>Iwan Setiawan HR, Ruslan, Asdar</td>
<td>296</td>
</tr>
<tr>
<td>Modification of Randomized Items Selection and Step-Size Based on Time Response Model to Reduce Item Exposure Level of Conventional Computerized Adaptive Testing</td>
<td>Iwan Suhardi</td>
<td>302</td>
</tr>
<tr>
<td>Characteristics of an Instrument of Vocational Interest Scales</td>
<td>Kumaidi</td>
<td>310</td>
</tr>
<tr>
<td>Rasch Model Analysis for Problem Solving Instrument of Measurement and Vector Subject</td>
<td>Mustika Wati, Yetti Supriyati, Gaguk Margono</td>
<td>315</td>
</tr>
</tbody>
</table>
Analysis of Mathematical Reasoning Ability of Elementary School Students Using Timss Test Design
Noening Andrijati

The Accuracy of Testees’ Ability Estimation of The Essay Test and Testlets in Mathematics Through The Graded Response Model (GRM) Application
Purwo Susongko, Wikan Budi Utami

The Comparison of Logistics Model on Item Response Theory: 1 Parameter (1pl), 2 Parameters (2pl), And 3 Parameters (3pl)
Rida Sarwiningsih, Heri Retnawati

Validity and reliability examination of indicators development materials instruction at Elementary School base on Curriculum 2013
Rochmiyati

Analysys Item Information Function on the Test of Mathematics
Rukli

Misuses Cronbach Alpha On Achievement Tests
Satrio Budi Wibowo

Item Discrimination of Two Tier Test on Hydrolysis of Salt
Sri Yamtinah, Haryono, Sulistyso Saputro, Bakti Mulyani, Suryadi BU

An Analysis of Test Quality by Using ITEMAN
Tia Nur Istianah, Desrin Lebafi

An Analysis of Person Fit Using Rasch Model
Yessica Mega Aprita, Yolandaru Septiana

Detecting Students Learning Difficulties Using Diagnostic Cognitive Tests
Yuli Prihatni

III. Sub Themes:

- Developing Instruments of Educational Assessment

Development and Implementation of Higher Order Thinking Skills Instruments in Physics Education
A. Halim, Yusrizal

Developing Picture Series and Vocabulary to Increase English Speaking Skill
Agustina Ellyana, Ketut Martini and Agus Risna Sari

Indonesian Adaptation Scale of Zung Self-Rating Anxiety Scale (SAS)
Alfiannor Luthfi Hasain

Development Hypothetical Model Resources Management Studies Teachers of Hindu Religion
Aris Biantoro, I Made Sutharjana, Wayan Sukarlinawati

Indonesian Adaptation of Organizational Commitment Questionnaire from Meyer & Allen, 2004
Baqiyatul Auladiyah
<table>
<thead>
<tr>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Creativity Problems Test Form Students Complete Description of Learning Connection with Learning Outcomes Counting Mathematics in Primary</td>
<td>411</td>
</tr>
<tr>
<td>Effectiveness Guided Discovery Approach Through Cooperative Learning Think Pair Share (TPS) Type in Terms of Students’ High Order Thinking Skill (HOTS)</td>
<td>418</td>
</tr>
<tr>
<td>Indonesian Adaptation on Scale of Readiness for Organizational Change</td>
<td>421</td>
</tr>
<tr>
<td>Developing Achievement Tests in Physics For Classroom Assessment</td>
<td>427</td>
</tr>
<tr>
<td>The Development of Evaluation Model Education Life Skill Program Out of School Education</td>
<td>434</td>
</tr>
<tr>
<td>Development of Performance Assessment in Guided Inquiry Learning to Improve Metacognitive Skills and Student’s Achievement</td>
<td>440</td>
</tr>
<tr>
<td>Design Student Development Work Sheet (Learning Cycle) 5E to Improve Student Learning Outcomes High School Class X</td>
<td>445</td>
</tr>
<tr>
<td>Development of Vocational Interest Scale: A preliminary study of the psychometrics properties*</td>
<td>449</td>
</tr>
<tr>
<td>Contextual Approach Using Pictures as a Media Increased Result and Motivation of Mathematical Learning (Mathematical Learning of Fractional Addition by Equalizing the Denominator)</td>
<td>455</td>
</tr>
<tr>
<td>The Content Validity of the Evaluation Model in the Affective Domain in Islamic Education Instruments</td>
<td>461</td>
</tr>
<tr>
<td>Developing Science Process Skill Instrument of Islamic Senior High Schools</td>
<td>467</td>
</tr>
<tr>
<td>Online Exam Model of Item Response Theory Based Cat Using Moodle Learning Management System</td>
<td>473</td>
</tr>
<tr>
<td>Developing an Accreditation Model of Secondary School</td>
<td>483</td>
</tr>
<tr>
<td>Developing an Instrument for Assessing the Performance of High School Physics Teacher</td>
<td>490</td>
</tr>
<tr>
<td>Analysis Instruments Test Reading for Academic Purpose Students of English Education Unisnu Jepara</td>
<td>496</td>
</tr>
</tbody>
</table>

