Sertifikat

Kutipan dari Keputusan Menteri Pendidikan dan Kebudayaan Republik Indonesia
Nomor: 040/P/2014, Tanggal 14 Februari 2014
Tentang Hasil Akreditasi Terbitan Berkala Ilmiah
Periode II Tahun 2013

Nama Terbitan Berkala Ilmiah
Jurnal Penelitian dan Evaluasi Pendidikan
ISSN: 1410-4725
Penerbit: Himpunan Evaluasi Pendidikan Indonesia (HEPI)
bekerjasama dengan Program Pascasarjana Universitas Negeri Yogyakarta

Ditetapkan sebagai Terbitan Berkala Ilmiah

TERAKREDITASI

Akreditasi sebagaimana tersebut di atas berlaku selama 5 (lima) tahun sejak ditetapkan.

Jakarta, 18 Februari 2014
Direktur Penelitian dan Pengabdian Kepada Masyarakat,
Direktorat Jenderal Pendidikan Tinggi

Prof. Agus Subekti, M.Sc., Ph.D.
NIP. 19600801 198403 1 002

Jurnal
Penelitian dan Evaluasi Pendidikan

HIMPUNAN EVALUASI PENDIDIKAN INDONESIA (HEPI)
in cooperation with
GRADUATE SCHOOL OF UNIVERSITAS NEGERI YOGYAKARTA
Kampus Karangmalang, Yogyakarta 55281. Phone. 0274 550836 Fax : 0274 520326

Website: http://journal.uny.ac.id/index.php/jpep
e-mail: jurnalhepi@uny.ac.id

9 772338 606001

1410 4725

Jurnal Penelitian dan Evaluasi Pendidikan

Volume 21, No 1, June 2017

Volume 21, No 1, June 2017

ISSN 1410-4725 (print)
ISSN 2338-6061 (online)
Akreditasi No. 040/P/2014

Jurnal
Penelitian dan Evaluasi Pendidikan

Jurnal
Penelitian dan Evaluasi Pendidikan

Publisher
HIMPUNAN EVALUASI PENDIDIKAN INDONESIA
In Cooperation With
PROGRAM PASCASARJANA UNIVERSITAS NEGERI YOGYAKARTA
(MOU Nomor 195 B/J.35.17/LK/04)

Correspondence: Graduate School of Universitas Negeri Yogyakarta
Kampus Karangmalang, Yogyakarta, 55281, Telp. (0274) 550835, Fax. (0274) 520326
Homepage: http://journal.uny.ac.id/index.php/jpep
e-mail: jurnalhepi@uny.ac.id

FOREWORDS

We are very pleased that *Jurnal Penelitian dan Evaluasi Pendidikan* is releasing its issue **Volume 21, No 1, June 2017.** We are also very excited that the journal has been attracting papers from many institutions in Indonesia. *Jurnal Penelitian dan Evaluasi Pendidikan* was first published in **1998** and since then regularly published online and in print twice a year: June and December.

Jurnal Penelitian dan Evaluasi Pendidikan with ISSN 1410-4725 (*printed*) and ISSN 2338-6061 (*online)* has been **re-accredited** by Indonesian Ministry of Education and Culture decision Number 040/P/2014 which is valid for 5 (five) years since enacted on 18 Februari 2014.

Jurnal Penelitian dan Evaluasi Pendidikan is a showcase of original, rigorously conducted educational evaluation, measurement and assessment from primary, secondary, and higher education institutions. Each issue of this journal is not limited to comprehensive syntheses of studies towards developing new understandings of educational evaluation, measurement and assessment only, but also explores scholarly analyses of issues and trends in the field.

Yogyakarta, June 2017

Editor in Chief

Table of Content

Jurnal Penelitian dan Evaluasi Pendidikan
Volume 21, No 1, June 2017 (1-10)
Online: http://journal.uny.ac.id/index.php/jpep

IMPLEMENTASI MODEL PEMBELAJARAN MATH-SCIENCE BERBASIS PERFORMANCE ASSESSMENT UNTUK MENINGKATKAN KEMAMPUAN BERPIKIR KRITIS SISWA DI DAERAH PERKEBUNAN KOPI JEMBER

Suratno [1]*, *Dian Kurniati* [1]
[1]FKIP Universitas Jember
[1]Jl. Kalimantan No. 37, Krajan Timur, Sumbersari, Jember, Jawa Timur 68121, Indonesia
* Corresponding Author. Email: suratno.fkip@unej.ac.id

Abstrak

Tujuan penelitian ini adalah untuk mengetahui peningkatan kemampuan berpkir kritis siswa kelas V SD di sekitar perkebunan kopi Garahan Jember melalui penerapan model pembelajaran math-science berbasis performance assessment. Kemampuan berpikir kritis dalam penelitian ini adalah kemampuan pembuktian, kemampuan generalisasi, dan kemampuan pemecahan masalah. Data dianalisis dengan pendekatan kuantitatif dan kualitatif. Uji coba penelitian ini diterapkan pada dua SD di sekitar perkebunan kopi yaitu MI Al –Amin Garahan dan SD Negeri Sidomulyo 03 Jember dengan subyek penelitian sebanyak 80 siswa. Data diperoleh dari hasil kinerja siswa selama mengerjakan post test pada materi math-science dan wawancara. Pada siklus pertama terdapat 8 siswa (10%) yang memenuhi semua indikator kemampuan generalisasi dan pembuktian, sedangkan kemampuan pemecahan masalah belum berkembang dengan maksimal. Pada siklus kedua terdapat peningkatan, yaitu terdapat 22 siswa (27.5%) yang mampu memiliki kemampuan pembuktian dan kemampuan generalisasi. Pada siklus ketiga terdapat 32 siswa (40%) yang mampu memiliki semua kemampuan berpikir kritis. Berdasarkan hasil penelitian, dapat disimpulkan bahwa terdapat peningkatan kemampuan berpikir kritis siswa.

Kata kunci: *math-science, performance assessment, kemampuan berpikir kritis*

THE IMPLEMENTATION OF MATH-SCIENCE LEARNING MODEL BASED ON PERFORMANCE ASSESSMENT TO IMPROVE STUDENTS' CRITICAL THINKING SKILL IN JEMBER COFFEE PLANTATION AREA

Abstract

The research is aimed t knowing the improvement of the critical thinking skills of the fifth grade students in Garahan Coffee plantation area through the implementation math-science learning model based on performance assessment. In this research, the critical thinking skills include proof skill, generalization skill, and problem solving skill. Data were analized by quantitative and qualitative approach. This research was implemented to two elementary schools in Garahan coffee plantation area, namely MI Al-Amin Garahan and SD Negeri Sidomulyo 03 Jember, with 80 students as the subject. Data were obtained from the result of students' post test on math-science and interview. In the first trial, there are 8 students (10%) who comply with all indicators of generalization and proof skills, while the problem solving skill has not optimally been developed. The result of the second trial has increased to 22 students (27.5%) who have proof and generalization skills. The third trial shows that 32 students (40%) comply with all indicators of critical thinking skills, namely proof, generalization, and problem solving skills. Based on the research findings, it can be concluded that there is improvement on students' critical thinking skill.

Keywords: *math-science, performance assessment, critical thinking skill*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v21i1.11799

Pendahuluan

Indonesia merupakan negara terbesar urutan ketiga penghasil kopi setelah Brazil dan Vietnam. Berdasarkan data statistik di Indonesia, jember merupakan produsen terbesar di Jawa Timur Indonesia. Produksi kopi di jember mencapai 3.105 tons pada tahun 2014 dan akan meningkat 18% setiap tahunnya (BPS Provinsi Jawa Timur, 2016). Berdasarkan data tersebut, pembelajaran kontekstual yang dapat memfasilitasi siswa untuk berpikir kritis untuk anak usia sekolah merupakan suatu keharusan agar produktivitas daerah perkebunan kopi di Jember dapat dioptimalkan.

Kemampuan berpikir kritis dapat diasah dengan cara membiasakan siswa terlibat secara aktif dalam menyelesaikan permasalahan yang membutuhkan kemampuan berpikir kritis siswa. Salah satunya dengan melibatkan kemampuan kinerja secara maksimal yang dimiliki oleh siswa. Akan tetapi, kenyataan di lapangan menunjukkan hal yang berbeda. Berdasarkan hasil wawancara peneliti dengan beberapa guru matematika dan IPA SD di Sidomulyo Jember dan MI AL-Amin Garahan, dapat dinyatakan bahwa siswa cenderung pasif dalam menyelesaikan permasalahan yang lebih kompleks. Siswa kurang tertarik untuk membuktikan suatu prinsip atau konsep, kurang tertarik untuk melakukan penyelidikan dan penggeneralisasian, serta kurang tertarik dalam menyelesaikan soal yang non rutin. Selain itu, berdasarkan penelitian sebelumnya bahwa kemampuan kinerja siswa sekolah dasar di sekitar perkebunan kopi Garahan Kabupaten Jember dalam menyelesaikan permasalahan *math-science* tergolong dalam level *Apprentice* (80%), level *Practititioner* (15%) dan level *Novice* (5%), sedangkan tidak terdapat siswa yang tergolong dalam level *expert* (Suratno & Kurniati, 2017). Sehingga dapat dikatakan bahwa kemampuan kinerja siswa SD di Garahan tergolong rendah dan sedang. Siswa lebih suka menyelesaikan permasalahan yang rutin bukan yang non rutin dengan tema lingkungan sekitar mereka. Selain itu, guru juga mengalami kesulitan dalam menyusun sebuah tugas dan instrumen yang akan diberikan kepada siswanya khususnya tugas yang terkait dengan kinerja yang dihubungkan dengan lingkungan mereka. Tugas yang dimaksud berupa rangkaian perintah yang meminta siswa untuk menyelesaikan permasalahan nyata yang berhubungan dengan lingkungan sekitarnya dengan melibatkan semua kemampuan kinerja yang dimilikinya. Permasalahan tersebut berkaitan dengan materi pembelajaran matematika dan IPA. Oleh karena itu perlu diterapkan suatu pembelajaran *math-science* yang berbasis pada kinerja siswa dengan menggunakan instrumen penilaian untuk mengukur kinerja siswa dalam meningkatkan kemampuan berpikir kritis siswa dalam menyelesaikan permasalahan yang kontekstual.

Dalam hal ini *performance assessment* adalah teknik mengases yang cocok untuk mengases kemampuan berpikir kritis siswa. Menempatkan siswa dalam situasi dunia nyata dan melibatkan siswa secara aktif untuk memaksimalkan kinerjanya melalui *performance Assessment* dalam menyelesaikan permasalahan matematika dan sains bukanlah hal yang mudah. Menyikapi hal ini, Wisconsin Education Association Council mengatakan bahwa guru dapat meminta siswa untuk menyelesaikan tugas yang berkaitan dengan dunia nyata yang disimulasikan dengan kinerja (Wisconsin Education Association Council, 1996). Sehingga untuk dapat melaksanakan *performance Assessment*, perlu untuk: (1) mendefiniskan konsep, pengetahuan dan kemampuan apa yang diases, (2) menentukan aktivitas kinerja yang akan ditunjukkan oleh siswa, (3) mengembangkan kriteria penilaian. Oleh karena itu siswa perlu diajak untuk tahu bagaimana sebuah tugas kinerja diases dan seperti apa cara menyelesaikan tugas tersebut yang seharusnya dilakukan dan tidak dilakukan melalui pencontohan *performance*.

Salah satu alternatif solusi terhadap masalah tersebut diatas adalah dengan menerapkan suatu model pembelajaran yang mampu meningkatkan kemampuan berpikir kritis siswa dalam menyelesaikan permasalahan atau tugas kinerja yang berkaitan dengan aplikasi matematika dan sains yang sesuai dengan lingkungan siswa. Model pembelajaran tersebut adalah Model Pembelajaran *math-science* berbasis *performance assessment* yang diharapkan mampu meningkatkan kemampuan berpikir kritis siswa SD khususnya di lingkungan perkebunan kopi.

Kemampuan berpikir kritis adalah suatu kemampuan yang dimiliki seseorang untuk menampilkan kemampuan kognitif dan disposisi intelektualnya yaitu untuk (1) mengidentifikasi secara efektif, analisis, dan evaluasi dari suatu argumen dan kebenaran, (2) menjelaskan pemikiran yang bisa dan prekonsepsi, (3) memformulasikan dan menampilkan alasan yang mendukung kesimpulan, dan (4) memberikan alasan yang logis dalam menyampaikan apa yang akan dilakukan (Bassham, 2011). Selanjutnya, terdapat tiga indikator berpikir kritis yaitu pembuktian, generalisasi, dan pemecahan masalah (Glazer, 2001). Pertama, kemampuan pembuktian adalah kemampuan untuk membuktikan suatu pernyataan secara deduktif (menggunakan teori-teori yang telah dipelajari sebelumnya). Adapun indikator kemampuan pembuktian yaitu (1) mampu menemukan kembali prinsip atau rumus matematika melalui uji coba, (2) mampu membuktikan kebenaran teori melalui pengamatan secara langsung, dan (3) mampu membuktikan penggunaan rumus matematika dalam menyelesaikan soal yang berkaitan dengan perkebunan kopi. Kedua, kemampuan generalisasi adalah kemampuan untuk menghasilkan pola atas persoalan yang dihadapi untuk kategori yang lebih luas. Indikator kemampuan generalisasi adalah (1) mampu menentukan pola bilangan berdasarkan pembuktian yang secara induktif dan (2) mampu menemukan pola umum yang ada di permasalahan pada tanaman kopi. Ketiga, indikator kemampuan pemecahan masalah adalah (1) kemampuan mengidentifikasi unsur yang diketahui, yang ditanyakan, dan memeriksa kecukupan unsur yang diperlukan dalam soal, (2) menyusun model matematika dan menyelesaikannya, serta (3) memeriksa kebenaran hasil atau jawaban.

Indikator berpikir kritis yang diterapkan pada penelitian ini diintegrasikan pada setiap permasalahan yang berkaitan dengan tema kopi melalui kinerja siswa yang ada pada *post test.* Model pembelajaran *math-science* berbasis *performance assessment* yang diterapkan pada pembelajaran IPA dan matematika di SD merupakan suatu model pembelajaran yang difokuskan pada kinerja siswa dalam menyelesaikan permasalahan sehari-hari.

Model pembelajaran *math science* berbasis *performance assessment* dalam meningkatkan kemampuan berpikir kritis dirancang untuk mampu mengajari siswa bagaimana memecahkan masalah matematika dan IPA yang terkait dengan perkebunan kopi dengan memodelkan/memberi contoh tentang bagaimana suatu kemampuan pemecahan masalah matematika dinilai. Untuk mendukung model ini diperlukan kelengkapan berupa instrumen untuk mengases kemampuan siswa berupa instrumen *performance assessment* yang meliputi: (1) *post test* sebagai alat penilaian untuk mengukur kognitif dan kinerja siswa, (2) rubrik penilaian sikap untuk menilai aspek afektif dan psikomotor siswa. Penggunaan instrumen harus dikomunikasikan dan dicontohkan terlebih dahulu kepada siswa. Pemilihan instrumen tersebut sesuai dengan teori yang menyatakan bahwa penilaian kinerja idealnya dilakukan melalui metode pengamatan langsung, sedangkan metode lainnya yang dapat dilakukan yaitu dengan metode simulasi komputer tes (Supahar & Prasetyo, 2015). Sehingga pada penelitian ini metode yang dipilih adalah metode tes yang dilakukan pada akhir pembelajaran yang disebut dengan *post test.* Dengan demikian, diharapkan siswa dapat meningkatkan kemampuannya untuk membangun kebiasaan berpikir secara disiplin dalam menghadapi masalah, mengetahui apa yang mereka perlukan untuk mengecek keakuratan, ketepatan dan kualitas pekerjaan mereka bahkan siswa dapat menilai sendiri pekerjaan mereka sebelum dikumpulkan kepada guru.

Tujuan pembelajaran tidak hanya ditekankan pada hasil belajar, tetapi lebih ditekankan pada proses yaitu bagaimana siswa menyelesaikan soal pemecahan masalah dengan kritis yang diberikan dengan caranya sendiri. Dengan mengerjakan soal dalam *post test math-science* yang bersifat *uncued world problem* secara individu, siswa diberi kesempatan untuk menyelesaikan soal dengan caranya sendiri. Masalah *uncued world problem* merupakan masalah yang otentik yang sangat mungkin ditemui dalam kehidupan sehari-hari, sifat *uncued* dari masalah diharapkan dapat membuat siswa tertantang untuk menyelesaikan dengan cara mereka sendiri (ketika siswa mencoba memecahkan masalah secara individu). Pembelajaran ini mempunyai kemungkinan sangat besar untuk dapat meningkatkan kemampuan pemecahan masalah siswa.

Konstruksi pengetahuan dalam model pembelajaran *math-science* berbasis *performance assessment* dalam meningkatkan kemampuan berpikir kritis siswa terjadi pada saat siswa bekerja untuk menyelesaikan soal-soal terkait dengan permasalahan di sekitar perkebunan kopi yang ada dalam *Math-Science performance task* baik ketika siswa menyelesaikan secara individu, berkelompok ataupun ketika kegiatan presentasi hasil pekerjaan dan pemberian contoh penilaian hasil pemecahan masalah. Soal-soal dalam *Math-Science performance task* dibuat sedemikian rupa sehingga siswa dapat mengkonstruk pengetahuannya sendiri. Adapun sintaks dari model pembelajaran *math-science* berbasis *performance assessment* adalah (1) Pra-pembelajaran, (2) orientasi, (3) pemecahan masalah secara individu, 4) pengorganiasian secara kelompok, 5) diskusi kelompok, (6) diskusi kelas, 7) pemberian contoh penilaian *performance*, (8) evaluasi, dan (9) pasca pembelajaran. Adapun secara rinci setiap sintaks tersebut disajikan pada Tabel 1.

Berdasarkan uraian tersebut, penelitian ini bertujuan untuk mengetahui peningkatan kemampuan berpkir kritis siswa kelas V SD di sekitar perkebunan kopi Garahan Jember melalui penerapan model pembelajaran *math-science* berbasis *performance assessment.*

Tabel 1. Sintaks Model Pembelajaran *Math-Science* Berbasis *Performance Assessment*

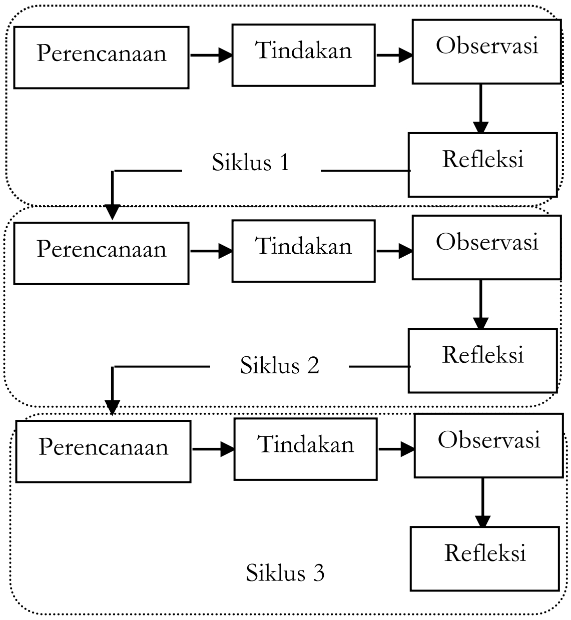
Sintaks	Aktivitas Siswa
Pra-pembelajaran	Mengerjakan soal pre-tes
Orientasi	▪ mendengarkan penjelasan guru, ▪ menjawab atau mengerjakan soal jika ada pertanyaan atau soal prasyarat yang disampaikan oleh guru ▪ Siswa membuat catatan, ▪ Siswa menerima soal tes dan rubrik ▪ Siswa bertanya jika ada penjelasan guru yang belum dimengerti dan menjawab pertanyaan yang disampaikan guru
Pemecahan masalah secara Individu	▪ Siswa secara individu mengerjakan *tes* dengan mengacu pada rubrik. Untuk menyelesaikan masalah, siswa dapat langsung mengerjakan di lembar tes
Pengorganisasian Kelompok	▪ Menempatkan diri dalam kelompok heterogen ▪ Membaca dan memahami tes
Diskusi kelompok	Siswa secara kelompok saling tukar pendapat dalam mengerjakan kembali tes dengan mengacu pada rubrik
Diskusi Kelas	Beberapa perwakilan kelompok menyajikan hasil diskusi kelompoknya
Pemberian contoh Penilaian *performance*	Mendengarkan penjelasan guru, membuat catatan, menerima soal tes dan rubrik, bertanya jika ada penjelsan guru yang belum dimengerti dan menjawab pertanyaan yang disampaikan guru
Evaluasi	Melakukan evaluasi hasil belajar
Pasca-Pembelajaran	Mengerjakan soal latihan

Metode Penelitian

Penelitian ini bertujuan untuk mengetahui peningkatan kemampuan berpikir kritis siswa kelas V SD di sekitar perkebunan kopi Garahan Jember melalui penerapan model pembelajaran *math-science* berbasis *performance assessment.* Data pada penelitian dianalisis dengan pendekatan kuantitatif dan kualitatif. Data diperoleh dari hasil kinerja siswa dalam menyelesaikan *post test* materi *math-science* dengan tema kopi dan hasil wawancara dengan siswa. Pendekatan kuantitatif digunakan untuk menentukan persentase ketuntasan siswa dalam mengerjakan *post test* dengan mengacu pada indikator berpikir kritis. Sedangkan, pendekatan kualitatif digunakan untuk menganalisis hasil wawancara siswa yang mengacu pada hasil pengerjaan *post test.* Instrumen yang digunakan pada penelitian ini adalah *post test* berbasis kinerja dengan tema kopi beserta rubrik penilaian kinerjanya dan lembar pedoman wawancara.

Penelitian ini diujicobakan pada 2 (dua) Sekolah Dasar di sekitar perkebunan kopi Garahan Jember, yaitu MI Al-Amin Garahan dan SD Negeri Sidomulyo 03 Jember dengan subjek sebanyak 80 siswa kelas V. Fokus penelitian ini adalah peningkatan kemampuan berpikir kritis siswa kelas V SD di sekitar perkebunan kopi garahan Jember, dengan indikator berpikir kritis yang digunakan adalah (1) kemampuan pembuktian, (2) kemampuan generalisasi, dan (3) kemampuan pemecahan masalah. Penelitian ini dilakukan dengan beberapa siklus hingga mencapai peningkatan kemampuan pemecahan masalah yang maskimal dari kondisi siswa.

Tema yang diberikan pada *post test* berbasis kinerja siswa adalah pemanfaatan lahan perkebunan kopi untuk menghasilkan kopi yang berkualitas baik dan berkuantitas banyak. Tema tersebut ditinjau dari dua bidang studi yaitu matematika dan IPA. Permasalahan yang terkait dengan matematika adalah bagaimana siswa mampu menghitung luas kebun kopi secara keseluruhan, menentukan jarak kebun dengan jalan penghubung, menentukan banyaknya tanaman kopi yang dapat ditanam semaksimal mungkin dengan menentukan jarak antar pohon, serta jarak perkebunan kopi dengan irigasi. Sedangkan permasalahan yang terkait dengan IPA yaitu menentukan jarak ideal dari antar pohon kopi ditinjau dari kelayakan ketersediaan air dan cahaya, ideal ukuran dan umur batang pohon kopi yang ditanam, kesuburan tanah, serta kebersihan lingkungan perkebunan kopi.



Gambar 1. Prosedur Penelitian

Penelitian ini dilaksanakan sebanyak 3 (tiga) siklus, dan masing-masing siklus terdiri dari tahap perencanaan, tindakan, observasi dan refleksi (Herawaty, 2009). Pada tahapan perencanaan, peneliti bersama-sama dengan guru kelas V di sekitar perkebunan kopi Garahan menentukan permasalahan yang dituliskan di *post test* berbasis kinerja, menentukan rubrik penilaian, dan menentukan pedoman wawancara. Pada tahapan tindakan, dilakukan ujicoba instrumen dan rubriknya kepada siswa kelas V di dua Sekolah Dasar. Bersamaan dengan kegiatan uji coba, peneliti melakukan pengamatan terhadap aktivitas kinerja siswa ketika menyelesaikan *post test.* Tahapan terakhir yaitu melakukan refleksi terhadap hasil kinerja siswa dengan mengecek kesesuaiannya dengan jawaban dari *post test*, dan melakukan wawancara untuk mengecek keabsahan data dari hasil kinerja siswa. Refleksi pada siklus pertama digunakan sebagai dasar untuk menyusun instrumen kinerja pada siklus kedua, begitu juga hasil refleksi siklus kedua sebagai dasar untuk menyusun instrumen pada siklus ketiga. Secara jelas tahapan pada penelitian ini dapat dilihat pada Gambar 1.

Hasil Penelitian dan Pembahasan

Pada siklus pertama, kedua, dan ketiga dilakukan empat tahapan penelitian, yaitu perencanaan, tindakan, observasi, dan refleksi. Pada siklus pertama, perencanaan dilakukan oleh 2 orang tim peneliti dan 20 orang guru kelas V SD di sekitar perkebunan kopi Garahan Jember pada tanggal 3 September 2016 dengan mendesain *post test* yang berbasis kinerja siswa untuk materi matematika dan IPA beserta rubrik penilaiannya. Permasalahan yang ada di soal *post test* didesain sedemikian sehingga setiap sub indikator dari indikator berpikir kritis yang ditingkatkan pada diri siswa dapat diamati. Adapun permasalahan yang disampaikan di soal *post test* untuk siklus pertama yaitu terkait dengan mengitung luas lahan maksimal yang dapat ditanami pohon kopi dan menentukan jarak tanam antar pohon supaya pohon kopi dapat berkembang dengan baik. Selain soal *post test* berbasis kinerja yang disusun pada tahap perencanaan, juga menyusun rubrik dari penilaian kinerja siswa.

Tahapan kedua adalah tahapan tindakan, yaitu guru melakukan pembelajaran dan memberikan soal *post test* yang sudah dikembangkan pada tahap pertama ke setiap siswa kelas V baik di MI Al-Amin Garahan ataupun SD Negeri Sidomulyo 03 Jember pada tanggal 17 dan 24 September 2016. Bersamaan dengan kegiatan siswa menyelesaikan soal *post test* tersebut, guru melakukan pengamatan terhadap aktivitas kinerja siswa selama mengerjakan soal *post test*. Pengamatan difokuskan pada indikator dari berpikir kritis.

Pada siklus pertama, terdapat 8 siswa dari 80 subjek penelitian atau 10% siswa yang memiliki kemampuan berpikir kritis dengan memenuhi dua indikatornya yaitu kemampuan pembuktian dan kemampuan generalisasi. Adapun skor *post test* dari 8 siswa tersebut lebih atau sama dengan 80 (skor ≥ 80). Kecenderungan secara umum siswa belum mampu memahami implementasi pembelajaran matematika dan IPA terhadap kehidupan sehari-hari yang dihubungkan dengan lingkungan mereka. Siswa masih belum bisa menentukan luas lahan maksimal jika lahan tersebut tidak berbentuk segiempat. Alasan yang disampaikan karena belum ada rumus luas bangun datar selain segitiga dan segiempat. Padahal seharusnya siswa bisa membagi lahan tersebut menjadi bangun segiempat atau segitiga dan menerapkan rumus luasnya. Selain itu siswa juga belum memahami jarak tanam antar pohon kopi jika menghasilkan kopi yang berkualitas baik. Hal tersebut disebabkan karena siswa jarang mengikuti atau mempelajari ke orang tua mereka ketika menanam kopi. Pemikiran seperti itu dimiliki oleh 72 siswa dari 80 siswa sebagi subjek penelitian. Sehingga pada siklus pertama belum maksimal untuk kemampuan berpikir kritis siswa dihubungkan dengan kinerja mereka terhadap perkebunan kopi.