*Firmanto Adi Nurcahyo

Note: The asterisk indicates a preliminary study.
Learning Evaluation Model Design with Multiple Choice Tests for Field Studies Exact Sciences  
*Nyenyep Sriwardani*  
502

Bhagavad Gita Video for Hinduism Education Lampung  
*Nyoman Siti, I Komang Arteyasa, Ni Made Indrayani*  
506

Development of Authentic Assessment Instrument at Grade Four Elementary School in Malang  
*Puri Selfi Cholilfah, Muhardjito, Eddy Sutadji*  
511

Model Employee Performance Evaluation of Economics Graduate Degree in Bali  
*Putri Anggreni*  
517

Hypothetical Model Development of Electrical Torso Learning Media Circulation System for Students Skill Formation of Critical Thinking and Scientific Attitude Senior High School in Lampung Timur  
*Ririn Noviyanti, Sisca Puspita Sari Nasution*  
523

Developing a Creative Thinking Assessment Model for Kindergarten Teachers  
*Risky Setiawan*  
531

Indonesian Adaptation Scale for Job Content Questionnaire (JCQ)  
*Sandra Jati Purwantari*  
539

Development of Assessment Instruments of Art Painting Production Integrated With Character for Assessing Learners’ Field Work Practice in Vocational High School  
*Trie Hartiti Retnowati, Djemari Mardapi, Bambang Prihadi*  
546

Analyzing the Quality of English Test Items of Daily, Mid Semester and Final School Examinations in Bandar Lampung: (Assessment and Evaluation in Language Teaching)  
*Ujang Suparman*  
556

Developing A Pedagogical Commitment Instrument  
*Wasidi*  
567

Adaptation and Construct Validation of the Indonesian Version of the Utrecht Work Engagement Scale  
*Yulia*  
574

**IV. Sub Themes:**

- **Program Evaluation for Improving Quality of Education**

  The Effectiveness of The Boarding Teacher Professional Development Program: an Approach of Process Evaluation  
  *Friyatmi*  
  579

  The Effect of Formative Test Types and Attitudes toward Mathematics on Learning Outcomes  
  *Hari Setiadi, Sugiarito, Rini*  
  584

  An Evaluation Model of Character Education in Senior High School  
  *Hari Sugiharto, Djemari Mardapi*  
  591
An Evaluation on the Implementation of Lesson Plans for Early Childhood Education Center (PAUD) Located Around IAIN Surakarta

Hery Setiyatna

The Effect of Cooperative Learning Model Type Group Investigation with Self Assessment Reinforcement and Learning Interest toward the Physics Learning Result of Students at Grade XI SMA Negeri 1 Watubangga Kolaka

I Gede Purwana Edi Saputra, H.M. Sidin Al

Effect of Cognitive and Emotif Techniques in Counseling Rational Emotif Behavior Therapy toward Tendency Aggressive Behavior Based on Type of Personality Among Students of SMP Negeri 4 Denpasar

I Wayan Susanta

THE EVALUATION OF THE SCHOLARSHIP DEGREE PROGRAM FOR THE ISLAMIC RELIGIOUS EDUCATIONAL TEACHERS AT SCHOOL

Ju’subaidi

The Influence of Teacher Pedagogical Competence and Emotional Intelligence towards Motivation and Physics Learning Result of Student at XI IPA Grade SMA Negeri 1 Watansoppeng

Kaharuddin Arafah, Adnani Yuni, Muris

Evaluating Policy Implementation Indicators in Decentralized Schools

Lilik Sabdaningtyas, Budi Kadaryanto

Identification Critical Thinking Skills of SMA Muhammadiyah 1 Banjarmasin Students to the Matter Dynamic Electricity

Misbah, Salijdah Mahtari, Sayid Muhammad Hasan

The Influence of the Socio-Cultural-Based Learning Device to Student Academic Performance

Muhammad Nur Wangid, Ali Mustadi

The Influence of Teacher Professional Competence and Interpersonal Intelligence Towards Motivation and Physics Learning Result of Student at XI MIA Grade Sma Negeri 1 Pangkajene

Murniay M, Kaharuddin Arafah, Subaer

Evalutation Study to Career Guidance Service-Program of Vocational High Schools in Banjarmasin

Nina Permatasari, Djaali, Ma‘ruf Akbar

Cipp Evaluation of The Learning in Cultural Dialogue During Unsoed Intercultural Summer-Camp

Oscar Ndayizeye, Agrégé TEFL

Evaluating Basic English Test Items for Non-English Students from Teachers Perspectives

Prihantoro

Is the German Language Text Too Short for the Senior High School Students?

Ryan Nuansa Dirga, Primardiana Hermilia Wijayati
Evaluation of Managerial Leadership Ability of Senior High School Headmasters in Sleman
Sabar Budi Raharjo, Lia Yuliana

Evaluation of Social Attitude Core Competence (KI-2) Implementation in State Elementary School in Yogyakarta
Siti Aminah, Yulian Sari

The Evaluation of The Foreign Language Intensification Program for the Students of UIN Allauddin Makassar
Sitti Mania

Evaluation of the Civilizing Moral Character Implementation in Elementary School
Sulthoni

The Evaluation of 2013 Curriculum Implementation on Thematic Integrative toward Math Subject for Elementary School In East Lombok
Syukrul Hamdi
The Effect of Number of Common Items on the Accuracy of Item Parameter Estimates with Fixed Parameter Calibration Method

Dina Huriaty
Mathematics Education Study Program, STKIP PGRI Banjarmasin
E-mail: dina_rty@yahoo.co.id

Abstract—This study aims to determine the effect of number of common items on the item parameter estimates accuracy in fixed parameter calibration methods on the assigned underlying ability distribution. This research is a simulation that was developed based on the item response theory with a three-parameter logistic model. There are the factors in simulation study: (1) three number of common items, namely 10, 20 and 30; and (2) three methods of calibration OWU-OEM (one prior weights updating and one expectation-maximization cycle), OWU-MEM (one prior weights updating and multiple expectation-maximization cycles), and MWU-MEM (multiple weights updating and multiple expectation-maximization cycles), (3) 500 samples for target groups, and (4) the distribution for target group, normal ability distribution. Base group generated to get the parameters of common-items. Twenty-five replications were generated for each condition. The accuracy of method parameter estimation measured by RMSE and absolute-Bias. The research results showed that on normal distribution, the increase of the number of common-items have effect on the accuracy of item parameter estimates on the three fixed parameter calibration method. The small RMSE that obtained of item parameters estimates are not always followed by a small absolute-biased. OWU-MEM method is the most accurate in estimating the parameters in the model with the number of common-items 30.

Keywords: common-items, estimation, calibration, fixed-parameter

I. BACKGROUND

When setting the matter for a test, developers can use the items that have been calibrated and added the new test items. In these circumstances, the problem that arises is how to put a new item parameter or parameters item in the previous test, whether it needs to be placed on a items scale that have been calibrated or on a new scale. How to put the estimation parameter from two separate groups into the same scale, can be done by calculating the estimated parameters for each group and then change the scale by using common-items. This can be done through the calibration process. These items which have been calibrated placed as a common-items on new test.

In study by Kim (2006), fixed calibration parameters are divided into five calibration methods. These method is distinguished by the number of ability prior that updated during EM cycle and number of EM cycles is used. These five methods are (1) NWU-OEM (no prior weights updating and one expectation-maximization cycle), (2) NWU-MEM (no prior weights updating and multiple expectation-maximization cycles), (3) OWU-OEM (one prior weights updating and one expectation-maximization cycle), (4) OWU-MEM (one prior weights updating and multiple expectation-maximization cycles), dan (5) MWU-MEM (multiple weights updating and multiple expectation-maximization cycles). According to Kim, MWU-MEM method is a accurate method in the third set distribution capabilities. On normal ability distribution N (0,1), NWU-MEM and OWU-MEM showed accurate performance on estimating parameter.

Dina & Mardapi (2014) study showed that of the five fixed parameter calibration methods, OWU-OEM is the most accurate method to estimate the parameters of the test item on the national exam in Mathematics.