Akan tetapi terdapat 8 siswa yang mampu memiliki kemampuan berpikir kritis dengan memenuhi dua indikator yaitu kemampuan pembuktian dan kemampuan generalisasi. Siswa cenderung memiliki kemampuan pembuktian dan kemampuan generalisasi yaitu sub-indikator (1) menentukan prinsip dan konsep yang digunakan untuk pemecahan masalah dengan coba-coba, (2) melakukan pembuktian secara benar dengan mengacu pada formula matematika dan IPA yang ada, dan (3) menggeneralisasi bentuk umum dari suatu pola lahan perkebunan kopi. Kedelapan siswa mampu menentukan bangun datar yang dapat digunakan pada perhitungan luas lahan kopi, sehingga siswa mampu menentukan konsep bangun datar yang digunakan dan prinsip luas dari bangun datar tersebut. Siswa juga mampu melakukan pembuktian secara benar meskipun masih coba-coba dengan membandingkan kenyataan di lingkungan mereka. Selain itu, siswa juga mampu menggeneralisasi bentuk umum dari pola bangun yang beragam dan mampu menentukan jarak antar pohon supaya pohon kopi dapat berkembang dengan baik.

Hasil analisis dari 80 subjek penelitian pada siklus pertama dijadikan dasar untuk mengembangkan instrumen pada siklus kedua. Adapun hal-hal yang perlu diperbaiki ketika menyusun instrumen berbasis kinerja siswa yang dihungkan dengan materi matematika dan IPA tema kopi yaitu (1) meminta siswa untuk melakukan observasi awal ke lingkungan perkebunan mereka dengan menanyakan proses penanaman kopi dan pemanfaatan lahan yang akan ditanami kopi, (2) mencatat semua aktivitas terkait dengan penanaman kopi, (3) tugas kinerja yang akan diberikan pada siklus kedua sebaiknya dikerjakan secara berkelompok dengan saling berdiskusi antarkelompok, dan (4) membuat LKS dengan permasalahan yang berbeda untuk setiap kelompok meskipun satu tema.

Pada siklus kedua, peneliti dan guru melakukan diskusi untuk menyusun instrumen penilaian beserta rubriknya dan permasalahan yang disampaikan di *post test* berdasarkan hasil refleksi siklus pertama. *Post test* pada siklus kedua dikerjakan oleh siswa secara berkelompok dengan setiap kelompok beranggotakan 4 siswa. Setiap kelompok diminta melakukan observasi awal terhadap metode sederhana menanam kopi dan memanfaatkan lahan kopi semaksimal mungkin. Tugas yang dikerjakan berupa proyek dimana siswa diminta menyusun denah perkebunan kopi yang didalamnya terdapat irigasi berupa sumur, tempat berteduh, dan tempat pembuangan sampah, serta jarak antar pohon kopi supaya berkembang dengan baik.

Kegiatan perencanaan tersebut dilakukan di MI Al-Amin Garahan Jember pada tanggal 1 Oktober 2016. Instrumen yang dikembangkan pada siklus kedua diujicoba ke MI Al-Amin Garahan dan SD Negeri Sidomulyo 03 Jember pada tanggal 8 dan 15 Oktober 2016. Pada awalnya guru meminta siswa untuk mempresentasikan hasil observasi awal kepada teman satu kelas dan meminta kelompok lain untuk menanggapinya. Kemudian dilanjutkan dengan pemberian *post test* yang berbasis kinerja dan proyek. Dari 80 siswa atau 20 kelompok belajar siswa, terdapat 22 siswa yang mengalami peningkatan kemampuan berpikir kritisnya dibandingkan siklus pertama yaitu dengan skor peningkatan paling sedikit 15 poin. Sedangkan 58 siswa lainnya belum meningkat kemampuan berpikir kritisnya, artinya mereka masih tetap pada posisi seperti siklus pertama. Siswa yang belum memiliki kemampuan berpikir kritisnya disebabkan karena siswa masih berpikir bahwa irigasi untuk perkebunan kopi berupa sungai, padahal pada *post test* siswa diminta mendesain posisi sumur untuk membantu irigasi dari sungai. Selain itu siswa juga belum memahami konsep dari luas bangun datar dan kelilingnya.

Hal tersebut berbeda dengan kondisi kemampuan 22 siswa lainnya. Siswa tersebut mampu menyelesaikan permasalahan yang ada pada *post test* dengan baik meskipun masih tahap coba-coba. Siswa sudah memulai menghubungkan permasalahan yang diberikan guru dengan kehidupan nyata mereka. Siswa mampu (1) menentukan formula baru berdasarkan formula yang sudah dipelajari di matematika terkait dengan keliling dan luas bangun datar, (2) membuktikan kebenaran jawabannya berdasarkan hasil observasi langsung ketika di rumah, (3) menentukan rumus atau formula yang dapat digunakan menyelesaikan permasalahan yang ada di *post test*, (4) menemukan suatu prinsip pemecahan masalah yang berlaku secara umum berdasarkan hasil pembuktikan pada indikator pertama, dan (5) menentukan generalisasi dari kesamaan permasalahan yang diberikan guru di sekolah dengan pengalaman langsung di rumah. Berdasarkan data tersebut, maka pada siklus kedua kemampuan berpikir kritis siswa sudah mengalami peningkatan meskipun tidak maksimal.

Adapun hasil refleksi dari kegiatan pada siklus kedua yaitu (1) siswa belum terbiasa mengerjakan soal yang non-rutin dengan mengacu pada proyek, (2) siswa belum mampu menganalisis prinsip apa yang dapat digunakan untuk menyelesaikan masalah, (3) siswa belum mampu melakukan pengerjaan penyelesaian masalah secara runtut dan memberi alasan dari setiap langkah pengerjaannya, (4) siswa belum memahami kebermanfaatan kebersihan lingkungan di perkebunan kopi, serta (5) siswa belum mampu menerapkan pelajaran IPA yang sudah diperoleh ke dalam kehidupan sehari-harinya.

Mengacu pada hasil refleksi siklus kedua, maka perlu dilakukan perbaikan pada siklus ketiga yaitu (1) memberikan soal yang non rutin secara kontinyu setiap mempelajari materi IPA dan matematika dengan mengaitkannya pada tema kopi, (2) memberikan petunjuk pemilihan prinsip-prinsip yang digunakan untuk menyelesaikan masalah yang diberikan guru, (3) melampirkan lembar jawaban dari *post test* dengan menuliskan setiap langkah secara rinci dan memberi pengarahan pentingnya menuliskan alasan dari setiap langkah, (4) memutarkan video kerusakan lahan baik perkebunan atau hutan yang diakibatkan oleh ulah manusia dan meminta siswa untuk mengimplementasikan pada lingkungannya, serta 5) siswa diberi permasalahan kehidupan sehari-hari yang berkaitan dengan tema kopi dan materi IPA.

Kegiatan perancangan siklus ketiga dilakukan oleh guru dan peneliti pada tanggal 22 Oktober 2016 di SD Negeri Sidomulyo 03 Jember. Kegiatan ini difokuskan pada hasil refleksi pada siklus kedua yaitu membiasakan siswa untuk mengerjakan soal non-rutin secara kontinyu dan membiasakan siswa untuk melakukan analisis secara rinci dengan memberikan alasan logisnya.

Siklus ketiga ini dilakukan di dua sekolah ujicoba yaitu MI Al-Amin Garahan Jember dan SD Negeri Sidomulyo 03 jember pada tanggal 29 Oktober 2016 dan 5 Nopember 2016. Pada siklus ketiga, siswa diminta untuk menyelesaikan soal non rutin artinya yang belum pernah siswa kerjakan tetapi sering mereka jumpai pada kehidupan sehari-hari mereka. Adapun permasalahan yang disampaikan pada *post test* adalah "Anda adalah seorang pemilik perkebunan kopi yang memiliki luas lahan sebesar 1 hektar. Pada lahan kebun ada harus dibuat suatu irigasi tetapi irigasi tersebut digunakan jika musim kemarau datang, dan pada lahan tersebut anda diwajibkan membuat jalan beraspal sebagai aktivitas berkebun. Bagaimanakah anda mendesain kebun anda untuk memaksimalkan banyaknya pohon kopi yang dapat ditanam? Sketsalah desain anda! Berapakah jarak antar pohon kopi supaya setiap kopi dapat menerima cahaya matahari dan cadangan air yang maksimal? Serta, bagaimana metode anda supaya perkebunan yang anda miliki dapat terjaga kebersihannya.

Untuk menyelesaikan permasalahan tersebutsiswa diberi lembar pengerjaan yang diberi petunjuk secara lengkap yaitu (1) identifikasi informasi apa yang diketahui dan apa yang ditanyakan, (2) tuliskan langkah-langkah pengerjaan untuk menjawab apa yang ditanyakan, (3) lakukan setiap langkah pengerjaan berdasarkan tahap kedua dengan memberikan alasan setiap langkah jawaban, dan (4) lakukan pengecekan kembali dan berikan jawaban lain selain jawaban yang sudah dan berikan sebelumnya. Lembar jawaban tersebut sangat membantu siswa untuk mengembangkan kemampuan pemecahan masalah. Apabila kemampuan pemecahan masalah mampu dikembangkan pada diri siswa, maka siswa memiliki kemampuan berpikir kritis yang maksimal.

Adapun hasil dari siklus ketiga yaitu terdapat 32 siswa dari 80 siswa atau 40% siswa mampu memiliki dan meningkatkan ke-mampuan berpikir kritisnya. Ketiga indikator berpikir kritis mampu dimiliki oleh 32 siswa dalam menyelesaikan soal berbasis proyek. Adapun indikator yang terpenuhi adalah kemampuan pembuktian, kemampuan generalisasi, dan kemampuan pemecahan masalah. Adapun skor dari 32 siswa tersebut lebih atau sama dengan 90 (skor $\geq$ 90). Sedangkan 48 siswa lainnya belum mampu menyelesaikan permasalahan yang diberikan. Hal tersebut disebabkan karena siswa memiliki kemampuan matematika dan IPA yang lemah serta mereka belum berani bertanya kepada temannya jika ada yang belum dipahami. Seorang guru sebagai fasilitator meminta mereka untuk bergabung dengan teman yang bisa agar mereka juga mampu menyelesaikan permasalahan tersebut. Data tersebut diperoleh ketika melakukan wawancara kepada siswa dan guru kelas.

Wawancara dilakukan dari setiap siklus untuk mengecek keabsahan data yang didapat dari hasil pengerjaan *post test* dan pengamatan. Berdasarkan hasil wawancara diperoleh data bahwa (1) guru belum terbiasa menggunakan penilaian berbasis kinerja, (2) guru belum memfokuskan pada kemampuan berpikir kritis siswa, (3) guru belum mengintegrasikan kemampuan matematika dan IPA untuk menyelesaikan permasalahan sehari-hari yang terkait dengan tema kopi, serta (4) guru kesulitan dalam mendesain instrumen penilaian yang berfokus pada kinerja siswa sehingga kemampuan berpikir kritis siswa dapat berkembang. Mengacu pada hasil wawancara itu maka perlu dilakukan suatu kegiatan yang membiasakan siswa untuk berpikir kritis terhadap suatu permasalahan yang dijumpai di kehidupan sehari-hari mereka khususnya terkait dengan kopi dan membiasakan guru untuk mengembangkan instrumen kinerja dengan memperhatikan kemampuan dan lingkungan nyata siswanya.

Hasil yang didapat pada ketiga siklus ini sejalan dengan hasil penelitian terdahulu yang pernah dilakukan peneliti yang menyatakan bahwa kecenderungan siswa dalam menyelesaikan masalah kemampuan generalisasi, kemampuan pemecahan masalah, dan yang terakhir kemampuan verifikasi (Suratno & Kurniati, 2016). Selain itu juga hasil penelitian ini sejalan dengan hasil penelitian sebelumnya yang menyatakan bahwa kemampuan berpikir tingkat tinggi (HOTS) siswa di kabupaten Jember mampu berkembang dengan baik meskipun belum maksimal, yaitu kemampuan logika dan penalaran, analisis, evaluasi, serta kreasi (Kurniati, Harimukti, & Jamil, 2016).

Simpulan

Berdasarkan hasil analisis dan pembahasan berkaitan dengan peningkatan kemampuan berpikir kritis siswa kelas V SD di sekitar perkebunan kopi Garahan Jember, dapat disimpulkan bahwa terdapat peningkatan dari siklus pertama ke siklus kedua, serta dari siklus kedua ke siklus ketiga. Pada siklus pertama dan kedua berturur-turut terdapat 10% dan 27.5% siswa yang mampu memiliki kemampuan berpikir kirtis dengan memenuhi dua indikatornya. Sedangkan pada siklus ketiga, terdapat 40% siswa yang memiliki kemampuan berpikir kritis dengan memenuhi ketiga indikatornya yaitu kemampuan pembuktian, kemampuan generalisasi, dan kemampuan pemecahan masalah.

Peningkatan tersebut juga terjadi pada kemampuan kritis yang dimiliki siswa, yaitu pada siklus pertama, siswa cenderung memiliki kemampuan pembuktian dan kemampuan generalisasi yaitu sub indikator (1) menentukan prinsip dan konsep yang digunakan untuk pemecahan masalah dengan coba-coba, (2) melakukan pembuktian secara benar dengan mengacu pada formula matematika dan IPA yang ada, dan (3) menggeneralisasi bentuk umum dari suatu pola lahan perkebunan kopi.

Pada siklus kedua, semua sub indikator dari indikator kemampuan pembuktian dan generalisasi telah dimiliki siswa. Sedangkan pada siklus ketiga, siswa cenderung memiliki ketiga kemampuan berpikir kritis dengan memenuhi semua sub-indikatornya dari masing-masing kemampuan tersebut.

Adapun saran bagi peneliti lanjut, hasil penelitian ini dapat digunakan sebagai dasar untuk melakukan penelitian terkait dengan proses berpikir kritis siswa khususnya siswa yang memiliki kesamaan dengan siswa di daerah perkebunan kopi Garahan Jember. Selain itu, perlu dilakukan uji coba berkali-kali untuk memperolah data yang akurat tentang perkembangan berpikir kritis siswa dengan mengacu pada instrumen penilaian kinerja.

Ucapan Terimakasih

Ucapan terimakasih disampaikan kepada pemberi anggaran dari kegiatan penelitian ini yaitu Hibah Pengabdian Berbasis Hasil Penelitian Tahun Anggaran 2016 oleh LPM Universitas Jember.

Daftar Pustaka

Bassham, G. (2011). *Critical thinking- a student's instruction*. New York:

McGraw-Hill Companies Inc.

BPS Provinsi Jawa Timur. (2016). Produksi perkebunan kopi. Retrieved October 3, 2016, from http://jatim.bps.go.id/linkTabelStatis/view/id/98.

Glazer, E. (2001). *Using Internet primary sources to teach critical thinking skills in mathematics.* California: Greenwood Press.

Herawaty, S. (2009). *Lesson study berbasis sekolah, guru konservatif menuju guru inovatif.* Malang: Bayumedia Publishing.

Kurniati, D., Harimukti, R., & Jamil, N. A. (2016). Kemampuan berpikir tingkat tinggi siswa SMP di kabupaten jember dalam menyelesaikan soal berstandar PISA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *20*(2), 142. https://doi.org/10.21831/pep.v20i2.8058

Supahar, & Prasetyo, Z. K. (2015). Pengembangan instrumen penilaian kinerja kemampuan inkuiri peserta didik pada mata pelajaran fisika SMA. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *19*(1), 96–108. Retrieved from https://journal.uny.ac.id/index.php/jpep/article/view/4560

Suratno, & Kurniati, D. (2016). Critical thinking of the elementary school students in coffee plantation area based on math science Exemplars task through performance assessment. In *International Conference on Education and Social Science (UK-ICESS)* (pp. 307–312). Malang: Kanjuruhan University.

Suratno, & Kurniati, D. (2017). Performance profile of the coffee plantation area students in solving the math-science problem. *Advanced Science Letters*, *23*(2), 1016–1018. https://doi.org/10.1166/asl.2017.7478

Wisconsin Education Association Council. (1996). Performance assessment. *Education Issues Series.* Retrieved from https://www.learner.org/workshops/socialstudies/pdf/session7/7.PerformanceAssessment.pdf

Jurnal
Penelitian dan Evaluasi Pendidikan

Jurnal Penelitian dan Evaluasi Pendidikan
Volume 21, No 1, June 2017 (11-20)
Online: http://journal.uny.ac.id/index.php/jpep

EVALUASI BUKU TEKS PELAJARAN BAHASA JEPANG TINGKAT DASAR "*MINNA NO NIHONGO*" (Studi Evaluasi di Universitas Darma Persada)

Hani Wahyuningtias
Universitas Darma Persada
Jl. Raden Inten II, Pd. Klp., Duren Sawit, Jakarta Timur, DKI Jakarta 13450, Indonesia
Email: haniwahyu37@gmail.com

Abstrak

Tujuan dari penelitian ini adalah untuk menentukan dan memperoleh pemahaman tentang kualitas buku teks Jepang "*Minna no Nihongo*". Metode yang digunakan dalam penelitian ini adalah metode evaluasi dengan teknik analisis isi. Model evaluasi yang digunakan adalah Evaluasi Berbasis Tujuan (*Goal Based Evaluation*) untuk mengukur dan menilai kualitas buku teks pelajaran bahasa Jepang. Dalam penelitian ini, dasar teoretis dari buku pelajaran yang dijelaskan oleh para ahli dieksplorasi dan dikembangkan oleh peneliti dalam bentuk konstruk instrumen untuk mengevaluasi buku teks bahasa asing. Instrumen ini terdiri dari empat komponen, yaitu: materi/isi, keterampilan berbahasa, penyajian, dan keterbacaan. Instrumen ini telah divalidasi oleh pakar perbukuan dan pakar bahasa Jepang, serta diujikan pada empat buku teks pelajaran seri "*Minna no Nihongo*" yang digunakan di Fakultas Sastra Jurusan Jepang Universitas Darma Persada. Berdasarkan hasil evaluasi, diketahui bahwa kualitas empat buku ditinjau dari empat komponen sekaligus dianggap baik. Namun, buku keterampilan membaca dan menulis ditinjau dari segi penyajian dianggap kurang baik. Instrumen ini diharapkan menjadi pelopor dalam mengevaluasi buku teks pelajaran bahasa asing yang digunakan di Indonesia.

Kata kunci: *evaluasi, buku teks pelajaran bahasa asing, instrumen penilaian buku teks pelajaran bahasa asing*

THE EVALUATION OF JAPANESE TEXTBOOK BASIC LEVEL "*MINNA NO NIHONGO*" (Evaluation Study at Darma Persada University)

Abstract

The purpose of this study is to determine and gain an understanding of the quality of Japanese textbooks "Minna no Nihongo". The method applied in this study is the evaluation method with content analysis techniques. Evaluation model used is Goal Based Evaluation (Objective Oriented Evaluation) to measure and assess the quality of Japanese lesson textbook.In this study, the theoretical basic of the textbooks described by the experts is explored and developed by researcher in the form of instruments construct on evaluating foreign language textbook. This instrument consists of four components, namely: the material/content, language skills, presentation, and readability. This instrument has been validated by book expert and Japanese Language experts, as well as tested in four textbooks series "Minna no Nihongo" used in the Faculty of Literature Japanese Department in University of Darma Persada. Based on the results of the evaluation, it is determined that the quality of four textbooks observed in terms of four components all at once is clarified favorable. But reading and writing skills books observed in terms of presentation only are considered unfavorable. This instrument is expected to be a pioneer in evaluating foreign language text books used in Indonesia.

Keywords: *evaluation, foreign language textbooks, assessment instruments of foreign language textbooks*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v21i1.11812

Pendahuluan

Buku teks merupakan salah satu media belajar yang berperan penting dalam dunia pendidikan. Pemilihan buku teks pelajaran bahasa Jepang di Universitas Darma Persada Fakultas Sastra Jepang tempat peneliti bekerja sampai saat ini ditentukan melalui rapat jurusan dan ditetapkan oleh Ketua Jurusan Jepang tanpa adanya proses evaluasi. Para guru sebaiknya diberikan pengetahuan dan keterampilan yang dibutuhkan untuk mengevaluasi dan mengadaptasi buku teks. Mereka juga harus dipersiapkan untuk menggunakan buku teks sebagai sumber untuk mengajar secara kreatif. Oleh karena itu dalam penetapan buku teks yang akan digunakan di suatu lembaga pendidikan diperlukan suatu pedoman yang dapat membantu para pendidik dalam memilih buku teks pelajaran yang sesuai untuk digunakan dalam proses pembelajaran di tempatnya bekerja.

Menurut Cunningsworth (1995, p. 7) buku teks adalah "*a resource in achieving aims and objectives that have already been set in terms of learner needs*". Kutipan ini menunjukkan bahwa buku teks merupakan sumber (*resource*) dalam mencapai tujuan dan sasaran yang sudah ditentukansebelumnya terkait dengan kebutuhan pelajar. Dengan demikian dapat disimpulkan bahwa buku teks pelajaran yang berkualitas adalah buku yang dapat menunjang kegiatan pembelajaran di kelas.

Padanan kata 'textbook' dalam bahasa Jepang adalah 'kyookasho' yang berarti buku pelajaran; buku teks. Menurut Takamizawa (2004, p. 46) buku teks pelajaran bahasa Jepang secara umum dibagi dua jenis yaitu: *kanji kana majiri bun tekisuto* dan *romaji tekisuto. Kanji kana majiri bun tekisuto* adalah buku teks yang ditulis dengan perpaduan huruf *kana* dan *kanji*, sedangkan *romaji tekisuto* adalah buku teks yang ditulis dengan huruf *romaji* (alphabet) yang digunakan pada masa awal pembelajaran. Buku teks bahasa Jepang yang ditulis dengan huruf alphabet (*romaji*) ditujukan bagi siswa yang berasal dari negara Eropa-Barat. Umumnya buku teks dengan huruf alphabet banyak digunakan pada buku percakapan (*kaiwa*). Pada umumnya, buku teks disertai dengan terjemahan dan penjelasan gramatikal. Namun dalam buku teks dengan huruf alphabet, adakalanya disertai dengan huruf *kana* dan *kanji*sebagai referensi siswa untuk belajar bahasa Jepang.

Di Universitas Darma Persada Fakultas Sastra Jepang, buku teks pelajaran "*Minna no Nihonggo"* telah digunakan mulai tahun 2004 sampai saat ini. Dalam masa yang cukup panjang ini buku digunakan secara berkelanjutan tanpa adanya proses evaluasi buku teks. Mengingat buku teks sebagai sumber pelajaran, diharapkan mengandung materi yang jelas, akurat, dan mutakhir. Oleh karena itu perlu dilakukan evaluasi buku teks untuk mengetahui kesesuaian isi buku teks dengan kurikulum yang berlaku. Dengan memperhatikan fungsi buku teks sebagai media dan sumber pembelajaran, peneliti akan mengevaluasi apakah buku teks pelajaran seri "*Minna no Nihongo*" yang digunakan di Universitas Darma Persada Fakultas Sastra Program Studi Sastra Jepang telah memenuhi syarat sebagai buku pelajaran bahasa asing yang berkualitas sehingga layak digunakan dalam proses kegiatan belajar mengajar di kelas.

Menurut (McGrath, 2002, p. 22) evaluasi buku teks meliputi apakah yang dicari dalam buku teks tersebut ada atau tidak. Ketika yang dicari itu ditemukan maka perlu untuk memberikan nilai pada temuan tersebut. Dengan demikian, evaluasi menyiratkan pengambilan keputusan (*judgment making*) yang cenderung bersifat subjektif. Cunningsworth (1995, p. 14) menjelaskan bahwa evaluasi menekankan pada kelebihan dan kelemahan khusus dalam buku teks yang sudah digunakan sehingga kelebihannya dapat dimanfaatkan secara optimal. Adapun kekurangan dari buku teks dapat diperkuat melalui adaptasi.

Tujuan pengajaran bahasa Jepang di perguruan tinggi yaitu menumbuhkan kemampuan berbahasa yang komunikatif yang meliputi keterampilan menyimak *(kiku),* berbicara *(hanasu),* membaca *(yomu)* dan menulis *(kaku).* Dalam bahasa Jepang empat jenis keterampilan di atas ini disebut dengan *gengo* *ginoo*. Menurut (Taniguchi, 2001, p. 35) *gengo ginoo* ini dikategorikan menjadi dua aspek yaitu: aspek keterampilan berbahasa bersifat produktif atau menghasilkan (*sanshutsuteki*) dan aspek keterampilan bahasa bersifat reseptif atau menerima (*juyooteki*). Kategori pembagian keterampilan berbahasa terangkum pada Tabel 1.

Tabel 1. Empat Keterampilan Berbahasa

	Aktif (*Sanshutsuteki*)	Pasif (*Juyooteki*)
Media Pendengaran	Berbicara (*hanasu*)	Menyimak (*kiku*)
Media Penglihatan	Menulis (*kaku*)	Membaca (*yomu*)

Buku teks bahasa asing terutama diharapkan dapat mengembangkan kompetensi komunikatif supaya siswa berani berkomunikasi sesuai dengan situasi dan kondisi yang dihadapinya. Jika tujuan pembelajaran adalah menjadikan siswa memiliki berbagai kompetensi, siswa perlu untuk menempuh pengalaman dan latihan serta mencari informasi. Alat yang efektif untuk itu adalah buku teks pelajaran sebab pengalaman dan latihan yang perlu ditempuh dan informasi yang perlu dicari, begitu pula cara menempuh dan mencarinya, disajikan dalam buku teks pelajaran secara terprogram. Oleh karena itu, buku teks yang dipakai perlu dievaluasi secara periodik. Tujuan pengevaluasian adalah untuk memutuskan apakah isi buku teks masih sesuai atau tidak dengan kurikulum dan perkembangan ilmu pengetahuan dewasa ini.

Berdasarkan uraian yang telah disampaikan tersebut, maka penelitian ini bertujuan untuk menentukan dan memperoleh pemahaman tentang kualitas buku teks Jepang "*Minna no Nihongo*".

Metode Penelitian

Metode penelitian yang diterapkan dalam penelitian ini adalah metode evaluasi dengan teknik analisis isi (*content analysis*). Menurut Weber (Moleong, 2011, p. 220), analisis isi adalah metodologi penelitian yang memanfaatkan seperangkat prosedur untuk menarik kesimpulan yang sahih dari sebuah buku atau dokumen. Adapun model evaluasi yang digunakan adalah evaluasi berbasis tujuan atau (*Objective Oriented Evaluation*) untuk mengukur dan menilai kualitas buku teks pelajaran bahasa Jepang. Menurut Wirawan (2012, p. 81), model evaluasi berbasis tujuan memokuskan pada pengumpulan informasi yang bertujuan mengukur pencapaian tujuan kebijakan, program, dan proyek untuk pertanggungjawaban dan pengambilan keputusan.

Dengan demikian, penilaian buku teks bahasa Jepang dengan menggunakan teknik analisis isi dianggap sesuai dalam rangka menarik kesimpulan melalui usaha menemukan karakteristik pesan buku teks yang dilakukan secara objektif dan sistematis. Berdasarkan hasil penilaian responden dalam wujud skor dan penilaian kualitatif terhadap buku teks akan diketahui kesesuaian buku teks dengan kriteria evaluasi buku teks. Dengan menggunakan model evaluasi berbasis tujuan akan dinilai dan dianalisis keadaan buku teks pelajaran seri "*Minna no Nihongo"* berdasarkan kriteria evaluasi yang ditetapkan dalam penelitian ini.

Evaluasi Buku Teks

Evaluasi merupakan kegiatan sistematis yang dilaksanakan untuk membantu audiensi agar dapat mempertimbangkan dan meningkatkan nilai suatu program atau kegiatan. McGrath (2002, p. 22) membedakan antara analisis dan evaluasi buku. Analisis adalah suatu proses menuju pendeskripsian yang bersifat obyektif dan dapat dipercaya, sedangkan evaluasi meliputi makna pengambilan keputusan *(judgement making)*. Analisis dilakukan untuk menyelidiki apa yang ada di buku itu sedangkan evaluasi lebih ditujukan pada apakah yang dicarinya ada atau tidak di buku itu. Jika ditemukan apa yang dicari, diberikanlah nilai atas temuan tersebut.

Langkah dalam penelitian ini adalah: (1) pengumpulan konsep/teori tentang buku teks; (2) penyusunan Instrumen penilaian buku teks; (3) validasi pakar (*Expert Judgement*); (4) pengumpulan data angket/kuesioner dan wawancara; (5) pengolahan data: (a) mencatat data, (b) memilah dan memeriksa data, (c) menganalisis dan menginterpretasi data, (d) mengadakan diskusi kelompok terarah/*Focus Group Discussion* (FGD); (6) penulisan laporan.