Based on the results of these two studies and the factors considered in the study of this simulation, this study will only be compared three methods of fixed parameters calibration, namely OWU-OEM (one prior weights updating and one expectation-maximization cycle), OWU-
MEM (one prior weights updating and multiple expectation-maximization cycles), dan MWU-MEM (multiple weights updating and multiple expectation-maximization cycles).

Estimate parameters of the common items, then placing it as a fixed parameter, items calibration and then comparing the level of accuracy in estimating the parameters are the background in this study. This study aimed to determine the effect of number of common items on the parameter estimates accuracy in fixed parameter calibration methods on the assigned underlying ability distribution.

The results of this study are expected to contribute to the development of theory in the field of measurement and testing, particularly for the development of item response theory. The results of this study can provide information about fixed parameter calibration process for various applications such as test score equating and items calibration on test development.

II. RESEARCH METHODS

This study is a simulation. The simulation carried out by conditioning as the real situation on the field. A simulation study is conducted to compare the accuracy of several different models in which analytically settlement can not be obtained easily (Harwell et al., 1997). A simulation study of the generation data involves factors that allegedly influence the accuracy of estimation parameter in this study, which is the number of common items on the capability distribution set. The sample size used in this simulation study is 500 samples for the target-group. The sample size is following minimal sample requirements if using three-parameter logistic model.

In this study, logistics model chosen is the three-parameter logistic model. It is based on the fact that the ability tests that use multiple choice format such as the format of the exam is an example of a situation where the three-parameter logistic model is suitable. Three-parameter logistic model is expressed as follows.

\[
P_i(\theta) = c_i + (1 - c_i) \frac{e^{D_i(\theta - b_i)}}{1 + e^{D_i(\theta - b_i)}},
\]


The ranges for item parameters on three parameter logistic model to estimate the item parameter are as follows. Item difficulty index \(b\) can be a real number, but normally estimate \(b\) is in the range of -3.0 and 3.0 (Hulin, Drasgow & Parsons: 1983, p. 35-36), while according to Hambleton, Swaminathan, & Rogers (1991, p. 13) item difficulty parameter index are in the range -2.0 and 2.0. Item discrimination index \(a\) has a value in the interval 0.30 to 2.0 (Hulin, Drasgow & Parsons: 1983, p. 35-36), while according to Hambleton, Swaminathan, & Rogers (1991, p. 15), item discrimination index on the scale (-\(\infty\), + \(\infty\)). Very rare to find item discrimination parameters index are more than +2. So that the item discrimination parameters are in the interval (0, 2). Asimptut restricted on the interval 0 < c <1. Estimates of pseudo-guessing index \(c\) are often close to 1 / m for items on the multiple choice test, where m is the number of choices (Hulin, Drasgow & Parsons: 1983, p. 35-36). Based on the limitations of the parameter index, this study set limitation for the item difficulty parameter index is [-3, 3], item discrimination parameter index is [0.3, 2] and pseudo-guessing parameter index is \(\leq 0.25\).

There are no guidelines on the number of common items on a test. Kolen & Brennan (1995, p. 248) provides a guide to get the accurate results in the dichotomous model of a single test, the necessary the number of common items at least 20% of the overall number of test items. In this simulation study defined common items are 10, 20, and 30 from a total of 40 rounds on each device that simulated or are respectively 25%, 50% and 75% common items. The distribution considered in this simulation is a normal ability distribution \(N(0, 1)\).

III. RESEARCH DESIGN SIMULATION

The factors that considered in the fixed parameter calibration simulation consisting of three types of common items are 10, 20, and 30; the ability distribution for target groups, namely the normal ability distribution; and three methods of fixed parameters calibration, are OWU-OEM (one prior weights updating and one expectation-maximization cycle), OWU-MEM (one prior weights updating and multiple expectation-maximization cycles), and MWU-MEM (multiple weights updating and expectation-maximization multiple cycles). Based on these factor, there are
3 × 3 = 9 conditions. Base-group raised to get the common-items parameters which then set as a fixed parameter.

Harwell (1997) considered that the use of multiple replication is the most appropriate way to conduct simulation studies on item response theory, but Cohen, Kane, & Kim (2001) argues that it is not necessary to have a very large number of replication. A simulation study for item response theory applied only a small amount of replication, which is at least 10 times replication. In this study, each condition performed 25 times replication. Generation of response data is done by Wingen 2.