Kriteria Evaluasi Buku Teks

Menyusun instrumen pada dasarnya adalah menyusun alat evaluasi, karena mengevaluasi adalah memperoleh data tentang sesuatu yang diteliti, dan hasil yang diperoleh dapat diukur dengan menggunakan standar yang telah ditentukan. Instrumen berfungsi sebagai alat dalam mengumpulkan data yang diperlukan. Peneliti berdasarkan landasan teoretis menyusun instrumen penilaian buku teks bahasa Jepang. Instrumen dalam penelitian ini berupa daftar pertanyaan atau pernyataan secara tertulis yang harus dijawab atau diisi oleh responden sesuai dengan petunjuk pengisiannya.

Pada Tabel 2 diuraikan kisi-kisi instrumen sesuai dengan komponen dan aspek yang akan dievaluasi. Kisi-kisi instrumen berisi komponen, indikator, danbobot untuk setiap komponen yang dievaluasi.

Tabel 2. Kisi-kisi Instrumen Penilaian Buku Teks Bahasa Jepang

No	Komponen	Indikator	Bobot
1	Materi/Isi	1. Mendukung kurikulum dan SAP 2. Orisinal, tidak mengandung diskriminasi gender dan tidak menimbulkan masalah SARAP 3. Memiliki kebenaran keilmuan, mutakhir, sahih, akurat 4. Menampilkan kondisi Jepang dewasa ini dan erat dengan konteks ke-jepang-an 5. Memaksimalkan kondisi yang sesuai dengan kondisi di luar Jepang	10
2	Keterampilan berbahasa	1. Menyimak 2. Berbicara Materi 3. Membaca Latihan 4. Menulis	10
3	Penyajian	1. Runtut, bergradasi, bersistem, lugas dan didukung ilustrasi 2. Seimbang dan berkesinambungan 3. Mengembangkan sikap spiritual dan sosial 4. Mengembangkan keterampilan berbahasa dan menumbuhkan motivasi untuk berkreasi dan berinovasi 5. Dilengkapi dengan buku pegangan guru 6. Dilengkapi dengan tes dan lembar kerja siswa 7. Dilengkapi dengan media belajar 8. Dilengkapi dengan bahan bergambar	10
4	Keterbacaan	1. Kemudahan membaca instrumen dan materi 2. Bahasa yang digunakan etis, komunikatif, fungsional, kontekstual, efektif, dan efisien. 3. Memiliki kandungan gramatikal dan kosakata yang sesuai dengan level, minat, dan kognisi siswa	10

Dalam pengevaluasian buku teks pelajaran seri "*Minna no Nihongo*" ini, responden diminta menuliskan skor. Skor didasarkan atas skala pengukuran *rating-scale*. Dengan *rating-scale* data mentah yang diperoleh berupa angka kemudian ditafsirkan dalam pengertian kualitatif. Hal yang penting bagi penyusun instrumen dengan *rating-scale* adalah dapat mengartikan setiap angka yang diberikan di setiap butir instrumen.Validasi instrumen dilakukan dengan meminta beberapa orang pakar dalam bidangnya untuk menilai instrumen yang disusun oleh peneliti. Pakar yang terlibat dalam penelitian ini adalah pakar perbukuan (satu orang) dan pakar bahasa Jepang (dua orang). Dalam hal ini, pakar diminta pendapatnya sehubungan dengan aspek-aspek yang akan diukur berlandaskan teori tertentu. Hal ini disebut dengan pendapat dari ahli (*Expert Judgment*).

Dalam penelitian ini skor didasarkan atas skala 1-2 (tidak terdapat kesesuaian); 3-5 (kurang lebih di bawah 50% sesuai); 6-8 (di atas 50% sesuai); 9-10 (seluruhnya sesuai); dan untuk pertanyaan yang dapat langsung dijawab dengan 'ya' atau 'tidak' diberikan dua pilihan jawaban yaitu: jika tidak terdapat kesesuaian diberikan skor 1 dan jika terdapat kesesuaian diberikan skor 10. Untuk bobot, umumnya berlaku seperti yang tertulis dalam instrumen penilaian buku teks. Namun, untuk komponen keterampilan berbahasa jumlah bobot sangat bergantung pada jumlah keterampilan yang terdapat dalam buku teks. Jika buku teks memiliki empat keterampilan, di bagian 'materi' setiap butir komponen keterampilan diberikan bobot 1, dan bagian 'latihan' diberikan bobot 1,5. Jika buku hanya mengandung satu buah keterampilan, di bagian 'materi' diberikan bobot 4,5 dan 'latihan' diberikan bobot 5,5 (lihat appendiks).

Persentase setiap komponen ditetapkan secara berbeda. Komponen materi diberikan persentase yang paling besar karena materi/isi merupakan skenario pembelajaran yang menjadi panduan bagi guru dalam menjalankan pembelajaran di kelas. Hal ini sesuai dengan pendapat Cunningsworth, (1995, p. 7) yang menyebutkan bahwa materi merupakan sumber aktivitas bagi pemelajar dalam praktik dan komunikasi interaktif. Keterampilan berbahasa menempati urutan kedua karena dalam pengajaran bahasa asing keterampilan ditujukan untuk memenuhi kebutuhan kemahiran berbahasa yang disajikan dalam satu kesatuan yang terpadu yaitu: menyimak, berbicara, membaca, dan menulis sesuai dengan situasi dan kondisi yang dihadapinya. Penyajian menempati urutan ketiga karena penyajian yang baik dapat melengkapi kesempurnaan sebuah buku. Adapun keterbacaan menempati urutan keempat karena keterbacaan berkaitan dengan kemudahan bahasa bagi siswa level yang dituju. Kesesuaian tingkat keterbacaan buku teks berpengaruh terhadap motivasi dan minat siswa untuk membacanya. Berdasarkan pertimbangan tersebut di atas, komponen materi diberikan persentase 40%, komponen keterampilan berbahasa diberikan persentase 30%, komponen penyajian diberikan persentase 20%, dan komponen keterbacaan diberikan persentase 10%. Kualitas buku teks bergantung pada persentase penilaian yang dicapai dari kriteria yang ditetapkan yaitu: 76-100 % (Baik); 51-75 % (Kurang baik); 26-50 % (Tidak baik); 0- 25 % (Sangat tidak baik). Data mentah yang diperoleh berupa angka ditafsirkan dalam pengertian kualitatif, ini disebut dengan *rating-scale* (Sugiyono, 2010, p. 141). Dalam penelitian ini penentuan kualitas buku teks didasarkan atas *rating-scale* untuk mengukur persepsi responden terhadap buku teks.

Hasil Penelitian dan Pembahasan

Berdasarkan hasil angket jawaban delapan orang responden diketahui kualitas keempat buku teks pelajaran seri "*Minna no Nihongo*" berdasarkan setiap komponen dan empat komponen sekaligus. Hasil penilaian buku teks terangkum pada Tabel 3 dan 4.

Tabel 3. Kualitas Buku Teks Pelajaran Seri "*Minna no Nihongo*" Berdasarkan setiap Komponen

Judul Buku Teks	Kualitas			
	Materi/Isi	Keterampilan Berbahasa	Penyajian	Keterbacaan
Minna no Nihongo Shokyu I Chookai Tasuku 25	80,8% (Baik)	86,9% (Baik)	77,8% (Baik)	80,8% (Baik)
Minna no Nihongo Shokyu I Dai 2 Ban	82,8% (Baik)	75,6% (Baik)	82,3% (Baik)	87% (Baik)
Minna no Nihongo Shokyu I Shokyu *de Yomeru Topikku 25*	81% (Baik)	76,8% (Baik)	65,6% (Kurang baik)	78.8% (Baik)
Minna no Nihongo Shokyu I Kanji Eigoban	81,3% (Baik)	81,9% (Baik)	66,5% (Kurang baik)	76,8% (Baik)

Tabel 4. Kualitas Buku Teks Pelajaran Seri "*Minna no Nihongo*" Berdasarkan Empat Komponen

Keterampilan Berbahasa	Judul Buku Teks	Kualitas Buku Teks
Menyimak	*Minna no Nihongo Shokyu I Chookai Tasuku 25*	82% (Baik)
Buku Inti (Empat keterampilan Berbahasa)	*Minna no Nihongo Shokyu I Dai 2 Ban*	81% (Baik)
Membaca	*Minna no Nihongo Shokyu I* Shokyu *de Yomeru Topikku 25*	76,5% (Baik)
Menulis	*Minna no Nihongo Shokyu I Kanji Eigoban*	78,1% (Baik)

Keterangan:

76-100%: Baik	51-75 %: Kurang baik
26-50 %: Tidak baik	0-25 %: Sangat tidak baik

Dengan demikian dapat disimpulkan bahwa empat buku teks pelajaran seri "*Minna no Nihongo*" ditinjau dari empat komponen sekaligus dianggap memiliki kualitas baik. Namun, buku keterampilan membaca dan menulis jika ditinjau hanya dari segi penyajian saja, dianggap kurang baik. Hal ini disebabkan buku teks tersebut dianggap kurang dapat mengembangkan sikap spiritual dan sosial, tidak dilengkapi dengan buku khusus pegangan guru, dan tidak disertai dengan media belajar seperti: CD dan DVD.

Berdasarkan empat buku teks pelajaran bahasa Jepang tingkat dasar "*Minna no Nihongo*" tersebut diketahui bahwa tiga di antaranya yaitu buku: keterampilan menyimak "*Minna no NihongoShokyu I Chookai Tasuku 25* (2003)", keterampilan membaca "*Minna no Nihongo Shokyu I* Shokyu *de Yomeru Topikku 25* (2000)", keterampilan menulis "*Minna no Nihongo Shokyu I Kanji Eigoban* (2000)*"* ditinjau dari segi materi khususnya butir keempat dianggap beberapa bagian yang terdapat di dalamnya kurang menampilkan kondisi sosiokultural Jepang dewasa ini. Kondisi sosiokultural dalam hal ini menyangkut keadaan riil masyarakat Jepang. Adapun buku inti yang berjudul "*Minna no Nihongo Shokyu I Dai 2 Ban*" karena sudah memasuki edisi kedua yaitu tahun 2012, isinya dianggap sudah sesuai dengan kondisi Jepang dewasa ini. Bagian yang 'kurang menampilkan kondisi sosiokultural bahasa asing yang sedang dipelajari' tersebut dapat dikembangkan dengan referensi lain yang mendukung sesuai kebutuhan di kelas dan target yang hendak dicapai dalam pembelajaran.

Berdasarkan empat buku teks pelajaran bahasa Jepang tingkat dasar "*Minna no Nihongo*" tersebut diketahui bahwa penggunaan sumber-sumber yang sesuai dengan kondisi di luar bahasa yang dipelajari dianggap kurang maksimal. Hal ini karena sumber-sumber di luar bahasa yang dipelajari sifatnya hanya sebagai pelengkap sehingga bagian tersebut ditampilkan berdasarkan kebutuhan. Adapun jika pengajar ingin mengadakan pengayaan materi tentang hal-hal di luar kondisi bahasa yang dipelajari, dapat memaksimalkan materi yang sedikit tersebut dengan mengompilasi dari rererensi pendukung lainnya seperti: informasi dari internet, program televisi Jepang maupun surat kabar Jepang yang dianggap dapat menunjang pembelajaran. Hal di atas sesuai dengan pendapat Garinger (2002, p. 2) yaitu: latihan dan kegiatan sebaiknya disajikan dalam format yang bervariasi sehingga secara terus menerus dapat memotivasi siswa.

Buku teks perlu mempertimbangkan aspek budaya pembelajaran. Buku teks bahasa Jepang yang digunakan oleh pemelajar di Indonesia sebaiknya bukan hanya berisi budaya Jepang semata, tetapi sebaliknya juga berisi muatan budaya lokal maupun budaya Barat. Bahan yang berisi budaya dari negeri pengguna bahasa Jepang akan ber-manfaat untuk memperkaya wawasan siswa dan memberikan gambaran sesungguhnya tentang penggunaan bahasa Jepang dalam kehidupan nyata sehari-hari. Bahan ajar yang mengandung budaya lokal seperti Indonesia akan membuat siswa merasa akrab dan tidak asing dengan hal-hal yang dijadikan topik dalam pembelajaran. Bahan ajar yang mengandung budaya Barat ditujukan untuk memperkaya wawasan siswa. Hal ini sesuai dengan pendapat McGrath (2002, p. 156) yaitu bahan ajar sebaiknya merefleksikan dunia luar dalam hal keaslian teks dan tugas. Berdasarkan paparan dan kutipan di atas, diharapkan siswa dengan pengetahuan yang diperoleh dari buku, mereka dapat menyerap, membandingkan, dan mengkaji nilai-nilai positif yang terkandung di dalamnya.

Bagi masyarakat Jepang, saat pergi makan bersama dengan teman adalah hal yang umum untuk membayar menu pesanan masing-masing. Hal ini tercermin dalam cuplikan percakapan (kaiwa) yang dikutip dari buku inti "*Minna no Nihongo Shokyu I Dai 2 Ban*" pelajaran 13.

すみません。<u>別々にお願いします。</u>
Sumimasen. <u>Betsu-betsu ni onegaishimasu</u>.
(3A Network, 2012)

Terjemahan:
Maaf. <u>Tolong (bayarnya) dihitung masing-masing.</u>

Namun, bagi orang Timur, seperti orang Indonesia mentraktir makan teman dianggap sebagai suatu keramahan untuk mengakrabkan diri dengan lawan bicara. Adapun jika ingin membayar masing-masing, umumnya hal tersebut tidak terucapkan secara langsung kepada lawan bicara. Sebenarnya di masyarakat Jepang ada juga konsep mentraktir orang lain di kesempatan tertentu asalkan sudah disepakati oleh si pembicara sebelumnya, misalnya dinyatakan dengan ungkapan '*konkaiwa boku ga <u>ogoru</u> yo*' (*This round's on me*).Sehubungan dengan pentingnya nilai sosiokultural, maka dalam pengevaluasian buku teks bahasa asing, unsur ini dimasukkan pada bagian komponen materi butir 4 dan 5. Dalam pembelajaran bahasa asing seorang anggota masyarakat perlahan-lahan belajar mengenal adat istiadat, tingkah laku, dan tata krama masyarakatnya. Oleh karena itu diperlukan sikap harmoni terhadap budaya bahasa sasaran untuk menghindari adanya benturan budaya. Dalam hal ini, guru diharapkan berperan untuk menanamkan dan mengembangkan sikap siswa dalam prinsip menghargai dan memberi ruang untuk menerima adanya perbedaan.

Berdasarkan empat buku teks pelajaran bahasa Jepang tingkat dasar "*Minna no Nihongo*" diketahui bahwa jika ditinjau dari segi penyajian butir ketiga yaitu pengembangkan nilai spiritual dan sosial, dianggap masih kurang karena pada hakikatnya empat buku tersebut merupakan buku keterampilan berbahasa asing yang lebih menekankan pada kemampuan menggunakan bahasa Jepang. Namun, bagian yang dianggap kurang dapat mengembangkan nilai spiritual dan sosial tersebut dapat dioptimalkan dengan cara mengkaji secara mendalam isi yang terkandung di dalamnya sambil dikompilasi dengan referensi lain yang menunjang materi pembelajaran. Nilai spiritual dan sosial sangat penting dihadirkan dalam buku teks pelajaran bahasa asing seperti: menolong orang tidak mampu dan orang sakit dengan bantuan keuangan, pendidikan, maupun kasih sayang terhadap sesama. Hal ini perlu diteladani sebagai sikap yang mulia. Tujuan pendidikan bukan hanya membentuk individu yang cerdas tetapi juga membentuk pribadi yang berkepribadian atau berkarakter. Individu yang cerdas tersebut diharapkan mampu mengembangkan potensinya yaitu kekuatan spiritual keagamaan, pengendalian diri yang baik, dan memiliki akhlak yang mulia. Oleh karena itu, walaupun sedikit dalam buku teks pelajaran bahasa asing perlu dimasukkan nilai-nilai luhur yang dapat diteladani oleh para siswa.

Berdasarkan empat buku teks pelajaran bahasa Jepang tingkat dasar "*Minna no Nihongo*" diketahui bahwa dua buku di antaranya yaitu buku keterampilan membaca "*Minna no NihongoShokyu I* Shokyu *de Yomeru Topikku 25"* dan keterampilan menulis "*Minna no Nihongo Shokyu I Kanji Eigoban"* tidak disertai dengan media belajar seperti: kaset, CD, dan DVD. Buku inti yang mengandung empat keterampilan berbahasa yaitu: "*Minna no Nihongo Shokyu I Dai 2 Ban"* dan buku menyimak "*Minna no Nihongo Shokyu I Chookai Tasuku 25"* dilengkapi dengan media belajar berupa kaset, CD, dan DVD sehingga sangat menunjang pembelajaran khususnya dalam rangka mengasah kemampuan berkomunikasi secara lisan.

Dari empat buku teks pelajaran bahasa Jepang tingkat dasar "*Minna no Nihongo*" tersebut diketahui bahwa hanya buku inti yaitu: "*Minna no NihongoShokyu I Dai 2 Ban"* yang dilengkapi dengan buku khusus pegangan guru yang berjudul "*Minna no Nihongo Shokyu I Oshiekata no Tebiki*". Buku keterampilan menyimak "*Minna no Nihongo Shokyu I Chookai Tasuku 25"* tidak disertai dengan buku khusus pegangan untuk guru. Namun, di bagian awal buku yaitu setelah kata pengantar dilampirkan 'bagi para guru yang menggunakan materi bahan ajar ini (*kono kyoozai o otsukai ni naru senseigata e*)' yang berisi: keistimewaan materi bahan ajar, cara penggunaan pada umumnya, dan langkah penggunaan sebanyak dua halaman. Buku keterampilan membaca "*Minna no Nihongo Shokyu I* Shokyu *de Yomeru Topikku 25*" tidak disertai dengan buku khusus pegangan untuk guru. Namun, melalui lembar 'cara penggunaan buku ini (*kono hon no tsukaikata*)' yang terdapat di bagian halaman depan buku dan delapan lembar halaman terpisah 'petunjuk bagi guru (*kyooshiyoo gaido*)' dapat dipelajari cara menggunakan buku keterampilan membaca tersebut. Buku keterampilan menulis "*Minna no Nihongo Shokyu I Kanji Eigoban"* tidak dilengkapi dengan buku khusus pegangan guru. Namun, buku ini dilengkapi dengan *booklet* referensi terpisah yang berisi target *kanji*, kosakata *kanji*, dan indeks.

Berdasarkan empat buku teks pelajaran bahasa Jepang tingkat dasar "*Minna no Nihongo*" tersebut *diketahui* bahwa empat buku tersebut dilengkapi dengan bahan bergambar. Adapun buku yang dilengkapi dengan kartu baca (*flashcards)* adalah buku "*Minna no NihongoShokyu I Dai 2 Ban"* yang merupakan buku inti dari buku teks pelajaran seri "*Minna no Nihongo*".

Simpulan

Berdasarkan hasil evaluasi dan pembahasan, peneliti menyimpulkan empat poin berikut ini. Pertama, buku teks perlu dievaluasi secara periodik oleh pihak penyelenggara pendidikan, pengajar, dan semua pihak yang terkait mengingat esensi dari buku teks itu sendiri yaitu sebagai sumber pembelajaran.

Kedua, materi yang sangat padat dalam empat buku tersebut dianggap dapat menimbulkan kesulitan pada guru maupun siswa jika diterapkan pada kelas non intensif. Bagi siswa SMA yang jam pelajaran bahasa Jepang terbatas seminggu sekali atau seminggu dua kali, sebaiknya tidak menggunakan buku ini sebagai buku utama, tetapi dapat menggunakannya sebagai materi tambahan atau pelengkap saja. Buku teks pelajaran bahasa Jepang untuk siswa SMA diharapkan buku yang lebih ringan kandungan materinya dalam rangka pengenalan bahasa Jepang terhadap siswa pemula dan penumbuhan minat pada siswa untuk belajar bahasa Jepang.

Ketiga, untuk siswa yang akan belajar secara intensif, baik mahasiswa Jurusan Sastra Jepang (S1), bahasa Jepang (D3), maupun mereka yang akan belajar di Jepang untuk melanjutkan pendidikan atau melakukan pelatihan (*training*) ke Jepang, empat buku tersebut dapat digunakan secara terpadu dalam rangka melatih dan meningkatkan kemampuan siswa berbahasa Jepang. Dengan waktu belajar yang intensif, buku ini akan lebih mudah untuk dipelajari dan dikuasai keterampilan bahasa yang terkandung di dalamnya.

Keempat, *Japan Foundation* (*Kokusai Kooryuu Kikin*) Jakarta sebagai lembaga yang memiliki kegiatan utama untuk mengembangkan bahasa Jepang di Indonesia diharapkan untuk terus mendorong dan memajukan pendidikan bahasa Jepang di Indonesia dengan mengenalkan buku teks karya penutur asli Jepang yang berkualitas dan membantu mensosialisasikan isi buku teks tersebut kepada pengajar bahasa Jepang orang Indonesia melalui kegiatan belajar bersama (*benkyookai*), analisis materi ajar (*kyoozai bunseki*), maupun metode pengajaran (*kyoojuhoo*).

Ketebatasan Penelitian

Penelitian ini telah diusahakan dan dilaksanan sesuai prosedur ilmiah, namun demikian masih memiliki ketebatasan yaitu sebagai berikut.

Pertama, dari empat buku teks yang mengandung empat keterampilan bahasa itu, buku edisi terbaru hanya pada buku inti yaitu buku yang berjudul "*Minna no Nihongo Shokyu I Dai 2 Ban*" (2012), sedangkan tiga buku teks lainnya masih edisi yang lama.

Kedua, adanya keterbatasan penelitian dengan menggunakan kuesioner yaitu terkadang jawaban yang dierikan responden ada bagian yang kurang mendalam.

Daftar Pustaka

3A Network. (2012). *Minna no Nihongo Shokyu I Dai 2-Han Honsatsu Kanji-Kana* (2nd ed.). Tokyo: 3A Network.

Cunningsworth, A. (1995). *Choosing your coursebook*. Oxford: Macmillan Heinemann English language teaching.

Garinger, D. (2002). Textbook Selection for the ESL Classroom Steps in the Selection Process. *Center for Applied Lingustics*. Retrieved from http://www.mcael.org/uploads/File/provider_library/Textbook_Eval_CAL.pdf

McGrath, I. (2002). *Materials evaluation and design for language teaching (Edinburgh textbooks in applied linguistics).* Edinburgh: Edinburgh University Press.

Moleong, L. J. (2011). *Metodologi penelitian kualitatif*. Bandung: PT Remaja Rosdakarya.

Sugiyono. (2010). *Metode penelitian pendidikan, Pendekatan kuantitatif, kualitatif dan R&D*. Bandung: Alfabeta.

Takamizawa, H. (2004). *Shin hajimete no nihongo kyouiku kihon yougo jiten*. Tokyo: Asuku.

Taniguchi, S. (2001). Nihongo nouryoku to wa nani ka. In *Nihongo Kyouikugaku wo Manabu Hito no Tame ni*. Kyoto: Sekaishisosha.

Wirawan. (2012). *Evaluasi: teori, model, standar, aplikasi, dan profesi*. Jakarta: Rajawali Press.

Jurnal
Penelitian dan Evaluasi Pendidikan

Jurnal Penelitian dan Evaluasi Pendidikan
Volume 21, No 1, June 2017 (21-33)
Online: http://journal.uny.ac.id/index.php/jpep

AN EVALUATION MODEL OF ISLAMIC LEARNING EDUCATION PROGRAM IN MADRASAH ALIYAH

Anidi
STKIP Pelita Nusantara Buton
Jl. Pahlawan Km. 4 Baubau. Kota. Kota Baubau - Prop. Sulawesi Tenggara, 93716 Indonesia
Email: said_anidi@yahoo.com

Abstract

The purpose of this study is: (1) to create the products as the evaluation model learning programs of Islamic education in madrasah aliyah, suitable, precise, and accurate, (2) to know how to test the suitability of the model to obtain the evaluation model of Islamic Education learning programs in madrasah aliyah which is valid, and reliable, and (3) determine the effectiveness of the application of evaluation model of Islamic learning programs that was developed. This was a research and development method study (R & D), referring to the model of Borg & Gall, the development process was simplified into three steps, namely: (1) the stage predevelopment of models, (2) the stage of development of the model, and (3) the operational phase of the field (test model). The data were collected using questionnaires, interview, observation, documentation, and test. Data were analyzed using quantitative and qualitative. The conclusion of this research and development is: (1) evaluation model of Islamic Religious Education Learning program developed comprising: (a) the evaluation procedures, (b) the evaluation instrument, and (c) the evaluation guide. (2) according to the experts, PAI teachers, and principals, procedures, instruments, and evaluation guidelines developed are already good, and can be used, d) the instrument developed is entirely valid (load factor > 0.3), reliable (CR > 0.7), and qualified as a fit model (RMSEA ≤ 0.08 and NFI, CFI and GFI > .90). second, the expert's opinion, PAI teachers and principals, the developed evaluation model (EPH model) is declared effective, practical, and easy to use, and supported by a valid and reliable instrument.

Keywords: *model of evaluation, evaluation process, evaluation outcome, Islamic learning education*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v21i1.12003

Introduction

People's lives are faced with the issue of decline in manners and ethics in social life, both in the family environment, school (madrasah), as well as in the surrounding environment. The phenomenon of public life today show the stronger need to optimize the implementation of Islamic Education in Madrasah, so able to become the basis of thinking and behave in life.

Religious Education today have many kinds of sharp criticism, because inability to cope the important issues in the life of societies (Sutrisno, 2006, p.5). The facts showed that the influence of non-educative foreign cultures global as well, such as the culture of materialism, consumerism and hedonism which is getting stronger (Muhaimin, 2009, p.51). Religion Education should get a common concern in building the character and moral of the nation.

Ideally, Islamic Religious Education (PAI) able to be the basis of other education, as well as being one of the measures in building national character (Nation character building) and the personality of the learner is balanced both in intelligence quotient (IQ), emotional quotient (EQ) and spiritual quotient (SQ) (Agustian, 2003, p. 175). More pronounced in Law Number 20 Year 2003 on National Education System, that the implementation of PAI in madrasas includes planning, implementation and assessment systems should be well planned, integrated, and sustainable with regard to all aspects of the student whether cognitive, affective and psychomotor.

Various efforts have been made by the government to improve the quality of National Education, namely through the development and improvement of curriculum, improvement of evaluation systems, development of learning materials, procurement of books and tools of learning, improvement of education infrastructures, increased the teacher kompentence, as well as improving the quality of school leadership. As a formal educational institutions, implementation of education in Madrasah should refer to the National Education Standards.

Islamic Religious Education Learning at the school is dependen on factors educators or teachers. Teachee is one component that has an important role Because in addition to as exemplary figure and teacher also as a facilitator, administrator, motivator, counselor, organizer, and Evaluators. The role of the teacher as a manager of learning (learning manager) requires teachers to manage and create a climate conducive learning and fun in order to form the religious competence of students as a whole. Additionally, the learder and Teachers PAI demanded to be resilient and creative in order to create a conducive learning environment Islamic Education.

The creation of a religious atmosphere in madrasah namely by utilizing a mosque or prayer room to pray and practice activities of worship, selebration Islamic great days, the religious broadcasting, convey the values of religious and moral messages to all subjects, utilizing the religion figures in the communities as a learning resource, and the availability of religious laboratory. All of them would help to achieve the learning objectives of the maxsimal Islamic education.

Improving the quality of education be conducted continuously by improving the quality of learning in all fields of study, so that the education goals can be achieved effectively and more efficient. Focus or direction in improving the quality of education is achieved the educational goals as the ability of an intact self-learners, which include the ability of academic or intellectual capital, and moral capacity or moral capital (Zamroni 2005, p. 1). Both the authorized capital is the capital required to improve the quality of education.

Duties and responsibilities of Islamic Education in Madrasah Aliyah not only on religious teacher, but also the responsibility of the madrassa as a whole. Madrasah environment should support and be a laboratory for the study of Islamic education. Improving the quality of learning requires efforts enhancement the whole evaluation learning program because the nature of the learning quality is the quality of the learning implementation program that has been designed.

Improving the quality ofthe program implementation requires enhancement in aspects of other programs, such as program designing and program components. Improving the quality of the learning program components includes: learning objectives, teachers, students, materials, sources of learning materials, as well as the learning facilities.

Based on the results of preliminary research conducted at several research sites which concluded that: (1) input as input in learning, namely learning completeness inadequate facilities, especially the subjects of jurisprudence requiring laboratory practice; (2) the learning process associated with the performance of teachers in the classroom, the teacher is still dominated learning; and (3) assessment of learning outcomes they prefer the cognitive, while affective and psychomotor were the less noticed.

Research conducted Schneider (Morrison, Mokasi, & Cotter, 2006, p. 5) concluded that the physical environment of the classroom or learning facilities had a significant impact on student learning and teacher performance. Classrooms were uncomfortable, hot, cold and many people were passing so as an obstacle to achieve the better learning. the next Results of research conducted by Sudjana (2002, p. 42) showed that 76.6% of student learning outcomes were influenced by the performance of teachers, it can be seen the ability of teachers to teach contributed 32.43%, mastery of the subject matter contributed 32, 38 % and the attitude of teachers to the subjects contributed 8.60%. While research conducted Wahyudi (2003, p. 1) proveded that there was strong correlation between student achievement in a class with moods or social environment created in the class.

Based on preliminary studies and empirical data mentioned above, to more optimize the results of the evaluation of the Islamic learning Education in Madrasah Aliyah, an evaluation of the learning program needs to be made more comprehensive, with scope of program not only the aspect of learning outcomes, but also included the aspect of the learning process.

Ideally Madrasah Aliyah according to Saleh, (2004, p.47) produced a profile that describe graduations as follows: (1) had the faith and piety in accordance with the teachings of his religion, (2) had a base value humanioran to implement a unity in life, (3) master academic ability and skills as well as the ethos of learning to continue education, (4) Converting academic ability and life skills in local and global communities, (5) the ability of expression, appreciated art and beauty, (6) has the ability to exercise, maintaining healthy, build endurance and physical fitness, (7) participated and nationality insight into the life of society, nation and state as a democracy.

Islamic Education in Madrasah Aliyah needs colaborate between knowing (knowing), doing (practice), and being (live), so that the necessary adaptation learning program evaluation that is implemented. Mardapi (2000, p. 2) explained that the evaluation in the field of education in review of the target can be divided in two parts, namely the evaluation of macro and micro evaluation. Evaluation of macro targets are education programs in general, the program planned to improve education. Evaluation of micro frequently used at the classroom level. So the micro evaluation targets are learning in the classroom program.

Evaluation of the learning program is more important is how the learning of Islamic Education could be evaluated in a professional manner, so provided the accurate and comprehensive information. It required a program evaluation model specifically developed to evaluate the components of the learning program that is implemented.

In order to achieve successful learning programs PAI in madrasa, requred the suitable evaluation model so could provide the accurate information to stakeholders and more important to the parties concerned, especially the leadership of madrasas, both in terms of content, scope and format of evaluation to improve the learning program.

This study developed an evaluation model of Islamic Religious Education Learning program (PAI) at Madrasah Aliyah. Evaluation model developed was limited to components: (1) the evaluation procedure (2) the evaluation instrument, (3) the evaluation guide. Therefore, the problem would be examined in this study, as follows: (1) how the evaluation model of learning programs PAI to provide the comprehensive information, axactly and accurate, and could contribute the benefit to the leadership and the teacher of Madrasah?, (2) how the suitability evaluation model of learning program PAI at Madrasah Aliyah valid and reliable? and (3) how the effectiveness of the learning program evaluation model developed?

Based on the formulation of the problem mentioned above, the purpose of research and development are as follows: (1) develop a model evaluation of learning programs that provide comprehensive information, exactly and accurate, (2) find out the suitability evaluation model of learning programs are valid and reliable, and (3) find out the effectiveness evaluation model of learning programs developed.

Research Methods

This research is development research (research and development), which aims to produce a product in the form of a model of evaluation, namely the learning evaluation program of Islamic education in Madrasah Aliyah. The research model used based the models of the research and development by Borg & Gall (1993, pp. 275-276), which consists of ten steps: (1) research and information colecting, (2) planning, (3) the first product development , (4) the primary trials, (5) the revision of major products, (6) the trial primary, (7) the revision of operational product, (8) trial field operations, (9) the revision of the product, and (10) the dissemination and implementation.

Ten steps which refer to the development of research models Borg & Gall to adjust into three steps, namely: (1) predevelopment stage models, (2) the stage of the model development, and (3) the stage field operational (test model). Pre-development stage models conducted with observations, interviews, and documentation by the teachers and the headmaster as well as reviewing relevant literature and research.

The development phase was done by determining the model of program evaluation with validation instrument colleagues (first draft), validation expert (expert judgment), and the drafting II. The field operational phase (test model) was done with Phase I trial (legibility), Phase II trial (feasibility), and Phase III trial (field operations).

The Subjects in this study consists of the students, teachers and headmaster. the sampling technique was done by purposive random sampling, namely the classified of the good quality, medium, and low. Madrasah was used as test subjects were 15 madrassas, which was divided into five good quality madrasas, 5 medium quality madrasa, and 5 Madrasa are low quality. Madrasah is spread several Ministry of Religious Regency/City region of Southeast Sulawesi province.

Data in this research are quantitative and qualitative data. Data collection techniques by used interviews, observations, questionnaires, and documentation. The research instruments was the assessment sheet, interview guidelines, observation guidelines, questionnaires, documentation guidelines, and sheets of observations, and tests.

Validity of the instrument in this study included content validity (content validity), and construct validity. The content validity obtained from the judging of the Experts and teacher of PAI towards the research instruments that had been prepared by using the formula Aiken's V. Testing of validity and reliability construct consist of the feasibility of the instrument by using using the correlation product moment and Alpha Cronbach, whereas the suitability model of the instrument used confirmatory factor analysis (CFA) by using lisrel program 8.70.

The parameters was used for testing the validity namely used the product moment correlation whereas reliability test used Cronbach Alpha. Validity viwed from coefficient r> 0.3 (Norušis, 1986, p. 12). To establish the reliability of the constructs, used Cronbach's Alpha formula is at least 0.7. While testing the suitability of theoretical models with empirical data evaluation model in this study refers to the criteria of Goodness of Fit (GOF) is the Root Mean Square Error of Approximation (RMSEA) <0:08; Chi-Square were obtained from the test has a probability greater than 0.05 (p> 0.05), and Normet Fit Index (NFI), Comparative Fit Index (CFI), and Googness of Fit Index (GFI)> 0.90.

Finding and Discussion

Pre-development Stage

The results of observation, interviews, and data documentation on pre-development stage became the base as need assessment for developing model evaluation program learning of Islamic education in Madrasah Aliyah. Results of studies development modal evaluation lerning program then combined with the results of the literatures related to the program evaluation model, then translated in initial model. This initial product was developed of the model CIPP that has two stages of evaluation namely a process, and learning outcomes, hereinafter the evaluation model program named EPH models, as presented in Figure 1.

Product Development Phase

Based on the results of the pre-development studies then performed subsequent drafting of the prototype model evaluation development of Islamic Religious Education Learning program at Madrasah Aliyah. The prototype consists of: (1) components and procedures learning program evaluation model of Islamic education in Madrasah Aliyah, (2) the istruments of learning process, and learning outcomes, and (3) the using of manual evaluation model learning programs. The products had been designed as a first prototype that was developed through self evaluation validated by experts of Islamic education, and expert PEP, (expert review) based on the contents, and language. The validation method used the peers instrument validation and the Delphi technique.

Experts and Teachers PAI provided feedback and agreed with the evaluation procedures components, instruments, and evaluation guidelines. The results of validation model evaluation program developed, the item's score used a calculation coefficient item Aiken's V, as presented in Table 1.

Table 1. Summary Of The Result Validation model Evaluation (Model EPH)

Evaluation Procedure		Learning Process		Learning Outcomes		Evaluation Guidlines	
Teacher	Expert	Teacher	Expert	Teacher	Expert	Teacher	Expert
0.88	0.79	0.79	0.88	0.75	0.88	0.88	0.88
0.83	0.75	0.75	0.79	0.75	0.79	0.71	0.79
0.75	0.75	0.75	0.83	0.75	0.79	0.75	0.83
0.83	0.83	0.75	0.75	0.75	0.75	0.79	0.79
0.75	0.75	0.79	0.83	0.75	0.83	0.79	0.83
0.83	0.83	0.88	0.88	0.79	0.79	0.79	0.83
0.75	0.75	0.88	0.88	0.71	0.75	0.71	0.75
0.75	0.75	0.88	0.88	0.75	0.75	0.71	0.75
0.75	0.71	0.71	0.71	0.75	0.75	0.71	0.79
		0.75	0.75	0.83	0.83	0.71	0.75
		0.75	0.75				
		0.79	0.83				

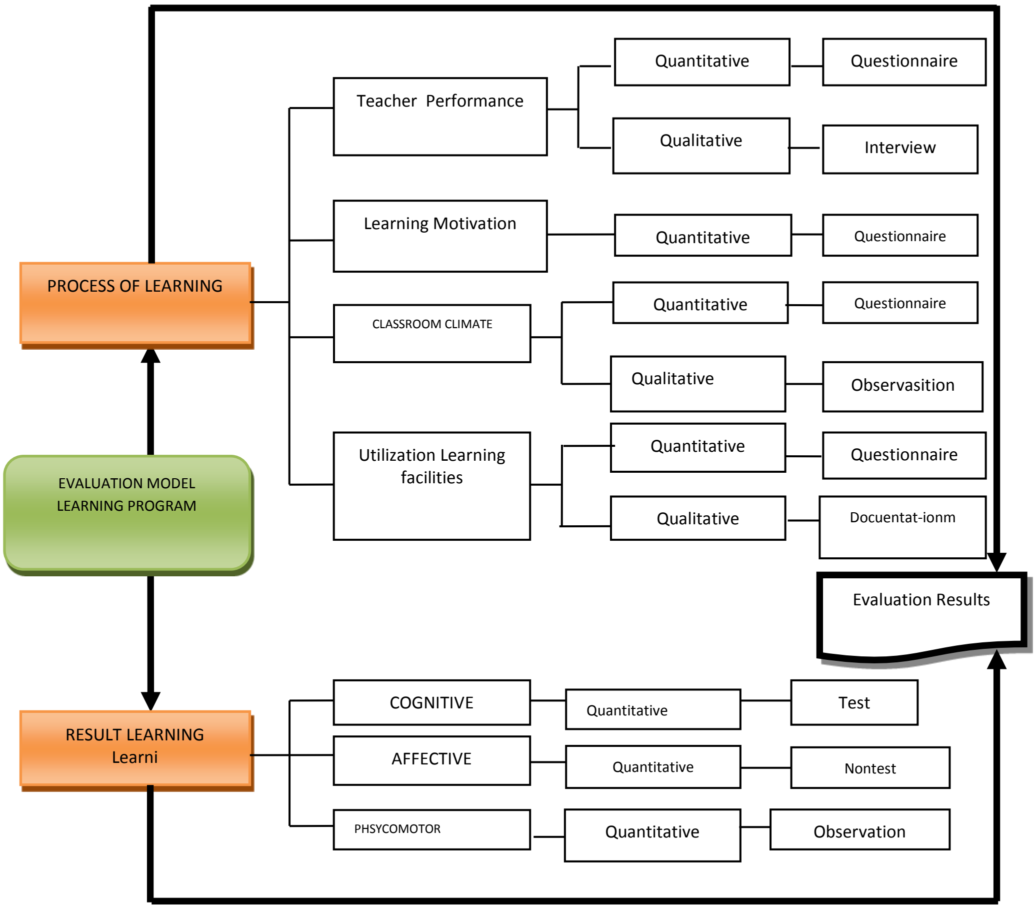


Figure 1. Model EPH

Based on the range of values Aiken's V between 0 and 1.00, then the Table 1, it could be said that the evaluation model program used has a good coefficient of content validity (content validity coefficient), because all Aiken's V above 0.7 (Azwar, 2013, p. 113).

The Trial Results EPH Model

Test Limited

Implementation of the limited trial conducted with 3 subjects trial namely headmaster and 13 teachers. The trial was designed to collect information related to the legibility of the model, in order to achieve the criteria of effective models, and practical. Information focused on the model regibility at the time used. The trial legibility evaluation models in the first phase were: (1) the legibility of the evaluation procedures, (2) the legibility of quality learning instruments, (3) instrument of learning outcomes, and (4) the legibility of the evaluation guidelines.

Assessment used a scale 5 the lowest score is 1, and the highest score 5. Based on the scores obtained on each item instrument, calculated the average score in points of the instrument. More results were presented in Table 2.

Table 2. Summary of Results Regibility Test Evaluation Model Program

Number	Legibilty	Mean Score
1.	Evaluation Procedure	4,55
2.	Process of Instrument	4.45
3.	Results of Instrument	4.40
4.	Evaluation Guidelines	4.40

Table 2 showed all the components of program evaluation models in category score of more than 4, it means that all components of the program evaluation model included in good classification and could be used.

Legibility Test Results Cognitive and Affective Domains

Legibility tests about the cognitive and affective limited involving students class X and class XII, the number of students 15 people from three madrasas. Trial limited legibility was focused on aspects of the instructions clarity answering test, communicative language used, the choice of words that easy to understand, the structure of sentences was not complicated, and the sentences were not multiple interpretations. Assessment using a scale 5 a minimum score1, maximum score 5. The results of summery mean score legibility aspect of the instrument can be presented in Table 3.

Table 3. Summary of legibility the Cognitive and Affective Instrument test

Number	Legibility	Mean Score
1.	Cognitive domain test	4,6
2.	Affective domain test	4,6

The mean whole score legibility trial results, that there were legibility test instrument about the cognitive and affective category of the mean total score more than 4. The mean total of this score if converted to the assessment criteria of quantitative data into qualitative data with a scale of 5 was set into a good category test.

The Main trial

At primary trial, there are two instruments which have been tested, namely: (1) instrument of learning process that includes: teacher performance, learning motivation, classroom climate, and utilization of learning facilities, and (2) the evaluation instrument learning outcomes consist of: instrument test questions cognitive, affective domain rubric tests and observation sheet rubric psychomotor.

Results of Instruments Learning Process

The main trial of the instruments learning process intended to indicate the feasibility of instruments, namely by estimating the instrument reliability and validity of the instruments used in this study. Based on the analysis with SPSS 19.00 for windows, can be indicated the index. The parameters used in testing validity by using the product moment correlation whereas reliability test with Cronbach Alpha. Validity viwed from coefficient r > 0.3 and reliability of Cronbach alpha coefficient > 0.7. Results summary SPPS all print out can be presented in table 4.

Table 4. Results of Testing Instrument Evaluation Model Learning Process

Aspect	Loading Factor	*Alpha Cronbach*
Teacher Performance	0.551-0.642	0.731-0.793
Learning Motivation (Students)	0.430-0.637	0.719-0755
Classroom Climate	Students 0.418-0.583	Teachers 0.447-0.674
Utility of learning Vasilities Fasilitas (Teacher)	0.446-0.843	0.889

Table 4 shows all the instruments components of the model evaluation learning process are valid and reliable, because the loading factor > 0.3 and reliability of Cronbach alpha coefficient> 0.7.

Results Instruments Learning Outcomes

Trial learning achievement test consist of: instrument of cognitive domain, affective rubric, and psychomotor observation rubric sheet. Trial cognitive learning outcomes instrument, designed to indicate the feasibility model of the instrument, namely by estimating the instrument validity and reliability of the instrument used in the study. The test results presented in Table 5.

Table 5. Results of Testing Instrument Evaluation Model Learning Outcomes

Instrument Test	loading factor	*Alpha Cronbach*
Cognitive	0.413 - 0.806	0.750 - 0.789
Affective	0.418 - 0.680	0.714 - 0.788
Psychomotor	-	0.6196 - 0.7256

Table 5. The results of the loading factor and Cronbach Alpha showed all instrument components of learning outcomes is valid and reliable, because the coefficient r > 0.3 and reliability of Cronbach alpha coefficient > 0.7.

The Results of Operations trials

CFA testing designed to evaluate the ability of the components were developed to be manifest, then used to reflect the variables evaluated. Two components tobe focus in learning program evaluation model of Islamic education is a learning process and results.

Learning Process

The learning process was reflected by the three latent constructs, namely teacher performance, student motivation and classroom climate. Results of second order CFA learning process viewed with some criteria Goodness of Fit is Root Mean Square Error of Approximation (RMSEA) <0:08; Chi-Square were obtained from the test has a probability greater than 0.05 (p> 0.05), and Normet Fit Index (NFI), Comparative Fit Index (CFI), and Googness of Fit Index (GFI)> 0.90. Visually model of CFA results were presented in Figure 2.

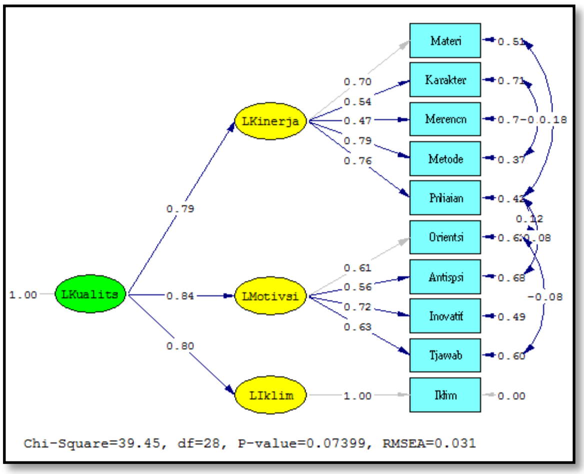


Figure 2. CFA Second Order Learning Process

Manifests of a latent variable beside to reflect significantly also should be unidimensional. These properties were evaluated by testing consruct reliability. The test results consruct reliability (CR) learning process are presented in Table 15. The test results Consruct Reliability Scond Order can be presented in Table 6.

Table 6. Test Results Consruct Reliability (CR) Second Order Learning Process

No	Conctruct	Manifests	λ	E	CR
1	Teacher Performance	Materials	0.700	0.5100	0.7917
		Character	0.540	0.7084	
		Planning	0.470	0.7791	
		Method	0.790	0.3759	
		Rating	0.760	0.4224	
2	Student learning motivation	Orientation	0.610	0.6279	0.7258
		Anticipation	0.560	0.6864	
		Inovation	0.720	0.4816	
		Responsibility	0.630	0.6031	
3	Classroom climate	climate	1	0.0000	1

Based on these results known that the construct has a coefficient CR> 0.7, it means significant in reflecting latent constructs reflected, so that the components of the learning quality process in the learning program evaluation model of Islamic education was proper to use (Hair, Black, Babin, & Anderson, 2010, p. 92).

Learning outcomes

Learning outcomes reflected by the results of learning in cognitive, affective, and psychomotor domains. The implementation of learning program evaluation model be adapted with domain measured. Subject to the implementation of the model to evaluate the cognitive domain was not the same as the affective and psychomotor domains. Therefore CFA testing performed on each construct.

Cognitive Domain

Evaluation of cognitive lesson domain PAI in Madrasah Aliyah consisted of subjects Qur'an Hadits, Morals, Fiqh and Islamic Cultural History (ICH).

Knowledge of Al- Qur'an and the Hadis reflected by ten manifest grouped into three constructs, namely definition, basic understanding of the content, and functionality of the Qur'an. Second order CFA results are presented in Figure 3. The coefficient of chi-square = 41.31 with p value = 0.10198 > 0.05 indicated measurement model Qur'an-Hadith suitable with the population. When viewed from the other GOF parameters such as NFI, CFI and GFI more than 0.9. and RMSEA ≤ 0.08 indicates that the model was fit so it was not necessary to change the model. Visually model of CFA results are presented in Figure 3.

Akidah akhlak was reflected by fourteen manifest, which are grouped into four constructs, namely the principles of Akidah, Tauhid, Syiriq, and Morals. Second order CFA results were presented in Figure 4. The coefficient of chi-square = 83.78 with p value = 0.16167 > 0.05 indicate that measurement model of akidah akhlak suitable with the population. GOF parameter values such as NFI, CFI and GFI more than 0.9. and RMSEA = 0.029 ≤ 0:08 indicates that the model fit. Visually model of CFA results are presented in Figure 4

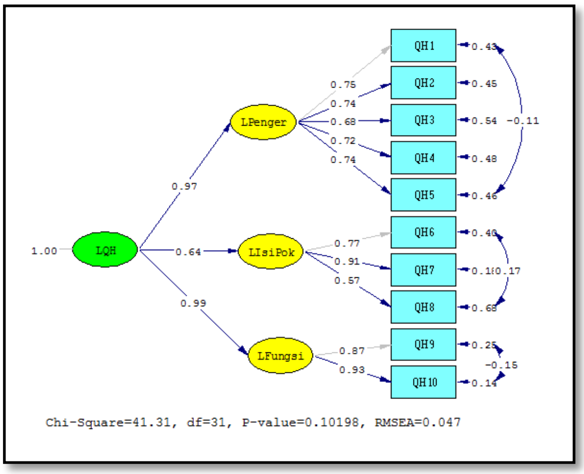


Figure 3. CFA second Order Qur'an Hadith

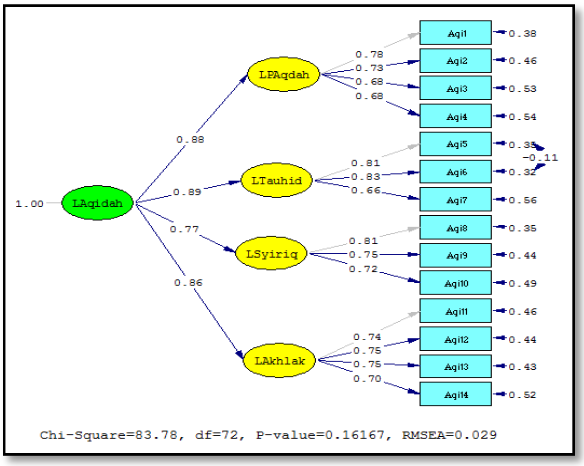


Figure 4. CFA Second Order Akidah Akhlak

Fikih was reflected by twelve manifest grouped into four constructs, namely Principle of Worship, alms Law, sacriface Wisdom, and the pilgrimage law. Results of second order CFA was presented in Figure 5. The coefficient chi-square = 61.56 with p value = 0.17116 > 0.05 indicates measurement model of fikih was suitable with the population. GOF parameter values such as NFI, CFI and GFI more than 0.9. and RMSEA = 0.030 ≤ 0.08 indicated that the model fit. Visually model of CFA results are presented in Figure 5.

Islamic Cultural History was reflected by the ten manifest, which are grouped into four constructs, namely: Understanding exemplary preaching Prophet in guiding the people, Understanding the leadership issue of Muslims after the Prophet's death, Understanding the probelem of islamic leadership after death Muhamad prophet, understanding the islamic development at the middle period clasical or the golden age and Understanding the development of Islamic in the medieval period/the dark age. Second order CFA results were presented in Figure 6. The coefficient of chi-square = 42.25 with p value = 0.06811 > 0:05 indicates that measurement model of Islamic Cultural History fits with the population. GOF parameter values such as NFI, CFI and GFI more than 0.9. and RMSEA = 0.064 ≤ 0:08 indicates that the model fit. Visually model of CFA results are presented in Figure 6.

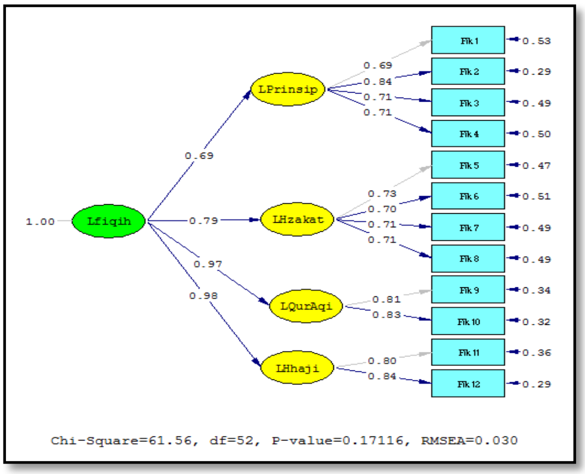


Figure 5. CFA Second Order Fikih

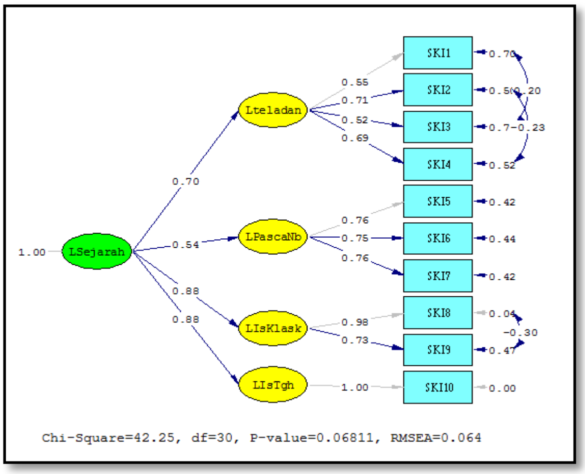


Figure 6. The Second Order CFA SKI

Testing unidimensionality at the second order gained coefficient Consruct reliability > 0.7 for each construct. it Means that each manifest used to reflect the construct was unidimensional (Hair et al., 2010, p. 92). Results Consruct Reliability Second Order. Learning Outcomes Cognitive Domains were presented in Table 7.

Table 7. The Results of Consruct Reliability Test (CR) Scond Order Learning Outcomes Cognitive Domains

Construct	λ	E	CR
Al-Qur'an Hadis	0.57-0.93	0.1351-0.6751	0.8050-0.8250
Akidah Akhlak	0.66-0.83	0.439-0.5376	0.8050-0.8250
Fikih	0.69-0.84	0.2944-0.5239	0.84-0.828
SKI	0.52-1	0-0.7296	1- 0.714

Based on the significant results of the manifest and latent constructs reflected; unidimensional nature manifest in the second group order, and constructs in the first order. Then the Islamic Cultural History component in the learning program evaluation model of Islamic education is proper to be used.

Affective Domain

Affective domain was reflected by twelve manifest grouped into four constructs, namely: Interests (label: L intreststst), Attitude (label: L attitude), Discipline (label: LDicipline), and Cooperation (label: L cooperation). Results of second order CFA wa presented in Figure 7. The coefficient of chi-square = 118.66 with p value = 0.08680 > 0:05 indicated that measurement model of affective domain was suitable with the population. GOF parameter values such as NFI, CFI and GFI more than 0.9. and RMSEA = 0.023 ≤ 0:08 indicates that the model fit. Visually model of CFA results were presented in Figure 7.

Testing unidimensionality at the second order gained coefficient Consruct reliability > 0.7 for each construct. it Means that each manifest used to reflect the construct was unidimensional (Hair et al., 2010, p. 92). Results Consruct Second Order Reliability Domains Affective presented in Table 8.