Parameter estimation accuracy of each replication on a simulation study were evaluated using criteria of RMSE (Root Mean Squared Error) (Kaskowitz & de Ayala, 2001; Kirisci, Hsu, & Yu, 2001; Li & Lissitz, 2000) and the absolute-bias. The smaller the RMSE, the more accurate the methods used compared to other methods. Beside RMSE, absolute-bias (a-bias) is also an indicator in assessing the accuracy of the calibration method. The smaller the absolute-bias indicate more accurate method to estimate the parameters. The actual parameter values on simulation study are obtained from the output of Wingen 2 (WGI and WGE). RMSE mathematically expressed by the following equation.

\[
RMSE = \sqrt{\frac{1}{n} \sum_{j}^{n} (\hat{X}_{ij} - X_i)^2}
\]

Mean square error (MSE) consists of two components, which is a measure of estimator variability (precision) and bias (accuracy). Bias of an estimator \( \theta \) of the parameter \( \theta \) is the difference between the expected value of \( \theta \) and \( \theta \) or bias = \( E (\theta) \) - \( \theta \). Estimator which have bias index equal to zero is called an unbiased estimator, or \( \theta \) is an unbiased estimator of \( \theta \) if MSE is equal to the variance of \( \theta \). Bias can be negative and positive. Negative bias indicates underestimated parameters and a positive bias shows over-estimated parameters. Bias is mathematically expressed as follows.

\[
Bias = \frac{1}{n} \sum_{j}^{n} (\hat{X}_{ij} - X_i)
\]

In this study bias value used is the absolute bias index, to be denoted by a-bias, which is expressed as follows.

\[
a\text{-bias} = \left| \frac{1}{n} \sum_{j}^{n} (\hat{X}_{ij} - X_i) \right|
\]

Software that is used as estimation tool in this study is Parscale. Parscale estimate the parameters of the model responses by Marginal Maximum Likelihood (du Toit: 2003, p. 611). Parscale used to calibrate the fixed parameters by the EM algorithm and can updates the prior ability distribution (Taehoon & Petersen, 2009 and Kim, 2006).

IV. PROCEDURE

Fixed parameter calibration procedure begins by setting the item parameters as common items. Common items Parameter are estimated on the base group, then placed as fixed parameters during the calibration process on the target group. Here is the procedure for each method.

1. In OWU-OEM method, a single cycle of EM applied once to new item when the initial prior ability has been updated once. Posterior value calculated after one EM cycle. As NWU-OEM method, in this method only common items that are used to calculate conditional likelihood for each testee.

2. The OWU-MEM method estimate the parameters in the new items through multiple cycles of EM, which way to update ability distribution once time. The first EM cycle run in a similar step to the NWU-OEM method, and the value of updated prior will used in the second cycle and cultivated it does not change during the next EM cycle. The initial prior ability renewed
one time after the last cycle of EM. As NWU-MEM method, both common items and new items are used to estimate the distribution of ability. The maximum number of EM cycles is set as 2000.

(3) The MWU-MEM method update prior ability distribution and find the maximum likelihood estimate on the new item parameters iteration. The first of EM cycle is first performed in the same step as in the method of NWU-OEM. The next step is to update the default value. In the second cycle of the EM, both the item and the ability parameter value are estimated together. After the last cycle of EM, the posterior latent value is obtained. MWU-MEM method uses all the information from the response of new items that use to obtain latent ability distribution and estimate the parameters. The maximum number of EM cycles is set as 2000.

V. RESULTS AND DISCUSSION

The RMSE and absolute-Bias on the factors considered in this study are presented in Table 1.

Table 1. The RMSE and absolute-Bias Index

<table>
<thead>
<tr>
<th>Item Parameter</th>
<th>CI</th>
<th>OWU-OEM</th>
<th>OWU-MEM</th>
<th>MWU-MEM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>RMSE</td>
<td>a-Bias</td>
<td>RMSE</td>
</tr>
<tr>
<td>b</td>
<td>10</td>
<td>0.2384</td>
<td>0.0209</td>
<td>0.2476</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.2340</td>
<td>0.0299</td>
<td>0.2402</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.2234</td>
<td>0.0399</td>
<td>0.2184</td>
</tr>
<tr>
<td>A</td>
<td>10</td>
<td>0.2329</td>
<td>0.1070</td>
<td>0.2317</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.2328</td>
<td>0.1313</td>
<td>0.2299</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.2316</td>
<td>0.1612</td>
<td>0.2244</td>
</tr>
<tr>
<td>C</td>
<td>10</td>
<td>0.1084</td>
<td>0.0528</td>
<td>0.1071</td>
</tr>
<tr>
<td></td>
<td>20</td>
<td>0.0965</td>
<td>0.0426</td>
<td>0.0951</td>
</tr>
<tr>
<td></td>
<td>30</td>
<td>0.0883</td>
<td>0.0399</td>
<td>0.0864</td>
</tr>
</tbody>
</table>