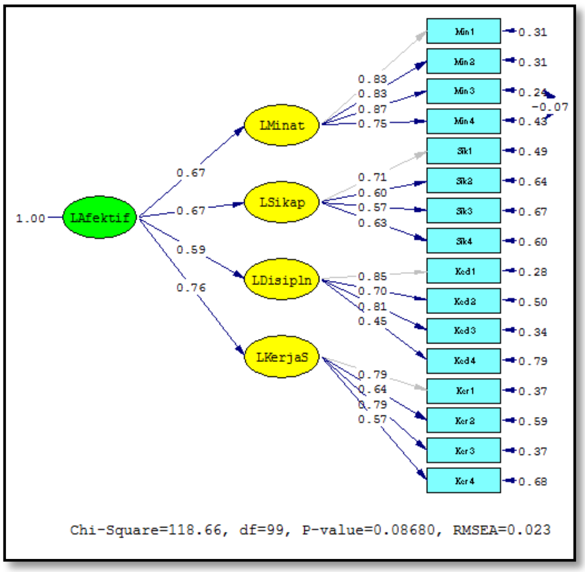


Figure 7. CFA Second Order Affective

Table 8. The results of Consruct Reliability Tes (CR) Scond Order Affective Domains

No	Construct	Manifest	λ	E	CR
1	Linterest	Min1	0.83	0.311	0.892
		Min2	0.83	0.311	
		Min3	0.87	0.2431	
		Min4	0.75	0.438	
2	Lattitude	Sik1	0.71	0.496	0.723
		Sik2	0.6	0.640	
		Sik3	0.57	0.675	
		Sik4	0.63	0.603	
3	Ldicipline	Ked1	0.85	0.278	0.804
		Ked2	0.7	0.510	
		Ked3	0.81	0.344	
		Ked4	0.45	0.798	
4	Lcooperating	Ker1	0.79	0.376	0.794
		Ker2	0.64	0.590	
		Ker3	0.79	0.376	
		Ker4	0.57	0.675	

Based on the significant results of the manifest and latent variables constructs was reflected, unidimensional nature manifest in the second group order. Then the affective component in the learning program evaluation model of Islamic education is proper to be used.

Psychomotor Domain

Psychomotor learning outcomes was presented through the practice of ablution and prayer. The measurement is reflected by nineteen manifest that grouped in two constructs, namely: the practice of ablution and prayer. the Second order CFA results were presented in Figure 8. The coefficient of chi-square = 148.08 with p value = 0.09726 > 0.05 indicates that Psychomotor Domains measurement model was suitable with the population. GOF parameter values such as NFI, CFI and GFI more than 0.9. and RMSEA = 0.041 ≤ 0.08 indicated that the model fit. Visually model of CFA results were presented in Figure 8.

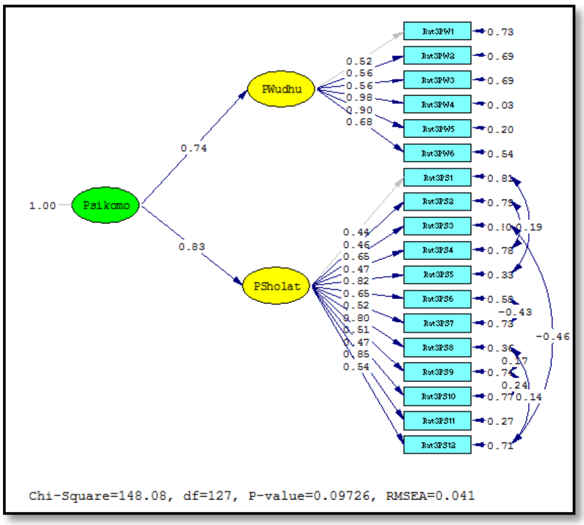


Figure 8. CFA Second Order Psychomotor

Testing unidimensionality at the second order gained coefficient Consruct reliability> 0.7 for each construct. Means each manifest used to reflect the construct is unidimensional (Hair et al., 2010, p. 92), as the result of second order reliability construct ablution and prayer practices were presented in Table 9.

Based on the significant results of the manifest and latent constructs was reflected unidimensional nature of the group's manifest in the second order, and constructs the first order. So the psychomotor component on evaluation model of Islamic education learning programs is proper to be used.

Table 9. The Results of Construct Reliability Test (CR) Sconde Order Learning Process

Number	Construct	Manifest	λ	E	CR
1	Ablution Practice	Rat3PW1	0.52	0.7296	0.860
		Rat3PW2	0.56	0.6864	
		Rat3PW3	0.56	0.6864	
		Rat3PW4	0.98	0.0396	
		Rat3PW5	0.90	0.19	
		Rat3PW6	0.68	0.5376	
2	Praying Paractice	Rat3PS1	0.44	0.8064	0.874
		Rat3PS2	0.46	0.7884	
		Rat3PS3	0.65	0.5775	
		Rat3PS4	0.47	0.7791	
		Rat3PS5	0.82	0.3276	
		Rat3PS6	0.65	0.5775	
		Rat3PS7	0.52	0.7296	
		Rat3PS8	0.80	0.36	
		Rat3PS9	0.51	0.7399	
		Rat3PS10	0.47	0.7791	
		Rat3PS11	0.85	0.2775	
		Rat3PS12	0.54	0.7084	

Study of the Last Product

Model of EPH has been tested, both quantitative and qualitative. The trial results quantitatively in medium scale by using SPSS program and field operations (operational field testing) used CFA confirmatory analysis program, the results showed that the model evaluation instrument EPH it have been fullfilled reliability coefficient, the items of have a instrument were valid, and the model was fit. Further test results are qualitatively EPH model evaluation, the results showed that the implementation of the evaluation model was practical, effective, and efficient.

Evaluation EPH models have been equipped components and evaluation procedures, the learning process instruments, instrument results (outputs) of learning, evaluation guidelines was very detailed and clear. Thus the model was applied in Madrasah Aliyah (MA), MTs (MTs), can also be applied to public schools (SMA and SMP).

Characteristics of Evaluation Model

(1) The model is used to evaluate the learning Islamic education program in Madrasah Aliyah

(2) The Evaluation model is the internal evaluation conducted by headmater and teachers to monitor the learning process and learning outcomes

(3) the Model evaluation learning program has two components evaluating learning programs, namely the evaluation of the learning process and evaluation of learning outcomes.

(4) Using of this model is not dependen on a particular learning program evaluation model that implemented by headmaster and PAI teachers.

(5) EPH model can be applied to Curriculum 2013

(6) This model is open for further development, and have not become the final product.

Implication of Evaluation Model

(1) Evaluation EPH models is very complete and comprehensive to evaluate learning process that includes evaluating teacher performance, student motivation, classroom climate, and utilization of learning facilities, conducted headmaster and teachers PAI

(2) Evaluation EPH models is very complete and comprehensive to evaluate learning outcomes include: evaluation of learning outcomes cognitive, affective, and psychomotor, committed teachers Islamic education

(3) This model is very practical to use, easy tobe understood by the user, and easy tobe implemented at the islamic school or public school.

Research Limitations

(1) The evaluation process have not involved the external parties that are independent but rely solely on the ratings of internal madrassas so it is possible there is an element of subjectivity assessment of implementing.

(2) Requires substantial costs mainly related to the implementation of testing, validation experts, practitioners, as well as regibility test. High financial also influenced by the location of the research is so far.

(3) Limitations of validator involved in this study were derived from a limited number of PAI teachers.

Conclusions

Conclusion based on results of this research consists of: (1) evaluation model of Islamic Education Learning program developed consist of: (a) the evaluation procedures, (b) the evaluation instrument (c) the evaluation guide; (2) according to experts, PAI Teachers and headmaster of Madrasa, procedures, instruments, and evaluation guidelines are already well developed, 3) The instrument developed entirely valid (load factor> 0.3), reliable (CR> 0.7), and qualify as a fit model (RMSEA ≤0.08 and NFI, CFI and GFI> .90). and 4) the assessment of the experts, PAI teacher and headmaster, EPH models asserted effective, practical, and easy to use, supported with a valid and reliable instrument.

Based on the conclusions of research, socould be suggested that: (1) the user should read and understand guidelines to the models that have been prepared, (2) the instruments are given to students adjusted to the phase of the evaluation is continuing, time, and the right conditions, (3) using models EPH, can be adapted to the purpose, useful and level of desired program, (4) in learning process teachers should be able to be a good guidance, maintain the performance, creating a conductive classroom climate, motivating, utilize of learning facilities in order to reach the good learning students achievements, and (5) this product should be used as guidelines in evaluating the learning process and learning outcome PAI because it able to reach the side of the cognitive, affective, and students' psychomotor comprehensively.

References

Agustian, A.G. (2003). *ESQ power*. Jakarta: Arga.

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

Borg, W. R., Gall, J. P., & Gall. M. D. (1993). *Applying educational research: A practical guide*. New York: Longman.

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). New York: Pearson Prentice Hall.

Mardapi, D. (2000). Evaluasi pendidikan. In *Konvensi Pendidikan Nasional 19-23 September 2000*. Universitas Negeri Jakarta.

Morrison, D. M., Mokashi, K., & Cotter, K. (2006). *Instructional quality indicator: Research foundations*. Cambrigde: Cambrigde University. Retrieved Mei 13, 2011, from www.co.nect.net.

Muhaimin. (2009). *Rekonstruksi pendidikan islam (dari paradigma pengembangan, manajemen kelembagaan, kurikulum hingga strategi pengembangan)*. Jakarta: PT Grasindo Persada.

Norušis, M. J. (1986). *SPSS/PC+ for theimbbc/xc/at*. Chicago: SPSS Inc.

Sudjana, N. (2002). *Dasar-dasar proses belajar dan mengajar*. Bandung: Sinar Baru Algesindo.

Republik Indonesia. Undang-Undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional (2003).

Sutrisno. (2006). *Pendidikan islam yang menghidupkan*. Yogyakarta: Kota Kembang.

Wahyudi. (2003). Penyusunan dan Validasi Kuesioner Iklim Lingkungan Pembelajaran di Kelas. *Jurnal Pendidikan dan Kebudayaan*, *043*.

Zamroni. (2005). Mengembangkan kultur sekolah menuju pendidikan yang bermutu. In *Seminar Nasional Mengembangkan Kultur Sekolah di Yogyakarta* Retrieved November 23, 2005.

Jurnal Penelitian dan Evaluasi Pendidikan
Volume 21, No 1, June 2017 (34-50)
Online: http://journal.uny.ac.id/index.php/jpep

PENGARUH UKURAN SAMPEL DAN INTRACLASS CORRELATION COEFFICIENTS (ICC) TERHADAP BIAS ESTIMASI PARAMETER MULTILEVEL LATENT VARIABLE MODELING: STUDI DENGAN SIMULASI MONTE CARLO

Muhammad Dwirifqi Kharisma Putra [1]*, *Jahja Umar* [1], *Bahrul Hayat* [1], *Agung Priyo Utomo* [2]
[1]UIN Syarif Hidayatullah Jakarta, [2]Sekolah Tinggi Ilmu Statistik
[1]Ciputat, Cempaka Putih, Ciputat Timur, Tangerang Selatan, Banten 15412, Indonesia
[2] Kampung Melayu, Jatinegara, RT.1/RW.4, Jakarta Timur, DKI Jakarta 13330, Indonesia
* Corresponding Author: muhammad.dwirifqi@gmail.com

Abstrak

Studi ini menggunakan simulasi *Monte Carlo* dilakukan untuk melihat pengaruh ukuran sampel dan *intraclass correlation coefficients* (ICC) terhadap bias estimasi parameter *multilevel latent variable modeling*. Kondisi simulasi diciptakan dengan beberapa faktor yang ditetapkan yaitu lima kondisi ICC (0.05, 0.10, 0.15, 0.20, 0.25), jumlah kelompok (30, 50, 100 dan 150), jumlah observasi dalam kelompok (10, 20 dan 50) dan diestimasi menggunakan lima metode estimasi: ML, MLF, MLR, WLSMV dan BAYES. Jumlah kondisi keseluruhan sebanyak 300 kondisi dimana tiap kondisi direplikasi sebanyak 1000 kali dan dianalisis menggunakan software Mplus 7.4. Kriteria bias yang masih dapat diterima adalah < 10%. Hasil penelitian ini menunjukkan bahwa bias yang terjadi dipengaruhi oleh ukuran sampel dan ICC, penelitian ini juga menujukkan bahwa metode estimasi WLSMV dan BAYES berfungsi lebih baik pada berbagai kondisi dibandingkan dengan metode estimasi berbasis ML.

Kata kunci: *multilevel latent variable modeling, intraclass correlation coefficients, Metode Markov Chain Monte Carlo*

THE IMPACT OF SAMPLE SIZE AND INTRACLASS CORRELATION COEFFICIENTS (ICC) ON THE BIAS OF PARAMETER ESTIMATION IN MULTILEVEL LATENT VARIABLE MODELING: A MONTE CARLO STUDY

Abstract

A monte carlo study was conducted to investigate the effect of sample size and intraclass correlation coefficients (ICC) on the bias of parameter estimates in multilevel latent variable modeling. The design factors included (ICC: 0.05, 0.10, 0.15, 0.20, 0.25), number of groups in between level model (NG: 30, 50, 100 and 150), cluster size (CS: 10, 20 and 50) to be estimated with five different estimator: ML, MLF, MLR, WLSMV and BAYES. Factors were interegated into 300 conditions (4 NG × 3 CS × 5 ICC × 5 Estimator). For each condition, replications with convergence problems were exclude until at least 1.000 replications were generated and analyzed using Mplus 7.4, we also consider absolute percent bias <10% to represent an acceptable level of bias. We find that the degree of bias depends on sample size and ICC. We also show that WLSMV and BAYES estimator performed better than ML-based estimator across varying sample sizes and ICC's conditions.

Keywords: *multilevel latent variable modeling, intraclass correlation coefficients, Markov Chain Monte Carlo method*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v21i1.12895

Pendahuluan

Dalam bidang penelitian sosial terkadang diperoleh struktur data yang merupakan data hirarki *(hierarchical)*. Data yang terstruktur hirarki merupakan data yang timbul karena individu-individu terkumpul dalam kelompok-kelompoknya, dimana individu-individu dalam kelompok yang sama memiliki karakteristik yang cenderung sama. Struktur hirarki mengindikasikan bahwa data yang dianalisis berasal dari beberapa *level*, dimana *level* yang lebih rendah tersarang pada *level* yang lebih tinggi. Permodelan *multilevel* umumnya digunakan pada data yang berstruktur hirarki (*clustered, nested*) (de Leeuw & Meijer, 2008; Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002). Ada dua masalah utama yang mungkin muncul akibat mengabaikan struktur hirarki. Pertama, jika analisis dilakukan dengan mengumpulkan data unit *level* rendah ke unit *level* tinggi (*aggregation*), maka banyak informasi yang hilang dan analisis statistika menjadi kehilangan kekuatan. Kedua, apabila seorang peneliti tidak hati-hati dalam menginterpretasi hasil maka akan menimbulkan kesalahan seperti menganalisis data pada salah satu *level* dan merumuskan kesimpulan pada *level* lain. (Kaplan, Kim & Kim, 2009; Muthén, 1994).

Metode analisis data pada umumnya menggunakan asumsi berupa penyederhanaan bahwa data yang telah diperoleh sebagai *simple random sample* dari suatu populasi tertentu. Ini melibatkan asumsi *independently and identically distribured observartion (IID)*. Data pendidikan banyak yang, bagaimanapun, diperoleh melalui, desain sampel yang kompleks, seperti *multistage sampling* yang melibatkan pengamatan berbentuk *cluster* di mana asumsi IID tidak realistis. (Muthén, 1991). *Multistage sampling* digunakan ketika sampel diambil secara acak dari unit yang lebih tinggi dan disampel dari populasi yang lebih besar dari unit tersebut. Hasil pengambilan sampel menggunakan metode *multistage sampling* pada data hierarki yang terstruktur (misalnya, siswa berkumpul di dalam satu unit kelas), membuat residual bergantung pada variasi *between-cluster*. Skor pada variabel yang diamati dari anak-anak dalam satu kelas mungkin lebih mirip daripada anak-anak di kelas yang berbeda, misalnya. Apabila mengabaikan struktur data hirarki dapat terjadi bias pada estimasi hubungan antar item (Geldhof, Preacher & Zyphur, 2014).

Selama bertahun-tahun, pemodelan *multilevel* telah mengalami perkembangan yang sangat pesat dan aplikasi untuk mengklaim bahwa pemodelan *multilevel* sekarang tegas berlindung di berbagai metodologi untuk penelitian ilmu sosial dan perilaku. Selain itu, saat ini telah banyak dilakukan untuk mengintegrasikan model bertingkat dengan *structural equation modeling* (SEM) sehingga memberikan metodologi umum yang dapat menjelaskan masalah kesalahan pengukuran, mediasi, dan simultanitas. Hal ini berguna, karena itu, untuk memberikan gambaran tentang metodologi, memeriksa perkembangan terbaru, dan menawarkan jalan untuk upaya penelitian masa depan (Kaplan, Kim & Kim, 2009).

Dalam berbagai kasus, penggunaan metode CFA yang di desain untuk data berstruktur *single-level* akan menyebabkan bias pada estimasi parameter dan *standard error* (Julian, 2001). MCFA yang merupakan submodel dari MSEM, telah dikembangkan untuk mengatasi permasalahan ini. (Muthén & Asparouhov, 2009). Tetapi teknik ini bukan merupakan teknik dengan sampel kecil. Secara khusus, teknik ini dapat digunakan untuk data yang memiliki jumlah kelompok yang cukup besar, setidaknya sekitar 50-100. Seperti yang kemukakan oleh Cronbach (Muthén, 1991), jika terbentur biaya, mungkin lebih baik mengamati siswa per kelas yang lebih sedikit untuk membantu menambah kelas yang lebih banyak. Sejalan dengan pendapat tersebut Hayes (2006) juga berpendapat bahwa permodelan *multilevel* merupakan prosedur dengan ”sampel yang besar”, yang berarti bahwa analisis matematika dan asumsi teoritis yang mendasari statistik yang dihasilkan oleh permodelan *multilevel* didasarkan pada analisis statistik yang dihitung dalam sampel yang besar. Seperti biasa, betapa besar cukup besar adalah pertanyaan yang sulit dijawab, karena tergantung pada banyak hal. Lebih parah lagi, ukuran sampel yang dibutuhkan adalah fungsi dari kedua jumlah unit *level* 1 dan jumlah unit *level* 2.

Dalam mendesain penelitian *multilevel*, peneliti harus memberi perhatian pada berbagai permasalahan terkait ukuran sampel dan bagaimana distribusi dari individu dalam unit yang akan diteliti (Heck & Thomas, 2015). Sejalan dengan pendapat tersebut, ukuran sampel yang cukup merupakan salah satu masalah yang paling penting dalam pemodelan *multilevel*, Kondisi desain yang paling dasar seperti sejumlah kelompok pada setiap tingkat analisis dan ukurannya menentukan kemampuan untuk memperoleh hasil estimasi koefisien regresi yang *unbiased* serta *standard errors* dengan besaran yang masih dapat diterima dan kekuatan tes (Łaszkiewicz, 2013; Snijders, 2005).

Selain itu, Busing (Łaszkiewicz, 2013) menemukan ukuran sampel yang tidak mencukupi (10 sampai 50 kelompok dengan 5 atau 10 orang) mungkin menjadi penyebab yang membuat model menjadi tidak kon-vergen. Meskipun sifat *asymptotic* dari estimator pada model *multilevel* (seperti REML atau IGLS), penggunaan ukuran sampel yang lebih besar menjamin pengurangan bias, di tengah ketertarikan mengetahui batas bawah dari sampel (Mass & Hox, 2005). Dengan demikian, ukuran sampel yang memadai (cukup) bisa dianggap sebagai sampel minimum, yang menjamin *unbiasedness* (atau lebih tepatnya: ukuran bias yang masih bisa diterima).

Perbedaan pada estimasi parameter dengan metode *single-level* dan dibandingkan dengan metode *clustering* (model *multilevel*) bergantung pada beberapa aspek dari data. (Pornprasertmanit, Lee & Preacher, 2014). Dengan pengukuran konstruk tingkat *cluster,* berbagai langkah-langkah pengukuran digunakan untuk mengevaluasi apakah respon dari item menunjukkan besaran dari *clustering* (pengelompokan) seperti yang diharapkan pada konstruk tingkat *cluster*. Salah satu ukuran yang sering digunakan adalah ICC. (Shrout & Fleis, 1979 dalam Stapleton, Yang & Hancock, 2016).

ICC merupakan salah satu hal yang harus diperhitungkan karena hal ini merupakan hal penting yang dapat merubah varians error pada model regresi linier sederhana (Kreft & de Leeuw, 1998). ICC merupakan faktor penting yang menjelaskan variabilitas keseluruhan yang dijelaskan oleh unit *level* 2 yang menjadi faktor penting dalam permodelan *multilevel* (Julian, 2001; Pornprasertmanit, Lee & Preacher, 2014). Terdapat keterkaitan antara rendahnya ICC dan besarnya bias pada estimasi parameter pada model *between-level* (Preacher, Zhang & Zyphur, 2011) serta rendahnya ICC dan kaitannya dengan rendahnya tingkat konvergensi (Kim, Kwok & Yoon, 2012 dalam Hsu et. al., 2016). Metode estimasi dengan estimator *maximum likelihood* juga tidak dapat bekerja dengan baik saat nilai ICC rendah ataupun jumlah individu pada *cluster* dibawah 50 orang. (Hox & Mass, 2004 dalam Stapleton, Yang & Hancock, 2016). ICC nilainya berkisar dari 0 ke 1, nilai ICC yang tinggi menunjukkan proporsi varians yang lebih besar dari varians pada tingkat "antar" dan bias sehingga kemungkinan besar jika sifat data bertingkat tersebut tidak diperhitungkan. (Dyer, Hanges & Hall, 2005).

Untuk menjawab permasalahan tersebut, baik permasalahan terkait ukuran sampel dan juga permasalahan terkait besarnya ICC dapat dilakukan studi dengan simulasi data yang memungkinkan peneliti menjawab pertanyaan-pertanyaan seputar masalah tersebut, salah satunya dengan menggunakan simulasi *Monte Carlo.* Terdapat beberapa penelitian yang menggunakan simulasi *Monte Carlo* pada analisis data *multilevel* seperti untuk menginvestigasi efek dari mengabaikan *clustering* dari data serta berfokus pada ICC dan faktor penting lain sebagai penyebabnya (Pornprasertmanit, Lee & Preacher, 2014),

Studi dengan simulasi data mulai populer pada akhir abad ke 19 dan awal abad ke 20 yang digunakan pada berbagai bidang ilmu pengetahuan. (Feinberg & Rubright, 2016). Terdapat berbagai metode dari simulasi data, salah satunya adalah simulasi *Monte Carlo* dengan algoritma yaitu *Markov Chain Monte Carlo* (MCMC) yang menggunakan pemilihan angka secara acak untuk menyelesaikan permasalahan *modeling* yang sulit ini, metode ini diperkenalkan pada bidang psikometri oleh Patz dan Junker tahun 1999 (Gelfand & Smith, 1990; Patz & Junker, 1999 dalam Feinberg & Rubright, 2016). Umumnya, studi menggunakan simulasi *Monte Carlo* yang digunakan pada bidang SEM untuk mempelajari sifat dari estimator dan uji statistik yang digunakan pada berbagai kondisi yang dimanipulasi oleh peneliti, seperti ukuran sampel, besarnya kesalahan spesifikasi model dan tidak normalnya data (Brown, 2006). Sehingga dari latar belakang yang telah dijelaskan sebelumnya, akan diuji pengaruh ukuran sampel dan *intraclass correlation coefficients* (ICC) terhadap bias estimasi parameter multilevel *latent variable modeling* dengan menggunakan metode berbasis pendekatan *Bayesian* yaitu *markov chain monte carlo* (MCMC).

Para peneliti telah lama menyadari permasalahan ini, dalam bidang pendidikan telah terjadi perdebatan yang dimulai mengenai permasalahan *'unit of analysis'* pada tingkat berbeda (Burstein, 1980). Sebelum permodelan *multilevel* berkembang dan menjadi metode penelitian yang umum digunakan, permasalahan yang diakibatkan karena mengabaikan struktur hirarki dari data masih dapat diterima dan dimengerti karena sulit untuk menyeselaikan masalah tersebut karena perangkat lunak yang belum tersedia (Goldstein, 2011). Tetapi pada saat ini telah banyak metode yang dapat digunakan untuk mengatasi permasalahan terkait metode analisis yang tepat untuk menganalisis data dengan struktur hirarki yang akan dijelaskan pada bagian selanjutnya.

Konsep Dasar Multilevel Latent Variable Modeling

Untuk mengatasi keterbatasan yang terkait dengan pendekatan bermasalah untuk analisis data yang berstruktur hirarki, Para peneliti telah membuat analisis untuk bentuk data hirarkis yang memungkinkan untuk pemodelan sesuai sistem organisasi seperti sekolah. Selain perkembangan statistik, kemajuan perangkat lunak sekarang memungkinkan estimasi relatif mudah untuk model *multilevel*, dan pemodelan seperti sekarang umum dilakukan dalam ilmu-ilmu sosial dan perilaku (Kaplan, Kim & Kim, 2009). Terdapat pengembangan dari analisis *multiple regression* yang dikenal dengan *multilevel modeling* (MLM) yang juga dikenal dengan *hierarchical linear modeling, random coefficient modeling* atau *mixed effect modeling* yang merupakan metode statistik berbasis regresi yang digunakan apabila data terstruktur hirarki atau dalam *cluster* (Kreft & de Leeuw, 1998; Raudenbush & Bryk, 2002) sedangkan pengembangannya pada penggunan variabel laten didalamnya yaitu *multilevel latent variabel modeling* (MLLVM) yang perkembangannya dirangkum oleh berbagai literatur (misal, Kaplan, Kim & Kim, 2009). Pendekatan ini didasarkan pada metode SEM. Muthén & Satorra (1995) menjelaskan pendekatan SEM yang digunakan pada data yang berasal dari survei yang kompleks dapat berupa agregasi dan disagregasi. Pendekatan berbasis SEM ini juga dikenal dengan nama *covariance structure analysis* yang menggunakan model matematika untuk menjelaskan matriks kovarians dari sekumpulan variabel dari faktor yang jumlahnya lebih sedikit dan dapat diperpanjang untuk menganalisis struktur yang ada. (Heck & Thomas, 2015; Muthén, 1994). Pada penelitian ini *multilevel latent variable modeling* yang dimaksud, dibatasi pada penggunaan metode *multilevel* CFA yang diperkenalkan oleh Muthén (1991, 1994).

Konsep Dasar Multilevel CFA

Stapleton, Yang & Hancock (2016) menjelaskan bahwa sebuah permasalahan muncul pada CFA apabila data dikumpulkan dari individu dengan struktur hirarki ataupun *nested*. Pengukuran pada tingkat individu diharapkan responnya relevan dengan hipotesis dari konstruk dan oleh karena itu skor pada pengukuran tersebut harus mencerminkan variabilitas individu.

Kaplan, Kim & Kim (2009) menjelaskan bahwa untuk memulai pendekatan *Multilevel Confirmatory Factor Analysis* (MCFA), dimulai dari mempertimbangkan model yang terurai p-dimensi vektor respon y_{ig} untuk siswa *i* di sekolah *g* ke dalam *sum of a grand* *mean* μ, bagian *between-group* ν_g dan bagian *within-group* u_{ig}. Sebagai berikut:

$$y_{ig} = \mu + \nu_g + u_{ig} \ldots (1)$$

Matriks kovarians untuk vektor respon dapat ditulis sebagai $\mathbf{y}_{ig}$:

$$\Sigma_T = \Sigma_B + \Sigma_W \ldots (2)$$

di mana $\boldsymbol{\Sigma}_T$ adalah matriks kovarians keseluruhan populasi, $\boldsymbol{\Sigma}_b$ adalah matriks kovarians populasi *between-group*, dan $\boldsymbol{\Sigma}_w$ adalah matriks kovarians populasi *within-group*. Jumlah sampel dapat didefinisikan sebagai:

$$\bar{y}_{.g} = \frac{1}{n_g} \sum_{i=1}^{n_g} \bar{y}_{ig} \ldots (3)$$

$$\bar{y} = \frac{1}{N} \sum_{g=1}^{G} \sum_{i=1}^{n_g} \bar{y}_{ig} \ldots (4)$$

$$S_W = \frac{1}{N-G} \sum_{g=1}^{G} \sum_{i=1}^{n_g} (\bar{y}_{ig} - \bar{y}_{.g})(\bar{y}_{ig} - \bar{y}_{.g})' \ldots (5)$$

$$S_B = \frac{1}{G-1} \sum_{g=1}^{G} n_g (\bar{y}_{.g} - \bar{y})(\bar{y}_{.g} - \bar{y})' \ldots (6)$$

di mana $\bar{\mathbf{y}}_{\cdot g}$ adalah *mean* sampel untuk kelompok *g*, $\bar{\mathbf{y}}$ adalah *grand mean*, S_w adalah matriks kovarians *sample pooled within group*, dan S_b adalah matriks kovarians *between groups*. Seperti aplikasi standar regresi linier pada data yang didapat dari pengambilan sampel secara *multistage*, penerapan analisis faktor juga harus memperhitungkan efek bersarang *(nested)*. Misalnya, sekumpulan item skala sikap digunakan untuk menilai persepsi siswa tentang iklim sekolah dan diberikan kepada siswa kemungkinan besar berlawanan dengan variabilitas antara sekolah. Mengabaikan variabilitas 'antar-sekolah' pada sejumlah siswa di sekolah-sekolah akan menghasilkan bias terhadap prediksi dalam parameter model analisis faktor. Oleh karena itu, diperlukan untuk memperpanjang metodologi *multilevel* dengan kerangka analisis faktor. (Kaplan, Kim & Kim, 2009). Untuk memulai, kita asumsikan bahwa vektor dari respon siswa dapat dinyatakan dalam *multilevel linear factor model* sebagai berikut:

$$y_{ig} = \nu + \Lambda_w \eta_{w_{ig}} + \Lambda_b \eta_{b_g} + \epsilon_{w_{ig}} + \epsilon_{b_g} \ldots (7)$$

dimana y_{ig} sama seperti yang didefinisikan sebelumnya, ν adalah *grand mean*, Λ_w adalah matriks muatan faktor untuk kelompok *within*, ηw_{ig} merupakan faktor yang bervariasi secara acak di seluruh unit kelompok *within*, Λ_b adalah matriks muatan faktor kelompok *between*, ηb_g merupakan faktor yang bervariasi acak di seluruh kelompok, ϵw_{ig} dan ϵb_g merupakan *within* dan *between group uniqueness*. Menggunakan asumsi standar analisis faktor linear, di sini diperluas untuk kasus *multilevel*, matriks kovarians total didefinisikan dalam Persamaan (2) dapat dinyatakan dalam hal model faktor parameter sebagai:

$$\Sigma_T = \Lambda_W \Phi_W \Lambda_W' + \Theta_W + \Lambda_b \Phi_b \Lambda_b' + \Theta_b \ldots (8)$$

dimana Φ_w dan Φ_b adalah matriks kovarians faktor untuk *within group* dan *between group* dan Θ_w dan Θ_b adalah matriks diagonal dari *unique variances* untuk bagian *within group* dan *between group*. Secara umum, biasanya mudah untuk menentukan struktur faktor untuk variabel *within school*. Hal ini juga mudah untuk memungkinkan variabel *within school* bervariasi *between school*. Kesulitan konseptual sering timbul untuk menjamin struktur faktor untuk menjelaskan variasi *between groups* (Kaplan, Kim & Kim, 2009). Sedangkan penelitian terbaru dari Geldhof, Preacher dan Zyphur (2014) menjelaskan bahwa MCFA merupakan ekstensi dari CFA yang dapat mengakomodir data *two-level* dan memungkinkan estimasi secara terpisah dan analisis matriks kovarians *within-* dan *between-cluster*.

Ada beberapa pendekatan dalam literatur metodologi untuk melakukan MCFA (mis, Muthén, 1994). terdapat metode baru yang dikembangkan oleh Muthén dan Asparouhov (2009, 2011) untuk melakukan analisis MSEM. MCFA adalah kasus khusus dari MSEM tanpa jalur struktural yang menghubungkan variabel laten, dalam banyak cara yang sama bahwa *single-level* CFA adalah kasus khusus dari SEM. Yuan dan Bentler (2007) menjelaskan berbagai keuntungan dari penggunaan metode *multilevel* SEM yaitu: (1) lebih mudah untuk mengetahui di *level* mana terjadi kesalahan spesifikasi; (2) indeks fit dari SEM konvensional dapat dengan mudah dikembangkan untuk mengevaluasi model pada *level* yang terpisah dari model *multilevel;* (3) diagnosa pada model SEM konvensional dari berbagai literatur akan dapat diaplikasikan dengan mudah untuk mengecek kesalahan spesifikasi dari model *multilevel;* (4) kesalahan spesifikasi pada salah satu *level* tidak mempengaruhi secara sistematis evaluasi dari model pada *level* lainnya.

Secara singkat, model *Multilevel Confirmatory Factor Analysis* (MCFA) dijelaskan sebagai kasus khusus dari model SEM menurut Muthén dan Asparouhov (2009) yang model ini dijelaskan oleh satu set dari tiga persamaan (notasi mereka):

$$\mathrm{Y_{ik}} = \Lambda_k \eta_{\mathrm{ik}} \ldots (7)$$

$$\eta_{\mathrm{ik}} = \alpha_k + B_k \eta_{\mathrm{ik}} + \zeta_{\mathrm{ik}} \ldots (8)$$

$$\eta_{\mathrm{k}} = \mu + \beta \eta_{\mathrm{ik}} + \zeta_{\mathrm{k}} \ldots (9)$$

di mana *i* dan kasus indeks *k* (unit *level 1*) dan *cluster* (unit *level 2*), masing-masing. Y_{ik} adalah vektor dari *p* variabel yang diukur; $\Lambda_k = \Lambda = [I_p\ 0_{pxm}\ I_p\ 0_{pxm}]$ adalah (p x (2*p* + 2*m*)) *factor loading matrix* yang menghubungkan Y_{ik} ke bagian laten *p* di kedua tingkat *within-* dan *between-cluster* dan *m* faktor umum pada kedua tingkat; η_{ik} adalah vektor panjang (2*p* + 2*m*) mengandung *p* laten bagian *within-cluster*, *m* yaitu faktor umum *within-cluster*, *p* laten bagian *between-cluster*, dan *m* faktor umum *between-cluster*; α_k adalah vektor panjang (2*p* + 2*m*) yang berisi item *intercepts p* dan *m* faktor umum *between-cluster*; B_k adalah (2*p* + 2*m*) x (2*p* + 2*m*) matriks yang mengandung muatan faktor *within-cluster*; η_k (r x 1) terdiri dari koefisien acak subskrip *k* dari α_k dan B_k, termasuk faktor umum *between-cluster*; μ (r x 1) berisi *mean* dari koefisien tersebut dan item *intercepts* (jika diinginkan); β (r x r) berisi muatan faktor *between-cluster*; ζ_{ik} berisi *unique factors* dan faktor umum residu untuk model *within-cluster*; dan ζ_k (r x 1) mengandung *unique factors* dan faktor umum residu untuk model *between-cluster*. Akhirnya, ζ_{ik} ~ MVN (0, ψ_w), dan ζ_k ~ MVN (0, ψ_B).

Geldhof, Preacher & Zyphur (2014) mengemukakan bahwa, meskipun model dasar MCFA dapat dijabarkan dalam berbagai cara, pada penelitian ini akan dibatasi hanya pada model faktor tanpa kovariat, hanya item kontinu, dan tidak ada *latent regression*. Selain itu, dalam kondisi ini hanya dipertimbangkan kasus di mana *item intercepts* dihilangkan, muatan faktor tidak berbeda secara acak di tingkat *cluster* (B_k = B), dan struktur faktor *configural* identik di seluruh tingkatan. Penyederhanaan ini menghasilkan bentuk Persamaan 7 dan 8 yang dibatasi menjadi:

$$\mathrm{Y_{ik}} = \Lambda \eta_{\mathrm{ik}} \ldots (10)$$

$$\eta_{\mathrm{ik}} = \alpha_k + \mathrm{B}\eta_{\mathrm{ik}} + \zeta_{\mathrm{ik}} \ldots (11)$$

Intraclass Correlation Coefficients (ICC)

Terdapat dua jenis dari ICC yang dapat dihitung: *latent factor* ICC dan *observed variable* ICC. (Hsu et al, 2016). *Latent factor* ICC merupakan hal yang lebih umum digunakan pada model *multilevel* CFA (Heck & Thomas, 2015; Muthen, 1994). dimana *latent factor* ICC dikomputasi dengan rumus:

$$Latent\ Factor\ \mathrm{ICC} = \frac{\mathrm{B}}{\mathrm{B} + \mathrm{W}} \ldots (12)$$

Dimana B merupakan proporsi varians *latent factor* pada *between-level* dan W pada *within-level*. Namun perlu dicatat bahwa *latent factor* ICC hanya bisa dikomputasi saat struktur dari model sama pada *between* dan *within-level* (misal, asumsi struktur model identik) sehingga dapat dilakukan manipulasi terhadap besarnya *latent factor* ICC dengan cara menetapkan besaran varians faktor pada *within level* sebagai konstan dan mengubah-ubah besarnya varians faktor pada *between level*. (Hsu et al., 2016). Sedangkan ICC jenis kedua yaitu *observed variable* ICC dapat dikomputasi dengan rumus:

$$Observed\ Variable\ \mathrm{ICC} = \frac{\mathrm{b}}{\mathrm{b} + \mathrm{w}} \ldots (13)$$

Dimana *b* = (muatan faktor *between-level*)^2 × varians faktor *between-level* + varians residual *between-level* dan *w* = (muatan faktor *within-level*)^2 × varians faktor *within-level* + varians residual *within-level*. (Hsu et al., 2016).

Bias Estimasi

Menurut kamus *American Psychological Association* (2015) bias estimasi terjadi apabila nilai yang diperoleh dari data sampel yang secara konsisten *underestimate* atau *overestimate* dari nilai sebenarnya dalam populasi berjumlah besar yang diteliti. Dengan kata lain, sebuah estimator dikatakan bias ketika rata-rata nilainya berbeda dari nilai parameter yang dimaksudkan untuk mewakili nilai yang sebenarnya. Hal ini disebut juga sebagai bias statistik. Marsh, Hau & Greyson (2005) menjelaskan bila nilai yang diharapkan dari statistik yang dilakukan pada sampel bervariasi secara sistematis dengan jumlah N, maka statistik yang dihasilkan akan menghasilkan bias dari parameter populasi sebenarnya.

Hal ini sejalan dengan penelitian sebelumnya (misalnya, Muthén, Kaplan & Hollis, 1987), yang juga mempertimbangkan ukuran mutlak presentase bias < 10% untuk mewakili tingkat bias yang dapat diterima. Karena estimasi parameter pada penelitian ini tidak diantisipasi akan terdistribusi secara normal, estimasi parameter median di setiap kondisi yang lebih baik dijadikan ukuran pemusatan statistik daripada *mean* estimasi parameter. Oleh karena itu bias dalam setiap kondisi dihitung sebagai ([estimasi median - parameter] / parameter) x 100. (Geldhof, Preacher & Zyphur, 2014). Dalam penelitian lainnya, Cai (2010b) menggunakan dua ukuran keakuratan dari estimasi parameter yaitu bias estimasi dan *root mean squared error* (RMSE), dimana pada sebuah parameter θ, bias estimasi didefinisikan sebagai M^{-1} $\sum_{i=1}^{M}(\theta - \widehat{\theta}\, i)$, dimana M merupakan jumlah replikasi *Monte Carlo* keseluruhan dan $\widehat{\theta}_{\,\mathrm{i}}$ adalah MLE dari $\widehat{\theta}$ pada replikasi ke i. Sedangkan RMSE dirumuskan menjadi $\sqrt{M^{-1}\sum_{i=1}^{M}(\theta - \widehat{\theta}\, i)^{2}}$. Tetapi, dalam kondisi dimana terdapat bias sistematis yang besar, RMSE menjadi tidak informatif sebagai ukuran dari efisiensi karena akan didominasi oleh bias. (Preacher, Zhang & Zyphur, 2011).

Metode Penelitian

Penelitian ini menggunakan metode *Markov Chain Monte Carlo* (MCMC) yang berdasar pada pendekatan *Bayesian*. Metode *Monte Carlo* merupakan penelitian dengan teknik simulasi di mana sampel dengan jumlah besar dan sifat tertentu yang ditentukan (misalnya, normalitas, ukuran, jenis model) dihasilkan oleh komputer untuk menilai penggunaan prosedur statistik atau parameter dalam berbagai kondisi. Sebagai contoh, seorang peneliti mungkin melakukan penelitian dengan metode *Monte Carlo* dengan sampel berjumlah besar yang terdistribusi normal dari berbagai ukuran sampel (misalnya, N = 50, 100, 200, 400, 800) di mana model struktural diterapkan untuk menggambarkan data. Hasil penelitian akan membantu peneliti menentukan kondisi di mana model berfungsi dengan benar (yaitu, sesuai dengan data) serta menunjukkan batas-batasnya (misalnya, tidak cocok dengan baik dengan sampel ukuran kurang dari 200). (*American Psychological Association*, 2015). Selanjutnya akan dijelaskan mengenai model populasi yang digunakan sebagai acuan proses pembangkitan data dan juga desain simulasi yang ditetapkan. Hal ini sebenarnya merupakan konsep-konsep terkait dasar-dasar pendekatan Bayesian seperti distribusi *prior, posterior* dan kaitannya dengan fungsi *likelihood*, namun penjelasannya tidak dijabarkan dengan rinci untuk memudahkan pelaporan. Penjelasan lengkap dapat dilihat pada berbagai literatur yang tersedia (misal, van De Schoot et al., 2014).

Model populasi yang digunakan pada penelitian ini merupakan model yang diambil dari penelitian Muthén (1994) yang berisi delapan indikator yang masing-masing diwakili satu faktor yaitu η_B pada *between level* dan η_W pada *within level.* Pada penelitian ini model MCFA pada bagian *between-level* memiliki struktur faktor yang muatannya berbeda dengan model *within-level.* Hal ini dilakukan karena mengikuti informasi yang didapat dari penelitian sebelumnya berupa data dan memilih untuk menggunakannya dibanding menciptakan kondisi baru dimana dilakukan variasi pada muatan faktor dan varians *residual* untuk menjadi model populasi sebagai proses pembangkitan data. Namun, berdasarkan model di atas, dapat disimpulkan bahwa model ini memiliki struktur faktor yang identik baik pada *between-level* maupun *within-level,* sehingga *latent factor* ICC dapat digunakan dalam penelitian ini karena struktur faktor yang identik tersebut. Berdasakan model ini, spesifikasi parameter bebas yang akan diestimasi dibagi menjadi dua bagian yaitu *between-level* dan *within level.* Spesifikasi parameter bebas pada *within-level* sebanyak 8 *lambda*, 8 *theta*. Sedangkan spesifikasi parameter bebas pada *between-level* yaitu: 8 *nu*, 8 *lambda* dan 8 *theta*. Sehingga keseluruhan berjumlah 40 parameter bebas yang akan diestimasi Adapun parameter bebas pada *between-level* dan *within-level* untuk melakukan simulasi data ditetapkan terlebih dahulu dimana hal utama yang ditetapkan adalah muatan faktor maupun varians *residual* yang digunakan sebagai *starting values* bagi estimasi parameter bebas yang dipaparkan pada Tabel 1.

Tabel 1. Muatan Faktor Model Populasi

Item	Muatan Faktor		Varians *Residual*	
	Between	*Within*	*Between*	*Within*
Y_1	0,97	0,52	0,05	0,72
Y_2	0,98	0,49	0,03	0,75
Y_3	0,92	0,32	0,15	0,89
Y_4	0,88	0,25	0,22	0,93
Y_5	0,89	0,34	0,20	0,88
Y_6	0,84	0,23	0,29	0,94
Y_7	0,80	0,26	0,36	0,93
Y_8	0,77	0,31	0,40	0,90

Berdasarkan informasi pada Tabel 1, untuk menciptakan kondisi ICC yang berbeda-beda, varians pada faktor *between-level* ditetapkan nilainya menjadi antara 0.06 sampai 0.26 sehingga menghasilkan kondisi ICC yang diinginkan yaitu 0.05, 0.10, 0.15, 0.20 dan 0.25 sebagai model awal yang digunakan untuk proses pembangkitan data untuk studi dengan simulasi data. Model ini merupakan model yang menjadi distribusi yang akan diuji pada berbagai kondisi yang diciptakan, adapun kondisi yang dimaksud, akan dijelaskan pada bagian selanjutnya.

Karena penelitian ini menggunakan metode simulasi data, maka harus ditetapkan terlebih dahulu kondisi simulasi data untuk melihat hasil dari simulasi berdasarkan kondisi yang berbeda-beda, kondisi yang ditetapkan harus ditetapkan berdasarkan cara yang tersedia maupun dengan kriteria tertentu untuk dapat diaplikasikan, meskipun hal ini sebenarnya menjadi kritik dari aliran klasik. Adapun kondisi yang ditetapkan sebagai desain simulasi pada penelitian ini yaitu:

1. Jumlah *Cluster*
   30, 50, 100 dan 150
2. Jumlah Observasi Pada *Cluster*
   10, 20 dan 50
3. *Intraclass Correlation Coeffcients* (ICC)
   0.05, 0.10, 0.15, 0.20 dan 0.25
4. Metode Estimasi
   Estimator yang digunakan pada penelitian ini adalah ML, MLF, MLR, WLSMV dan BAYES.

Dengan kondisi simulasi yang ditetapkan 4 × 3 × 5 × 5 akan menghasilkan kondisi sebanyak 300 yang masing-masing akan direplikasi sebanyak 1000 kali. Penentuan kondisi ini juga diikuti dengan observasi maupun komparasi dengan keadaan pada kondisi nyata dimana kondisi-kondisi tersebut umum digunakan pada aplikasinya dengan data yang didapat dari lapangan. Hal ini dilakukan agar hasil yang didapat dari penelitian ini yang dilakukan menggunakan studi dengan simulasi dapat diperbandingkan ataupun dikaitkan dengan pengaplikasiannya yang telah umum dilakukan.

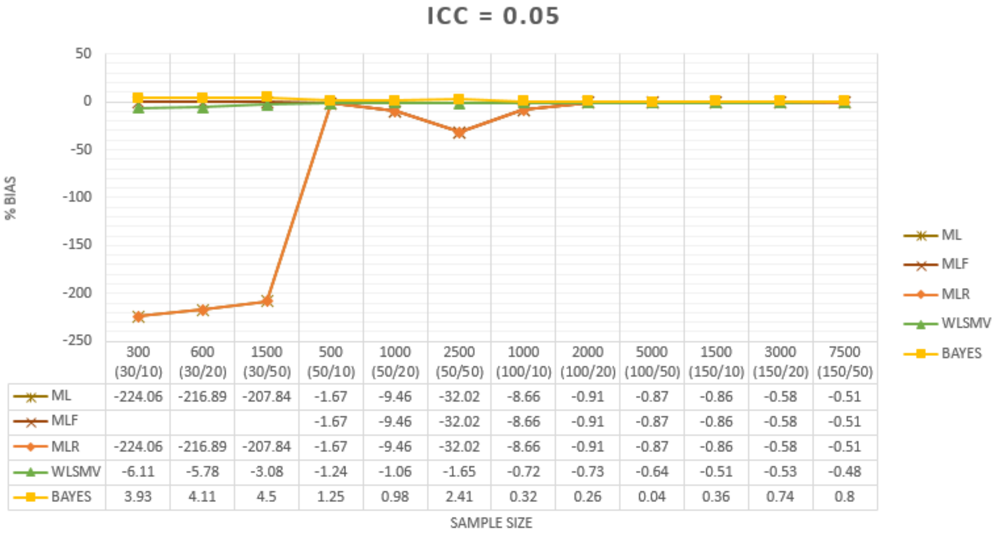
Hasil Penelitian

Akan dipaparkan pengelompokan hasil estimasi yang menunjukkan besaran bias pada masing-masing kondisi ICC yaitu 0.05, 0.10, 0.15, 0.20 dan 0.25 yang membandingkan lima metode estimasi yaitu ML (*maximum likelihood*), MLF (*full information maximum likelihood*), MLR (*maximum likelihood robust*), WLSMV (*robust weighted least square*) dan BAYES (*Bayesian estimation*). Adapun pengelompokannya dijelaskan masing-masing pada grafik yang akan ditampilkan.

Pengelompokan hasil simulasi data berdasarkan kondisi dengan ICC sebesar 0.05 pada kondisi ukuran sampel maupun kelima metode estimasi yang digunakan yaitu ML, MLF, MLR, WLSMV dan BAYES digambarkan dalam grafik yang masing-masing berisikan 60 kondisi yang memuat informasi mengenai besaran bias yang terjadi pada kondisi ICC ini, adapun grafik yang dimaksud berada pada Gambar 1.

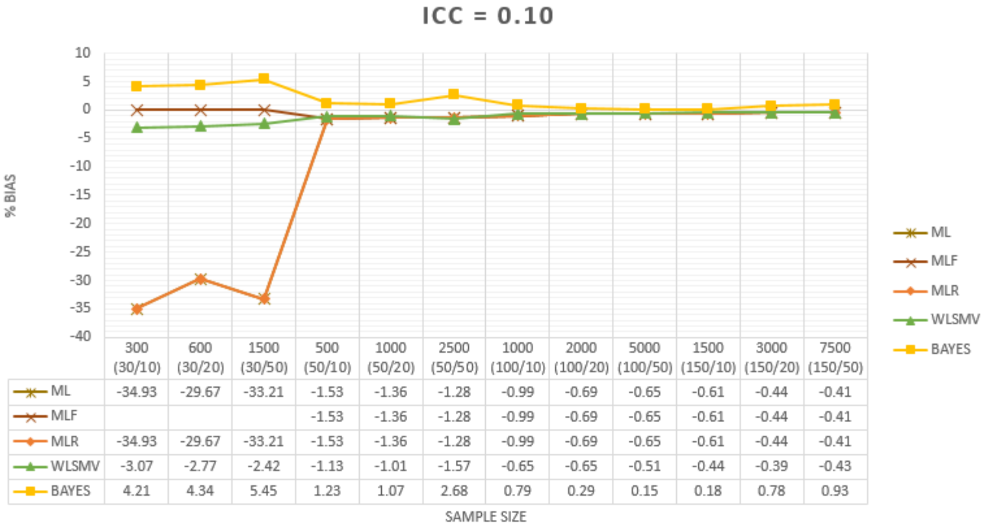
Berdasarkan grafik pada Gambar 1, dengan kondisi ICC sebesar 0.05 terlihat bahwa metode estimasi ML dan MLR memberikan hasil estimasi yang bias pada kondisi sampel 30/10, 30/20, 30/50 dan 50/50.

Sedangkan pada kondisi sampel lainnya tidak terjadi bias yang melebihi batas yang telah ditetapkan. Lalu, dengan metode estimasi BAYES dan WLSMV hasil estimasi parameter yang diberikan berbeda dengan ML & MLR pada kondisi sampel yang sama dimana metode estimasi BAYES dan WLSMV memberikan hasil estimasi yang lebih baik dan tidak bias pada kondisi ukuran sampel manapun. Sedangkan metode estimasi MLF tidak berfungsi pada kondisi sampel 30/10, 30/20 dan 30/50 dimana meskipun metode estimasi ini sama-sama berbasis *maximum likelihood* seperti ML dan MLR tetapi proses kalkulasi yang berbeda menganggap bahwa kondisi ukuran sampel tersebut tidak memenuhi syarat untuk melakukan estimasi parameter. Namun pada kondisi sampel yang lain, metode estimasi MLF dapat digunakan dan memberikan hasil yang sama dengan ML ataupun MLR.



	300 (30/10)	600 (30/20)	1500 (30/50)	500 (50/10)	1000 (50/20)	2500 (50/50)	1000 (100/10)	2000 (100/20)	5000 (100/50)	1500 (150/10)	3000 (150/20)	7500 (150/50)
ML	-224.06	-216.89	-207.84	-1.67	-9.46	-32.02	-8.66	-0.91	-0.87	-0.86	-0.58	-0.51
MLF				-1.67	-9.46	-32.02	-8.66	-0.91	-0.87	-0.86	-0.58	-0.51
MLR	-224.06	-216.89	-207.84	-1.67	-9.46	-32.02	-8.66	-0.91	-0.87	-0.86	-0.58	-0.51
WLSMV	-6.11	-5.78	-3.08	-1.24	-1.06	-1.65	-0.72	-0.73	-0.64	-0.51	-0.53	-0.48
BAYES	3.93	4.11	4.5	1.25	0.98	2.41	0.32	0.26	0.04	0.36	0.74	0.8

Gambar 1. Bias pada Kondisi dengan ICC 0.05



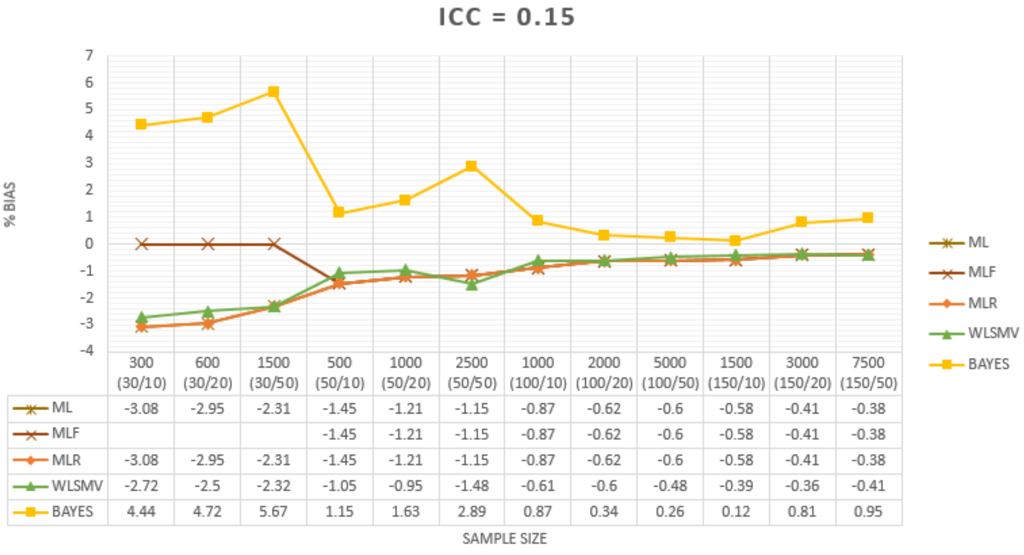
	300 (30/10)	600 (30/20)	1500 (30/50)	500 (50/10)	1000 (50/20)	2500 (50/50)	1000 (100/10)	2000 (100/20)	5000 (100/50)	1500 (150/10)	3000 (150/20)	7500 (150/50)
ML	-34.93	-29.67	-33.21	-1.53	-1.36	-1.28	-0.99	-0.69	-0.65	-0.61	-0.44	-0.41
MLF				-1.53	-1.36	-1.28	-0.99	-0.69	-0.65	-0.61	-0.44	-0.41
MLR	-34.93	-29.67	-33.21	-1.53	-1.36	-1.28	-0.99	-0.69	-0.65	-0.61	-0.44	-0.41
WLSMV	-3.07	-2.77	-2.42	-1.13	-1.01	-1.57	-0.65	-0.65	-0.51	-0.44	-0.39	-0.43
BAYES	4.21	4.34	5.45	1.23	1.07	2.68	0.79	0.29	0.15	0.18	0.78	0.93

Gambar 2. Bias pada Kondisi dengan ICC 0.10

Berdasarkan grafik di Gambar 2, dengan kondisi ICC sebesar 0.10 terlihat bahwa metode estimasi ML dan MLR memberikan hasil estimasi yang bias pada kondisi sampel 30/10, 30/20 dan 30/50. Sedangkan pada kondisi sampel lainnya tidak terjadi bias yang melebihi batas yang telah ditetapkan. Lalu, dengan metode estimasi BAYES dan WLSMV hasil estimasi yang diberikan berbeda dengan ML & MLR pada kondisi sampel yang sama dimana metode estimasi BAYES dan WLSMV memberikan hasil estimasi yang lebih baik dan tidak bias melebihi kriteria yang ditetapkan pada kondisi ukuran sampel manapun. Sedangkan metode estimasi MLF tidak berfungsi pada kondisi sampel 30/10, 30/20 dan 30/50 dimana meskipun metode estimasi ini sama-sama berbasis *maximum likelihood* seperti ML dan MLR tetapi proses kalkulasi yang berbeda menganggap bahwa kondisi dengan ukuran sampel tersebut tidak memenuhi syarat untuk melakukan estimasi parameter. Namun pada kondisi sampel yang lain, metode estimasi MLF dapat digunakan dan memberikan hasil yang sama dengan ML ataupun MLR.