On normal ability distribution N (0, 1), the increase of the number of common-items affect to the decrease value of the b-RMSE on the three methods. The smallest RMSE value appears on OWU-MEM on 30 common-items model. This shows that the increase of the number of common items increases the accuracy of the item difficulty parameter estimation on the three fixed parameter calibration methods.

MWU-MEM method has the smallest a-bias of the item difficulty parameter on the 10 common items model; while on 20 and 30 common items model, the OWU-OEM method has smallest a-bias of the item difficulty parameter. The increase number of common items does not affect on absolute bias indeks decrease of each method. This is shown in Figure 1.

Figure 1. RMSE and a-Bias of Item Difficulty Parameter
The figure 2 shows that on normal ability distribution N (0.1), the increase of the number of common items give an effect to the a-RMSE on OWU-OEM, OWU-MEM, and MWU-MEM method. All methods on sample size of 500 model indicates that if the number of common items increase then a-RMSE will decrease.

As well as a-bias of item difficulty parameters estimates, on a a-bias of item discrimination index shows that the increase the number of common items will increase a-bias indeks. OWU-OEM method has the smallest a-bias index on 10 common items model. OWU-MEM method is the most accurate method to estimate the item discrimination parameters on 30 common items model.

The figure 3 shows that on the three fixed-parameter calibration methods, c-RMSE is getting smaller as the number of common items increase. OWU-MEM method has the smallest RMSE on 30 common items model. A-bias of pseudo-guessing on all methods are getting smaller as the number of common items increase. MEM OWU method is also the most accurate method to estimate pseudo-guessing parameters based on a-bias index.

This results lead to the conclusion that number of common items can increase the accuracy of item estimation parameter on all fixed parameter calibration methods. These results is relevance with Kim (2006) study, which is the value of RMSE-b and RMSE-a tends decreased when common items increased. According Taehoon & Petersen (2009), when the number of common items increased, each MSE of method deacreased dramatically.

The results of this study showed different results with the results of the study on real data has been conducted by Dina & Mardapi (2014) which the OWU-OEM method is the most accurate method to estimate the parameters. This can be explained that in the real data where only 6 common items, then when analyzed against the accuracy of the method on a model with a common item 10, it appears that the method OWU-OEM shows the smallest RMSE index of item difficulty and a-bias item discrimination. From these results it can be concluded that in models
with a small number of common items, OWU-OEM methods is an accurate method to estimate the parameters.

VI. CONCLUSIONS AND RECOMMENDATIONS

Based on the results of research and discussion, we concluded that the results showed in normal distribution, increasing the number of common-items affect the accuracy of item parameter estimation on three fixed parameter calibration methods. Small RMSE which obtained from item paramete estimation are not always followed by a small a-bias index. OWU-MEM method is the most accurate method to estimate the parameters in 30 common-items.

The implication of this study can be a benchmark in the development of an item bank. The results provide information to developers about how to obtain calibrated test items using a variant of fixed parameter calibration method. The study provides some alternative methods in the process of calibration items and place a number of common items that can improve the accuracy of item estimation parameter which developed for the benefit of the implementation of the test. Those items that have been calibrated can be used to develop an item bank. The existence of item bank can make it easier for the planning and administration the next test.

Limitations in this study are shown in some circumstances, small RMSE values are not followed by a small a-bias, so that the accuracy of the determination of the criteria of a method based on a biased RMSE and smallest by three parameter logistic model can not be met strictly. Prototype results of this simulation study can be a reference in the selection method that can estimate the parameters of the desired items accurately.

BIBLIOGRAPHY

understood and can implement the 2013 Curriculum properly. It is suggested that the government and policy makers should give special attention to the input aspect, especially in terms of facilities, and help the parents to understand the instructional activities carried out in the implementation of the 2013 Curriculum, especially in rural elementary schools, through various programs involving participation of parents, school components and the community in the hope that there will be no gap between the regency town, the district town and the rural area in terms of inputs and outputs.

REFERENCES