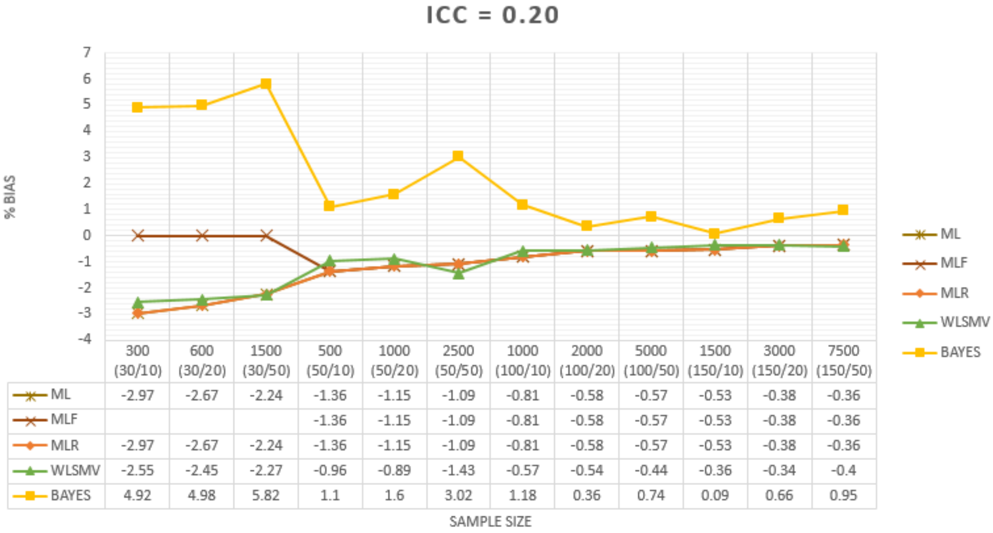
Berdasarkan grafik pada Gambar 3, secara keseluruhan pada kondisi ICC 0.15 pada seluruh kondisi ukuran sampel tidak terjadi bias yang melebihi batas yang ditetapkan yaitu > 10%. Berdasarkan arah dari bias meskipun sangat kecil, pada metode estimasi ML, MLR, WLSMV bias yang terjadi arahnya negatif dan pada metode estimasi BAYES arahnya positif dan lebih besar dibandingkan metode yang berbasis *maximum likelihood*. Tetapi terdapat pengecualian dimana MLF tidak dapat berfungsi dengan baik pada kondisi sampel yang tidak memenuhi syarat. Terdapat kesamaan pola dimana semakin besar jumlah sampel maka semakin rendah bias yang terjadi seperti dapat dilihat pada gambar 3.

Berdasarkan grafik di Gambar 4, secara keseluruhan pada kondisi ICC 0.20 pada seluruh kondisi ukuran sampel tidak terjadi bias yang melebihi batas yang ditetapkan yaitu > 10%. Berdasarkan arah dari bias meskipun sangat kecil, pada metode estimasi ML, MLR, WLSMV bias yang terjadi arahnya negatif dan pada metode estimasi BAYES arahnya positif dan lebih besar dibandingkan metode yang berbasis *maximum likelihood*. Tetapi terdapat pengecualian dimana MLF tidak dapat berfungsi dengan baik pada kondisi sampel yang tidak memenuhi syarat. Terdapat kesamaan pola dimana semakin besar jumlah sampel maka semakin rendah bias yang terjadi seperti dapat dilihat pada grafik sebelumnya.



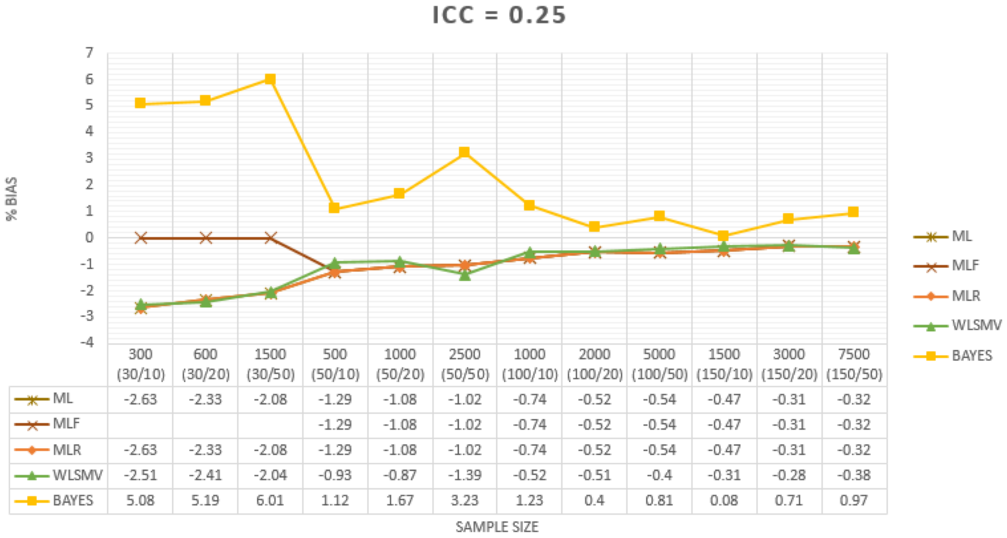
	300 (30/10)	600 (30/20)	1500 (30/50)	500 (50/10)	1000 (50/20)	2500 (50/50)	1000 (100/10)	2000 (100/20)	5000 (100/50)	1500 (150/10)	3000 (150/20)	7500 (150/50)
ML	-3.08	-2.95	-2.31	-1.45	-1.21	-1.15	-0.87	-0.62	-0.6	-0.58	-0.41	-0.38
MLF				-1.45	-1.21	-1.15	-0.87	-0.62	-0.6	-0.58	-0.41	-0.38
MLR	-3.08	-2.95	-2.31	-1.45	-1.21	-1.15	-0.87	-0.62	-0.6	-0.58	-0.41	-0.38
WLSMV	-2.72	-2.5	-2.32	-1.05	-0.95	-1.48	-0.61	-0.6	-0.48	-0.39	-0.36	-0.41
BAYES	4.44	4.72	5.67	1.15	1.63	2.89	0.87	0.34	0.26	0.12	0.81	0.95

Gambar 3. Bias pada Kondisi Dengan ICC 0.15



	300 (30/10)	600 (30/20)	1500 (30/50)	500 (50/10)	1000 (50/20)	2500 (50/50)	1000 (100/10)	2000 (100/20)	5000 (100/50)	1500 (150/10)	3000 (150/20)	7500 (150/50)
ML	-2.97	-2.67	-2.24	-1.36	-1.15	-1.09	-0.81	-0.58	-0.57	-0.53	-0.38	-0.36
MLF				-1.36	-1.15	-1.09	-0.81	-0.58	-0.57	-0.53	-0.38	-0.36
MLR	-2.97	-2.67	-2.24	-1.36	-1.15	-1.09	-0.81	-0.58	-0.57	-0.53	-0.38	-0.36
WLSMV	-2.55	-2.45	-2.27	-0.96	-0.89	-1.43	-0.57	-0.54	-0.44	-0.36	-0.34	-0.4
BAYES	4.92	4.98	5.82	1.1	1.6	3.02	1.18	0.36	0.74	0.09	0.66	0.95

Gambar 4. Bias pada Kondisi dengan ICC 0.20



	300 (30/10)	600 (30/20)	1500 (30/50)	500 (50/10)	1000 (50/20)	2500 (50/50)	1000 (100/10)	2000 (100/20)	5000 (100/50)	1500 (150/10)	3000 (150/20)	7500 (150/50)
ML	-2.63	-2.33	-2.08	-1.29	-1.08	-1.02	-0.74	-0.52	-0.54	-0.47	-0.31	-0.32
MLF				-1.29	-1.08	-1.02	-0.74	-0.52	-0.54	-0.47	-0.31	-0.32
MLR	-2.63	-2.33	-2.08	-1.29	-1.08	-1.02	-0.74	-0.52	-0.54	-0.47	-0.31	-0.32
WLSMV	-2.51	-2.41	-2.04	-0.93	-0.87	-1.39	-0.52	-0.51	-0.4	-0.31	-0.28	-0.38
BAYES	5.08	5.19	6.01	1.12	1.67	3.23	1.23	0.4	0.81	0.08	0.71	0.97

Gambar 5. Bias pada Kondisi dengan ICC 0.25

Berdasarkan grafik di Gambar 5, secara keseluruhan pada kondisi ICC 0.25 pada seluruh kondisi ukuran sampel tidak terjadi bias yang melebihi batas yang ditetapkan yaitu > 10%. Berdasarkan arah dari bias meskipun sangat kecil, pada metode estimasi ML, MLR, WLSMV bias yang terjadi arahnya negatif dan pada metode estimasi BAYES arahnya positif dan lebih besar dibandingkan metode yang berbasis *maximum likelihood*. Terdapat kesamaan pola dimana semakin besar ukuran sampel maka semakin rendah bias yang terjadi seperti dapat dilihat pada grafik sebelumnya.

Hal ini menunjukkan bahwa ke lima metode estimasi yang digunakan merupakan metode estimasi yang tidak bias, konsisten dan efisien untuk mengestimasi parameter data yang berstruktur *multilevel* apabila ukuran sampel dari data memenuhi kriteria yang telah ditetapkan oleh penelitian-penelitian terdahulu terkait ukuran sampel yang dibutuhkan meskipun terdapat pengecualian dimana pada kondisi sampel tertentu MLF tidak dapat berfungsi dengan baik. Berlawanan dengan metode estimasi MLR, metode estimasi BAYES merupakan metode yang memiliki kelebihan untuk melakukan estimasi pada kondisi data yang tidak ideal yaitu berupa ukuran sampel kecil ataupun ICC yang rendah. Metode estimasi BAYES juga membutuhkan waktu komputasi yang lebih panjang dibanding metode estimasi ML, MLF, MLR & WLSMV. Kelima metode ini merupakan metode yang memiliki dasar yang berbeda satu sama lain sehingga memiliki kelebihan dan kekurangan masing-masing, tetapi secara keseluruhan kelima metode ini merupakan metode estimasi yang baik untuk digunakan dalam mengestimasi parameter data berstruktur *multilevel.* Selanjutnya, untuk melakukan perbandingan hasil yang didapat dari penelitian ini dengan penelitian sebelumnya yang relevan, akan dijelaskan lebih lanjut pada bagian selanjutnya pada subbab diskusi.

Pembahasan

Berdasarkan hasil analisis data yang dilakukan, penelitian ini menunjukkan bahwa terdapat berbagai temuan yang menarik pada analisis data yang berstruktur hirarki. Analisis data *multilevel* adalah topik yang kompleks karena mengacu pada kontribusi dari berbagai bidang metodologi penelitian. Dua perspektif dapat membedakan, yaitu pengambilan sampel dan parameter yang berbeda-beda. Dari perspektif *sampling*, data yang bertingkat dapat dilihat seperti yang diperoleh dengan *cluster sampling*. Analisis perlu untuk menentukan variasi stokastik yang mencerminkan skema *sampling*, seperti membuat model yang terurai dari variasi kelompok (*cluster*) dan komponen individu. (Muthén, 1994).

Hasil pengambilan sampel menggunakan metode *multistage sampling* pada data hierarki yang terstruktur (misalnya, siswa berkumpul di dalam satu unit kelas), membuat residual bergantung pada variasi *between-cluster*. Skor pada variabel yang diamati dari anak-anak dalam satu kelas mungkin lebih mirip daripada anak-anak di kelas yang berbeda, misalnya. Apabila mengabaikan struktur data hirarkis dapat terjadi bias pada estimasi hubungan antar item. (Geldhof, Preacher & Zyphur, 2014), ICC bersamaan dengan ukuran sampel dan juga informasi jumlah individu dalam *cluster* merupakan informasi yang penting dalam permodelan *multilevel* (Stapleton, Yang & Hancock, 2016). Penelitian tersebut menunjukkan bahwa ICC merupakan faktor penting yang perlu diperhatikan dalam menganalisis data *multilevel*, dimana hasil dari penelitian ini menunjukkan bahwa kondisi ICC berpengaruh terhadap bias estimasi parameter data *multilevel.*

Terdapat temuan yang menarik pada kondisi ICC 0.05 dimana apabila dilakukan perbandingan pada dua kondisi sampel yaitu 50/20 dan 100/10 yang sama-sama sebesar 1000 observasi terjadi bias pada kondisi 50/20 sebesar -9.46 tetapi biasnya menurun pada kondisi 100/10 yaitu sebesar -8.66, hal ini menunjukkan bahwa struktur dari sampel mempengaruhi bias dimana jumlah *cluster* yang lebih besar membuat estimasi parameter data *multilevel* menjadi lebih baik dibanding *cluster* yang lebih sedikit. Hal ini sejalan dengan pendapat Cronbach (1976, dalam Muthén, 1991) yaitu jika terbentur permasalahan, mungkin lebih baik mengamati siswa per kelas yang lebih sedikit untuk membantu menambah kelas yang lebih banyak dimana kelas sebagai *cluster* yang jumlahnya lebih banyak akan memberikan hasil estimasi yang terbukti lebih baik dan tidak bias. Besaran bias yang terjadi pada kondisi dengan sampel kecil seperti 30/10, 30/20 & 30/50 pada metode estimasi MLR menunjukkan bahwa MLR tidak berfungsi dengan baik apabila ICC kecil dan juga jumlah individu dalam cluster lebih kecil dari 50 sesuai dengan hasil penelitian terdahulu. (Hox & Mass, 2004, dalam Stapleton, Yang & Hancock, 2016).

Hasil penelitian ini juga sejalan dengan penelitian Pornprasertmanit, Lee & Preacher (2014) yang menyatakan bahwa bias akan semakin rendah saat jumlah *cluster* semakin besar apabila menggunakan metode estimasi MLR. Temuan ini juga sejalan dengan pendapat Heck & Thomas (2015) yang menyatakan bahwa ukuran sampel pada tingkat grup (jumlah grup) umumnya lebih penting dari ukuran sampel tingkat individu (jumlah observasi dalam *cluster*) yang sejalan dengan Łaszkiewicz (2013) dan

Cronbach (1976, dalam Muthen, 1991) yang berpendapat bahwa hal yang penting bukan jumlah yang besar pada pengamatan per unit namun jumlah besar kelompok tampaknya lebih penting untuk mendapatkan estimasi yang akurat.

Temuan ini juga membuktikan kesimpulan yang diambil dari penelitian sebelumnya oleh Preacher, Zhang & Zyphur (2011) yaitu semakin meningkatnya jumlah grup, ukuran *cluster* dan ICC yang berada dalam batas menunjukkan hasil yang baik pada penggunaan metode MSEM terkhusus pada bias yang terjadi dengan metode estimasi MLR. Hal ini sejalan dengan penelitian sebelumnya yang menyatakan bahwa jika jumlah *cluster* besar dan ICC bernilai setidaknya 0.10, MSEM merupakan metode yang paling efisien seiring bertambahnya jumlah *cluster* dan metode ini memberikan estimasi yang lebih efisien dibandingkan penggunaan nilai rata-rata dari kelompok apabila ukuran kelompok yang besar digunakan (Lüdtke, Marsh, Robitzsch, Trautwein, Asparouhov, & Muthén, 2008).

Pada kondisi dengan ukuran sampel kecil sebesar 30/10, 30/20, 30/50 dan 50/50 metode estimasi MLR memberikan hasil estimasi yang bias, hal ini sejalan dengan penelitian terdahulu yang menyatakan bahwa dengan ukuran sampel yang kecil, metode ML menghasilkan estimasi yang bias (Morris, 1995 dalam Heck & Thomas, 2015). Lebih lanjut, untuk mengatasi permasalahan ukuran sampel yang kecil, metode estimasi *Bayesian* merupakan pendekatan alternatif yang dapat digunakan (Heck & Thomas, 2015) yang pada penelitian ini terbukti bahwa dalam kondisi ukuran sampel kecil, metode estimasi *Bayesian* memberikan hasil estimasi yang lebih baik dari MLR dengan ukuran bias yang masih dapat diterima. Temuan ini juga membuktikan hasil pene-litian sebelumnya yang menyatakan bahwa pada kondisi data yang tidak ideal, metode estimasi BAYES dapat bekerja dengan baik (Lindley & Smith, 1972; Smith, 1973; Morris, 1995; dalam Heck & Thomas, 2015). Metode estimasi BAYES juga memberikan keuntungan dimana model *multilevel* SEM sulit diestimasi dengan metode estimasi berbasis ML karena membutuhkan jumlah integrasi yang banyak, sehingga metode estimasi BAYES sangat baik untuk digunakan (Asparouhov & Muthen, 2014).

Pada kondisi ukuran sampel yang lain selain ukuran sampel 30/10, 30/20, 30/50 dan 50/50 metode estimasi ML, MLF, MLR terbukti konsisten, *unbiased* dan efisien dimana hasil penelitian ini menunjukkan bahwa semakin besar ukuran *cluster* dalam kondisi ICC lebih besar dari 0.05, maka semakin kecil bias yang terjadi pada estimasi parameter data *multilevel*. Sedangkan metode estimasi BAYES memberikan hasil yang berbeda, dimana metode ini membutuhkan waktu komputasi yang lebih lama dibandingkan penggunaan metode estimasi MLR.

Penelitian ini juga menjawab permasalahan yang terkait dengan penggunaan metode estimasi MLF dimana seharusnya metode berbasis *maximum likelihood* seperti ML, MLF dan MLR menghasilkan hasil yang sama, tetapi pada penelitian ini terdapat temuan bahwa pada beberapa kondisi ukuran sampel dan ICC tertentu MLF tidak dapat berfungsi. Hal ini terjadi karena menurut penelitian terdahulu, bagaimanapun untuk ukuran sampel yang kecil/menengah MLF akan mengalami *overestimate* pada *standard error*-nya. Hal ini biasanya terjadi pada kasus dimana perbandingan antara jumlah unit independen (observasi pada model satu *level* atau *cluster* pada model *multilevel*) dan jumlah parameter kurang dari 10. (Asparouhov & Muthen, 2012). Dalam penelitian ini terdapat gambaran jumlah ukuran sampel yang pasti dimana MLF tidak dapat berfungsi yaitu ukuran sampel 30/10, 30/20 dan 30/50. Sehingga pada penelitian mendatang untuk melakukan analisis MLLVM tidak disarankan untuk menggunakan metode estimasi MLF pada ukuran sampel tersebut.

Terkait penggunaan metode estimasi WLSMV pada penelitian ini, hasil yang didapat menunjukkan jika performa dari WLSMV lebih baik dari metode estimasi berbasis *maximum likelihood* bila tidak terjadi bias yang melebihi kriteria yang ditetapkan pada seluruh kondisi ukuran sampel dan ICC yang tersedia. Hal ini sejalan dengan penelitian yang menunjukkan bahwa WLSMV menghasilkan estimasi yang lebih baik dibandingkan estimator berbasis ML pada seluruh kondisi sampel (Beauducel & Herzberg, 2006).

Sejalan dengan pendapat yang dikemukan Meuleman & Billiet (2009) studi ini menggambarkan penggunaan studi dengan simulasi *Monte Carlo* yang berperan baik dalam menggambarkan akurasi dari estimasi yang dilakukan. Tetapi pendekatan *Bayesian* yang banyak menawarkan keuntungan dalam penggunaannya pada ukuran data yang kecil perlu diperhatikan bahwa efisiensi dari estimasi bergantung pada informasi yang tersedia sebagai *prior* dari parameter, tanpa informasi ini, estimasi yang dilakukan tidak akan memberikan hasil yang lebih baik dibandingkan metode estimasi lain dan bahkan dapat menghasilkan hasil yang lebih buruk (Heck & Thomas, 2015).

Terdapat berbagai keterbatasan dari penggunaan metode MSEM seperti permasalahan tentang konvergensi dan waktu komputasi, tetapi metode estimasi tingkat *advance* dan kekuatan alat untuk melakukan komputasi akan menyelesaikan permasalahan ini karena kompleksitas dari model ataupun desain dari penelitian yang berkembang akan terus memaksa untuk melakukan pengembangan pada prkateknya dimana para peneliti akan bereksperimen mengenai model-model yang akan di estimasi tetapi tetap berpegang pada bagaimana penelitian tersebut mungkin dilakukan dalam prakteknya. (Preacher, Zhang & Zyphur, 2016). Studi dengan simulasi data menggunakan metode MCMC menjawab persoalan untuk bagaimana efisien dan komputasi berdasar pada algoritma dengan bantuan komputer. (Cai, 2010b). Tetapi, tidak dapat dipungkiri juga apabila waktu yang dibutuhkan dalam proses komputasi *Bayesian* meningkat karena teknik dengan *iterative sampling* digunakan. (Van de Schoot, Kaplan, Denissen, Asendorpf, Neyer & Aken, 2014).

Hasil penelitian ini sejalan dengan penelitian Muthen & Asparouhov (2012) yang memperkenalkan aplikasi pendekatan*Bayesian* dalam penggunaannya pada SEM yang bernama BSEM yang merupakan metode yang mudah dan cepat untuk menganalisis keterkaitan antara *loading* dimana pengaplikasian metode estimasi *Bayes* pada penelitian ini menghasilkan estimasi yang baik pada kondisi ICC dan juga ukuran sampel manapun. Namun, terdapat kelemahan yang sama seperti penelitian tersebut dimana analisis yang melibatkan kovarians *residual* menyebabkan komputasi yang sangat sulit karena konvergensi dari MCMC akan menjadi lambat.

Penelitian ini memberikan contoh bagaimana hipotesis yang dibuat dari distribusi *prior* informatif yang diuji dengan pendekatan *Bayesian* sejalan dengan penelitian sebelumnya yang menjelaskan bahwa hipotesis pada pendekatan klasik dan hipotesis informatif merupakan hal yang berbeda. bukan hanya dari bagaimana hipotesis tersebut dirumuskan namun juga bagaimana hipotesis tersebut dievaluasi. Hipotesis nol pada pendekatan klasik dapat dievaluasi menggunakan *p-value* sedangkan hipotesis informatif dapat dievaluasi menggunakan seleksi maudel *Bayesian*, hasil utama dari seleksi model *Bayesian* adalah distribusi *posterior* (Hoijtink, Klugkist & Boelen, 2008). Tetapi pendekatan ini dikritik karena mengasumsikan bahwa setiap parameter memiliki distribusi pada populasi, bahkan termasuk kovarians dimana aliran *frequentist* tidak setuju dengan asumsi ini karena mereka berasumsi bahwa pada populasi hanya ada satu nilai tetap dari parameter (Van de Schoot, Kaplan, Denissen, Asendorpf, Neyer & Aken, 2014).

Simpulan

Berdasarkan hasil penelitian dan pembahasan, dapat disampaikan simpulan sebagai berikut. Hasil penelitian ini menunjukkan bahwa bias yang terjadi dipengaruhi oleh ukuran sampel dan ICC, penelitian ini juga menujukkan bahwa metode estimasi WLSMV dan BAYES berfungsi lebih baik pada berbagai kondisi dibandingkan dengan metode estimasi berbasis ML.

Daftar Pustaka

American Psychological Association. (2015). *APA dictionary of psychology* (2nd ed.). Washington, DC: American Psychological Association

Asparouhov, T., & Muthén, B. O. (2003). Full-information maximum-likelihood estimation of general two-level latent variable models with missing data. *Mplus Working Paper*. Los Angeles, CA: Muthén & Muthén, Inc

Asparouhov, T., & Muthén, B. O. (2012). Saddle points. *Mplus Working Paper*. Los Angeles, CA: Muthén & Muthén, Inc

Asparouhov, T., & Muthén, B. O. (2014). General random effect latent variable modeling: Random subjects, items, contexts, and parameters. *Mplus Working Paper*. Los Angeles, CA: Muthén & Muthén, Inc

Beauducel, A. & Herzberg, P. Y. (2006) On the Performance of Maximum Likelihood Versus Means and Variance Adjusted Weighted Least Squares Estimation in CFA, *Structural Equation Modeling: A Multidisciplinary Journal*, 13(2), 186-203

Brown, T. A. (2006). *Confirmatory factor analysis for applied research*. New York, NY: Guilford Press.

Burstein, L. (1980). The analysis of multilevel data in educational research and evaluation. *Review of research in education,* 8(1), 158-233

Cai, L. (2010b). Metropolis-hastings robbins-monro algorithm for confirmatory item factor analysis. *Journal of Educational and Behavioral Statistics*, 35(3), 307-335

de Leeuw, J. & Meijer, E. (2008), *Handbook of multilevel analysis*. New York, NY: Springer.

Dyer, N. G., Hanges, P. J., Hall, R. J. (2005). Applying multilevel confirmatory factor analysis techniques to the study of leadership. *The Leadership Quarterly*. 16, 149–167

Feinberg, R. A. & Rubright, J. D. (2016). Conducting simulation studies in psychometrics. *Educational Measurement: Issues and Practice*, 35(2), 36-49

Geldhof, G. J., Preacher, K. J., Zyphur, M. J. (2014). Reliability estimation in a multilevel confirmatory factor analysis framework. *Psychological Methods*, 19(1), 72–91

Goldstein, H. (2011). *Multilevel Statistical Models* (4th ed.). London: Wiley.

Hayes, A. F. (2006). A primer on multilevel modeling. *Human communication research*, 32, 385–410

Heck, R. H., & Thomas, S. L. (2015). *An introduction to multilevel modeling techniques* (3rd ed.). New York, NY: Routledge.

Hoijtink, H., Klugkist, I., & Boelen, P. A. (2008). *Bayesian evaluation of informative hypotheses*. New York, NY: Springer

Hsu, H., Lin, J., Kwok, O., Acosta, S., & Willson, V. (2016). The impact of intraclass correlation on the effectiveness of level-specific fit indices in multilevel structural equation modeling: a monte carlo study. *Eduactional and Psychological Measurement*, 1-27

Julian, M. (2001). The consequences of ignoring multilevel data structures in nonhierarchical covariance modeling. *Structural Equation Modeling*, 8, 325–352.

Kaplan, D., Kim, J. S., and Kim, S. Y. (2009). Multilevel latent variable modeling: current research and recent developments. in r. e. millsap & a. maydeu-olivares (Eds.), *The SAGE Handbook of Quantitative Methods in Psychology* (pp. 592-613). Thousand Oaks, CA: Sage

Kreft, I., & de Leeuw, J. (1998). *Introducing multilevel modeling*. Thousand Oaks, CA: Sage.

Łaszkiewicz, E. (2013). Sample size and structure for multilevel modelling: monte carlo investigation for the balanced design. *Quantitative methods in* *economics*, 14(2), 19-28

Lüdtke, O., Marsh, H. W., Robitzsch, A., Trautwein, U., Asparouhov, T., & Muthén, B. (2008). The multilevel latent covariate model: A new, more reliable approach to group-level effects in contextual studies. *Psychological Methods*, 13, 203-229

Marsh, H. W., Hau, K-T & Grayson, D. (2005). Goodness of Fit Evaluation in Structural Equation Modeling. In A. Maydeu-Olivares & J. McArdle (Eds.), *Contemporary Psychometrics. A Festschrift for Roderick P. McDonald* (pp. 275-340). Mahwah NJ: Erlbaum.

Mass, C. J. M. & Hox, J. J. (2005). Sufficient sample sizes for multilevel modeling. *Methodology*, 1(3), 86-92

Meuleman, B., & Billiet, J. (2009). A Monte Carlo sample size study: How many countries are needed for accurate multilevel SEM?. *Survey Research Methods, 3,* 45–58.

Muthèn, B. O. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators, *Psychometrika,* 49, 115–132.

Muthèn, B. O. (1989). Latent variable modeling in heterogenous populations, *Psychometrika,* 54, 557–585.

Muthén, B. O. (1991). Multilevel factor analysis of class and student achievement components. *Journal of Educational Measurement*, 28(4), 338-354

Muthén, B. O. (1994). Multilevel covariance structure analysis, *Sociological Methods & Research,* 22, 376

Muthén, B. O. (1997). Latent variable modeling of longitudinal and multilevel data, *Sociological Methodology,* 27, 453-480

Muthén, B. O., & Asparouhov, T. (2009). Growth mixture modeling: analysis with non-Gaussian random effects. In G. Fitzmaurice, M. Davidian, G. Verbeke, & G. Molenberghs (Eds.), *Longitudinal data analysis*. Boca Raton, FL: Chapman & Hall/CRC Press.

Muthén, B. O., & Asparouhov, T. (2011). Beyond multilevel regression modeling: Multilevel analysis in a general latent variable framework. In J. Hox & J. K. Roberts (Eds.), *Handbook of advanced multilevel analysis*. New York, NY: Taylor & Francis.

Muthén, B. O., & Asparouhov, T. (2012). Bayesian structural equation modeling: a more flexible representation of substantive theory. *Psychological Methods*, 17(3), 313-335

Muthén, B. O., Kaplan, D., & Hollis, M. (1987). On structural equation modeling with data that are not missing completely at random. *Psychometrika, 52,* 431–462.

Muthén, L. K., & Muthén, B. O. (1998–2015). *Mplus user's guide: Statistical analysis with latent variables* (7th ed.). Los Angeles, CA: Muthén & Muthén.

Muthén, B. O. & Satorra, A. (1995). Complex sample data in structural equation modeling. *Sociological Methodology*, 25, 267-316.

Pornprasertmanit, S., Lee, J., & Preacher, K. J. (2014). Ignoring clustering in confirmatory factor analysis: some consequences for model fit and standardized parameter estimates. *Multivariate Behavioral Research*, 49, 518-543

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2011). Alternative methods for assessing mediation in multilevel data: The advantages of multilevel sem. *Structural Equation Modeling*, 18, 161-182

Preacher, K. J., Zhang, Z., & Zyphur, M. J. (2016). Multilevel structural equation models for assessing moderation within and across levels of analysis. *Psychological Methods*, 21 (2), 189-205

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: applications and data analysis methods* (2nd ed.). Thousand Oaks, CA: Sage.

Snijders, T. A. B. (2005). Power and Sample Size in Multilevel Linear Models. In: B. S. Everitt and D.C. Howell (eds.), *Encyclopedia of Statistics in Behavioral Science*. Volume 3, 1570–1573. Chicester, UK: Wiley.

Stapleton, L. M., Yang, J. S., & Hancock, G. R. (2016). Construct meaning in multilevel settings. *Journal of Educational and Behavioral Statistics*, 20 (10), 1-40

van de Schoot, R., Kaplan, D., Denissen, J., Asendorpf, J. B., Neyer, F. J., & Van Aken, M. A. (2014). A gentle introduction to Bayesian analysis: Applications to research in child development. *Child Development, 85,* 842-860.

Yuan, K-H. & Bentler, P. M. (2007). Multilevel covariance structure analysis by fitting multiple single-level models, *Sociological Methodology,* 31(1), 53-82

Jurnal Penelitian dan Evaluasi Pendidikan
Volume 21, No 1, June 2017 (51-64)
Online: http://journal.uny.ac.id/index.php/jpep

PENGEMBANGAN TES KEMAMPUAN BERPIKIR KRITIS PADA MATERI OPTIK GEOMETRI UNTUK MAHASISWA FISIKA

Shan Duta Sukma Pradana [1]*, *Parno* [1], *Supriyono Koes Handayanto* [1]
[1]Universitas Negeri Malang
[1]Jl. Semarang No.5, Sumbersari, Kec. Lowokwaru, Kota Malang, Jawa Timur 65145
* Corresponding Author. Email: shanduta.sp@gmail.com

Abstrak

Penelitian ini merupakan jenis penelitian & pengembangan yang dilakukan dengan tujuan untuk mengembangkan tes kemampuan berpikir kritis. Penelitian ini menggunakan model ADDIE dengan urutan tahapan penganalisisan, perencanaan, pengembangan, pengimplementasian, dan pengevaluasian, tetapi pada penelitian ini hanya dilakukan sampai tahap pengimplementasian. Tes yang dikembangkan dalam penelitian ini terdiri dari lima belas butir soal uraian. Validasi terhadap butir soal tes dilakukan dua kali, yaitu validasi isi dan validasi empiris. Hasil validasi isi menunjukkan bahwa nilai rata-rata butir soal tes sebesar 3,394 berkategori baik, sedangkan hasil validasi empiris menunjukkan bahwa ada sebelas soal berkategori valid dan empat soal berkategori tidak valid. Sebelas soal yang berkategori valid memiliki nilai koefisien reliabilitas Cronbach Alpha sebesar 0,67. Hasil implementasi tes menunjukkan hasil bahwa nilai rata-rata kemampuan berpikir kritis mahasiswa sebesar 27,20 dari 100,00 $(S_D = 11{,}66)$ dengan nilai tertinggi 71,05 dan nilai terendah 2,63. Hal ini menunjukkan bahwa kemampuan berpikir kritis mahasiswa masih kurang.

Kata kunci: *tes kemampuan bepikir kritis, kemampuan berpikir kritis*

DEVELOPING CRITICAL THINKING SKILLS TEST IN GEOMETRICAL OPTIC FOR PHYSICS STUDENT

Abstract

The kind of this study is the research and development that aims to develop critical thinking skills test. This study uses ADDIE model that has five steps: analyze, design, develop, implement, and evaluate, but in this study only done until the implementing step. This test consists of fifteen items essay test. Validation of this test performed twice, content validation and empirical validation. Content validation shows that this test have good categorized with average score 3,394, whereas empirical validation shows that there are eleven items that have valid categorized and four items that have invalid categorized. The valid items have reliability coefficient Alpha Cronbach 0,67. The result of implementing step shows that students have average of critical thinking skill score 27,20 from 100,00 $(S_D = 11{,}66)$ with the higest score 71,05 and the lowest score 2,63. This result shows that students' critical thinking skills are still lacks.

Keywords: *critical thinking skills test, critical thinking skills*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v21i1.13139

Pendahuluan

Kemampuan berpikir kritis merupakan salah satu tuntutan yang harus dipenuhi pada pembelajaran saat ini. Perhatian pembelajaran terhadap kemampuan berpikir kritis disebabkan oleh pengaruhnya bagi orang dalam mengikuti perkembangan ilmu pengetahuan dan tekonolgi yang saat ini berkembang sangat pesat (Luthvitasari, Putra, Linuwih, 2012). Selain itu, kesuksesan dan profesionalitas seseorang juga sangat dipengaruhi oleh kemampuan berpikir kritis yang dimilikinya Quitadamo, Faiola, Johnson, & Kurtz, 2008). Penelitian yang dilakukan oleh Frijters, Dam, & Rijlaarsdam, (2008), menyatakan bahwa jika seseorang memiliki kemampuan berpikir kritis yang kurang, maka orang tersebut akan kesulitan untuk bersaing di dunia global. Pada sisi lain, jika seseorang yang memiliki kemampuan berpikir kritis yang baik, maka orang tersebut dapat ikut serta berperan sebagai konsumen sains (National Research Council, 2012).

Saat ini, banyak penelitian yang mengkaji tentang kemampuan berpikir kritis. Pengkajian tersebut tentu saja memerlukan tes pengukuran agar dapat mengukur kemampuan berpikir kritis dengan tepat. Pengukuran kemampuan berpikir kritis seseorang dapat dilakukan dengan menggunakan tes pilihan ganda berasalan, tes keterampilan (Ennis, 1993; Ennis, 1996), dan tes uraian (Ennis, 1993). Pada penelitian ini dipilih pengembangan tes uraian untuk mengukur kemampuan berpikir kritis mahasiswa fisika.

Jenis penelitian & pengembangan tes kemampuan berpikir kritis telah pernah dilakukan sebelumnya. Ennis (1993) pernah mengembangkan tes kemampuan berpikir kritis, tetapi bebas materi. Selain itu, penelitian & pengembangan tes kemampuan berpikir kritis yang terkait materi juga pernah dilakukan, seperti penelitian Kartimi & Liliasari (2012) yang mengembangkan tes kemampuan berpikir kritis berbentuk pilihan ganda tetapi pada materi termokimia dan Amalia & Susilaningsih (2014) yang mengembangkan tes kemampuan berpikir kritis berbentuk uraian pada materi asam basa. Selain itu, penelitian terkait pengembangan alat ukur kemampuan berpikir kritis juga pernah dilakukan oleh Amarila, Habibah, & Widiyatmoko, (2014) berbentuk tes pilihan ganda, isian singkat, dan uraian pada mata pelajaran IPA tingkat SMP dan Jazuli & Wardani (2015) berbentuk tes uraian pada mata pelajaran IPA tingkat SMP. Akan tetapi, untuk materi fisika masih jarang dilakukan penelitian & pengembangan tes kemampuan berpikir kritis.

Tes yang dikembangkan dalam penelitian & pengembangan ini berupa tes uraian kemampuan berpikir kritis untuk mahasiswa fisika pada materi optik geometri. Pemilihan bentuk tes uraian didasarkan pada karakteristik materi optik geometri yang lebih banyak membuat diagram daripada perhitungan. Selain itu, banyak kesulitan yang dialami dalam mempelajari materi optik geometri, seperti konsep tentang pemantulan dan pembiasan (Aydin, Keleş, & Haşiloğlu, 2012; Galili & Hazan, 2000) dan pembentukan bayangan dari peristiwa pemantulan dan pembiasan (Chang et al., 2007; Parker, 2006; Galili & Hazan, 2000). Tujuan penelitian ini adalah mengembangkan tes kemampuan berpikir kritis untuk mahasiswa fisika pada materi optik geometri dalam mata kuliah Fisika Dasar III. Tujuan lainnya adalah mengetahui validitas dan reliabilitas dari tes yang telah dikembangkan. Selain itu, sebagai tahap implementasi, penelitian ini juga bertujuan untuk mengetahui kemampuan berpikir kritis mahasiswa fisika Univeritas Negeri Malang, khusunya pada materi optik geometri.

Metode Penelitian

Penelitian ini termasuk jenis penelitian & pengembangan dengan mengadaptasi langkah penelitian model ADDIE dari Branch (2009). Terdapat lima tahapan dalam model tersebut, yaitu: (a) *analyze* (penganalisisan), (b) *design* (perencanaan), (c) *develop* (pengembangan), (d) *implement* (pengimplementasian), dan (e) *evaluate* (pengeveluasian). Pada penelitian ini, dilakukan hanya sampai tahap keempat yaitu pengimplementasian. Total waktu yang diperlukan dalam penelitian ini adalah tujuh bulan (bulan Juni sampai Desember 2016).

Tahap pertama adalah penganalisisan. Pada tahap ini dilakukan telaah terhadap tes pengukuran kemampuan berpikir kritis yang sudah ada. Ennis (1993) telah mengembangkan tes uraian untuk mengukur kemampuan berpikir kritis, tetapi tes uraian tersebut bersifat umum. Selain itu, Kartimi & Liliasari (2012) juga telah mengembangkan tes kemampuan berpikir kritis tetapi pada materi termokimia, Amalia & Susilaningsih (2014) pada materi asam basa serta Amarila et al, (2014) dan Jazuli & Wardani (2015) pada mata pelajaran IPA tingkat SMP. Tes berpikir kritis yang berkaitan dengan materi optik geometri untuk mahasiswa fisika masih belum ada. Oleh karena itu dilakukan pengembangan tes kemampuan berpikir kritis pada materi optik geometri untuk mahasiswa fisika.

Tahap kedua adalah perencanaan. Pada tahap ini dilakukan pemilihan patokan dalam pengembangan tes kemampuan berpikir kritis. Patokan yang dipilih adalah lima aspek kemampuan berpikir kritis yang dikememukakan oleh Ennis (1987), yaitu: (a) memberikan penjelasan dasar, (b) membangun keterampilan dasar, (c) menyimpulkan, (d) memberikan penjelasan lanjut, dan (e) strategi dan taktik. Kelima aspek tersebut kemudian menjadi patokan dalam mengembangkan butir soal.

Tahap ketiga adalah pengembangan. Tes kemampuan berpikir kritis dikembangkan dari kelima aspek kemampuan berpikir kritis yang menjadi patokan. Terdapat lima belas butir soal uraian yang dikembangkan dalam penelitian ini. Lima belas soal tersebut mewakili lima aspek kemampuan berpikir kritis. Kelima aspek kemampuan berpikir kritis dan rincian butir soal yang mewakilinya ditunjukkan pada Tabel 1.

Setelah pengembangan tes selesai, tes tersebut kemudian divalidasi. Validasi dilakukan dua kali, yaitu validasi isi dan validasi empiris. Validasi isi dilakukan oleh dua orang dosen Jurusan Fisika FMIPA Universitas Negeri Malang, yang terdiri dari satu dosen ahli materi fisika dan satu dosen ahli pendidikan fisika. Validasi ini meliputi empat aspek, yaitu: (a) kesesuaian butir soal dengan indikator, (b) tingkat kesukaran butir soal (konsep soal), (c) penggunaan bahasa dalam butir soal, dan (d) kebenaran konsep kunci jawaban. Selain itu, validasi isi juga dilakukan untuk mendapatkan saran terhadap butir soal tes dari ahli. Analisis data dari hasil validasi isi dilakukan dengan metode deskripsi rata-rata. Selain itu, butir soal tes juga direvisi berdasarkan saran dari ahli.

Tabel 1. Lima Aspek Kemampuan Berpikir Kritis dan Rincian Butir Soal yang Mewakilinya

No	Aspek Kemampuan Berpikir Kritis	Butir Soal
1	Memberikan penjelasan dasar	1, 2, 3, dan 4
2	Membangun keterampilan dasar	5, 6, dan 7
3	Menyimpulkan	8, 9, dan 10
4	Memberikan penjelasan lanjut	11 dan 12
5	Strategi dan taktik	13, 14, dan 15

Setelah dilakukan tahap revisi, maka dilakukan tahap validasi empiris. Validasi empiris dilakukan terhadap mahasiswa S1 Pendidikan Fisika dan S1 Fisika FMIPA Universitas Negeri Malang angkatan 2015 yang dipilih secara acak sebanyak 68 mahasiswa. Validasi ini digunakan untuk mengetahui validitas dan reliabilitas butir soal. Validitas butir soal dianalisis dengan menggunakan perhitungan koefisien korelasi antara skor butir soal uraian dengan total soal uraian yang dirumuskan (Djaali & Muljono, 2008, p. 86). Butir soal tes dikatakan valid jika $r_{it} > r_{tabel}$. Reliabilitas butir soal dianalisis dengan menggunakan perhitungan koefisien Cronbach Alpha (Djaali & Muljono, 2008, p. 89).

Tahap keempat adalah pengimplementasian. Pada tahap ini, tes yang sudah diketahui validitas dan reliabilitasnya digunakan untuk mengukur kemampuan berpikir kritis mahasiswa fisika. Jumlah responden adalah 109 mahasiswa fisika Universitas Negeri Malang dengan rincian 87 mahasiswa prodi S1 Pendidikan Fisika dan 22 mahasiswa prodi S1 Fisika.

Hasil Penelitian dan Pembahasan

Penelitian & pengembangan ini diawali dengan menentukan tujuan penelitian, yaitu mengembangkan tes kemampuan berpikir kritis dan mengetahui validitas dan reliabilitas tes tersebut. Setelah itu, dilakukan pemilihan patokan yang digunakan dalam pengembangan butir soal tes. Pada penelitian & pengembangan ini dipilih patokan yaitu kemampuan bepikir kritis yang dikemukakan oleh Ennis (1987) yang terdiri dari lima kemampuan berpikir kritis, seperti yang telah dijelaskan sebelumnya. Setelah pengembangan selesai, dilakukan validasi isi untuk mengetahui skor rata-rata butir soal dan mendapatkan saran untuk perbaikan butir soal. Penskoran validasi isi menggunakan lembar penilaian dengan rentang nilai 1 sampai 4.

Hasil Validasi Isi

Validasi isi meliputi empat aspek seperti yang telah dijelaskan sebelumnya. Hasil validasi isi ditunjukkan pada Tabel 2.

Pembahasan Validasi Isi

Dari data Tabel 2, diketahui bahwa semua butir soal memiliki kriteria valid. Rata-rata nilai butir soal adalah 3,394 yang menunjukkan kategori valid dan layak untuk digunakan. Hal ini juga didukung oleh hasil penelitian oleh Jazuli & Wardani (2015) yang mengembangkan alat evaluasi dengan nilai rata-rata hasil validasi isi 3,627 dan layak untuk digunakan. Saran dari hasil validasi isi hanya terdapat pada enam butir soal saja, yang terdiri dari dua jenis saran, yaitu perbaikan jawaban dan kesesuaian soal dengan kemampuan berpikir kritis. Saran terhadap butir soal selengkapnya ditunjukkan pada Tabel 3.

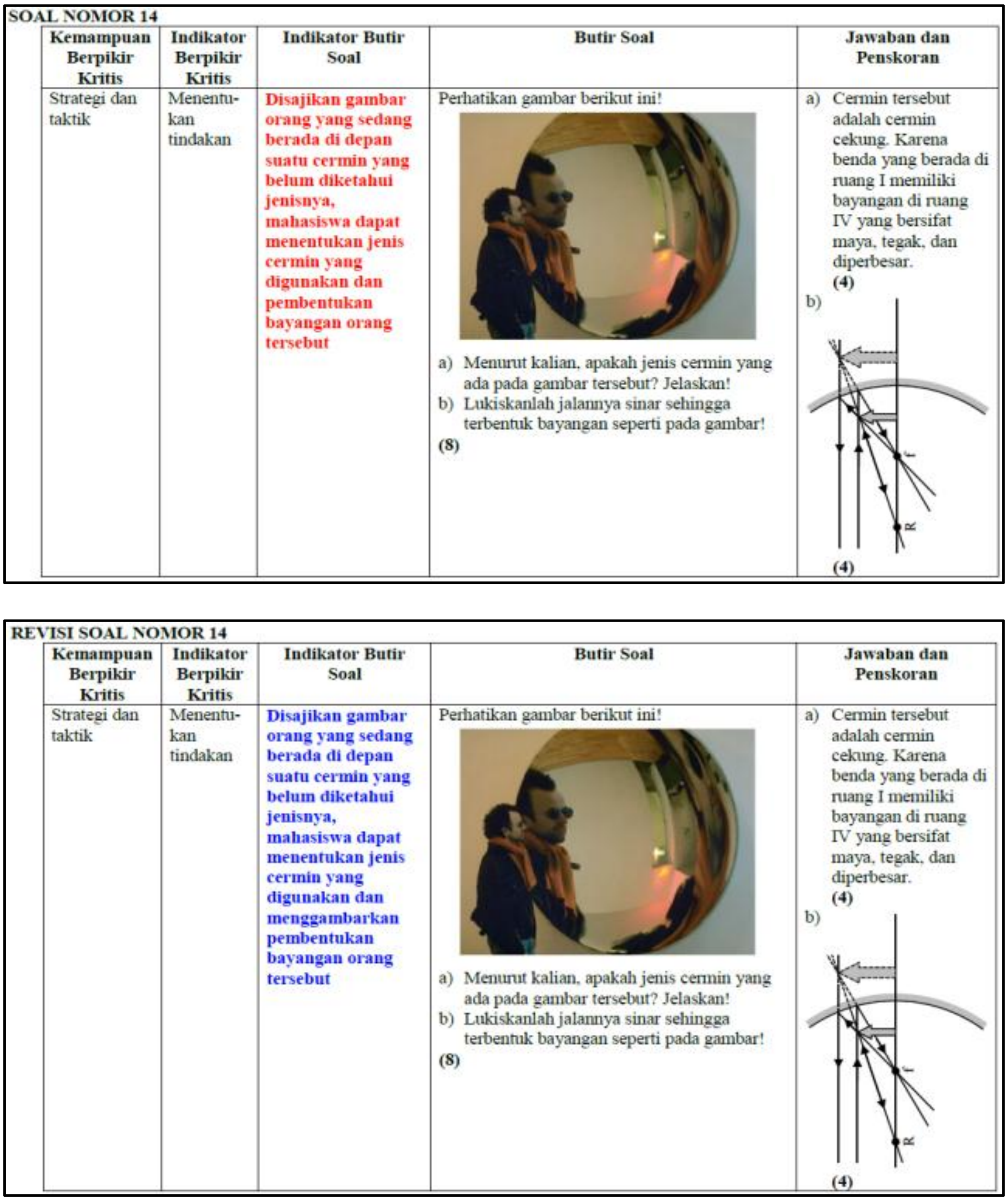
Tabel 2. Rekapitulasi Hasil Validasi Isi

	Pernyataan					
Butir soal	Kesesuaian butir soal dengan indikator	Tingkat kesukaran butir soal sesuai dengan jenjang mahasiswa S1	Butir soal menggunakan bahasa yang mudah dimengerti dan tidak menimbulkan penafsiran ganda	Kebenaran konsep kunci jawaban	Rata-rata	Keterangan
1	4	2,5	3	3	3,125	valid
2	3,5	2,5	3,5	3,5	3,25	valid
3	3,5	3,25	4	3,25	3,5	valid
4	3,5	3,5	4	3,5	3,625	valid
5	3,5	3	3	3,5	3,25	valid
6	3	3	3,5	3,5	3,25	valid
7	3	3	3,5	3,5	3,25	valid
8	3	4	3,5	3,5	3,5	valid
9	4	3,5	3,5	3,5	3,625	valid
10	3,5	3,5	3,5	3,5	3,5	valid
11	3,5	3	3	3,5	3,25	valid
12	3	3	3,5	3,5	3,25	valid
13	3,5	4	3	3,25	3,475	valid
14	3	3,5	3,5	3,5	3,375	valid
15	3	3,5	3,5	3,5	3,375	valid
Rata-rata	3,292	3,375	3,417	3,479	3,394	valid

Tabel 3. Saran dari Hasil Validasi Isi

Butir soal	Saran
2 dan 13	Jawaban disesuaikan dengan pertanyaan
6, 12, 14, dan 15	Butir soal disesuaikan dengan indikator butir soal

Berdasarkan hasil validasi isi, revisi dilakukan terhadap butir soal tes. Contoh revisi terhadap butir soal dapat dilihat pada Gambar 1. Meskipun kelima belas butir soal dinyatakan baik berdasarkan validasi isi, hal ini belum cukup kuat untuk menyimpulkan bahwa butir soal tersebut valid dan reliabel untuk mengukur kemampuan berpikir kritis mahasiswa. Validasi isi tersebut hanya terbatas pada kesesuaian materi dengan kemampuan berpikir kritis. Hasil validasi isi belum bisa menunjukkan bagaimana respon mahasiswa terhadap butir soal tersebut. Oleh karena itu, perlu adanya validasi lanjutan untuk mengetahui tingkat validitas dan reliabilitas butir soal yang telah dikembangkan serta mengetahui respon mahasiswa terhadap butir soal.



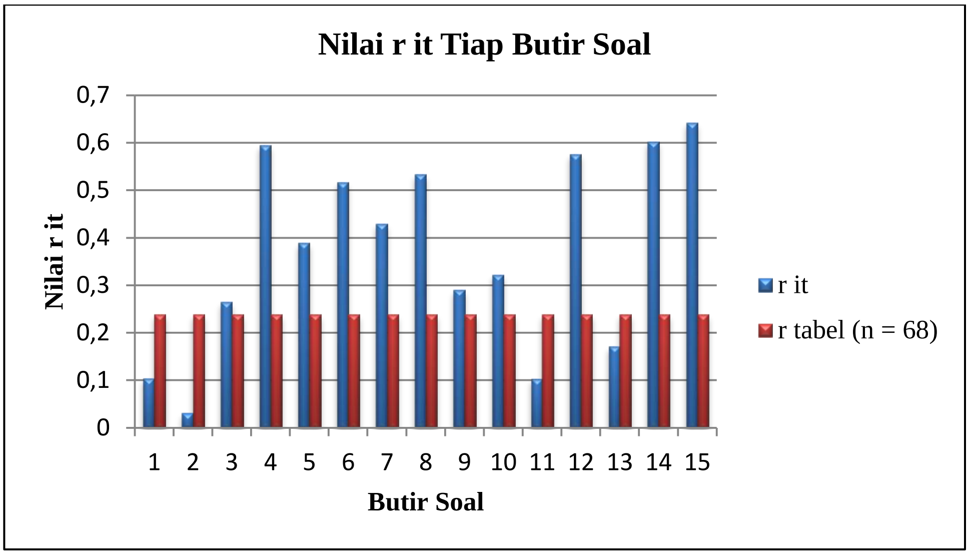
SOAL NOMOR 14

Kemampuan Berpikir Kritis	Indikator Berpikir Kritis	Indikator Butir Soal	Butir Soal	Jawaban dan Penskoran
Strategi dan taktik	Menentukan tindakan	**Disajikan gambar orang yang sedang berada di depan suatu cermin yang belum diketahui jenisnya, mahasiswa dapat menentukan jenis cermin yang digunakan dan pembentukan bayangan orang tersebut**	Perhatikan gambar berikut ini! a) Menurut kalian, apakah jenis cermin yang ada pada gambar tersebut? Jelaskan! b) Lukiskanlah jalannya sinar sehingga terbentuk bayangan seperti pada gambar! **(8)**	a) Cermin tersebut adalah cermin cekung. Karena benda yang berada di ruang I memiliki bayangan di ruang IV yang bersifat maya, tegak, dan diperbesar. **(4)** b) **(4)**

REVISI SOAL NOMOR 14

Kemampuan Berpikir Kritis	Indikator Berpikir Kritis	Indikator Butir Soal	Butir Soal	Jawaban dan Penskoran
Strategi dan taktik	Menentukan tindakan	**Disajikan gambar orang yang sedang berada di depan suatu cermin yang belum diketahui jenisnya, mahasiswa dapat menentukan jenis cermin yang digunakan dan menggambarkan pembentukan bayangan orang tersebut**	Perhatikan gambar berikut ini! a) Menurut kalian, apakah jenis cermin yang ada pada gambar tersebut? Jelaskan! b) Lukiskanlah jalannya sinar sehingga terbentuk bayangan seperti pada gambar! **(8)**	a) Cermin tersebut adalah cermin cekung. Karena benda yang berada di ruang I memiliki bayangan di ruang IV yang bersifat maya, tegak, dan diperbesar. **(4)** b) **(4)**

Gambar 1. Contoh Revisi dari Hasil Validasi Isi



Gambar 2. Nilai Validitas Tiap Butir Soal

Hasil Validasi Empiris

Validasi empiris dilakukan dengan subjek mahasiswa S1 Pendidikan Fisika dan S1 Fisika FMIPA Universitas Negeri Malang angkatan 2015 yang dipilih secara acak sebanyak 68 mahasiswa. Mahasiswa diberi waktu 100 menit untuk mengerjakan butir soal tes secara mandiri. Setelah mahasiswa mengerjakan tes, jawaban mahasiswa kemudian dikoreksi dan dianalisis. Hasil analisisnya ditunjukkan pada Tabel 4. Gambar 2 menunjukkan perbandingan nilai r_{it} dan r_{tabel} untuk setiap butir soal.

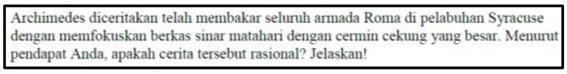
Tabel 4. Hasil Analisis Validasi Empiris

Butir soal	r_{it}	r_{tabel} (n = 68)	Keterangan
1	0,1043	0,2387	Tidak valid
2	0,0312		Tidak valid
3	0,2648		Valid
4	0,5949		Valid
5	0,3899		Valid
6	0,5167		Valid
7	0,4295		Valid
8	0,5334		Valid
9	0,2899		Valid
10	0,3224		Valid
11	0,1035		Tidak valid
12	0,5756		Valid
13	0,1712		Tidak valid
14	0,6025		Valid
15	0,6422		Valid

Pembahasan Validasi Empiris

Berdasarkan Tabel 4 dan Gambar 2, diketahui bahwa terdapat empat butir soal yang tidak valid, yaitu butir soal nomor 1, 2, 11, dan 13. Hal ini berarti bahwa keempat butir soal ini tidak dapat mengukur kemampuan berpikir kritis yang dimiliki oleh mahasiswa. Berdasarkan analisis terhadap jawaban mahasiswa, diketahui bahwa butir soal nomor 2 terlalu mudah, sehingga hampir semua mahasiswa dapat menjawabnya dengan benar. Pada sisi lain, butir soal nomor 1, 11, dan 13 terlalu sulit, sehingga hampir semua mahasiswa menjawabnya dengan salah.

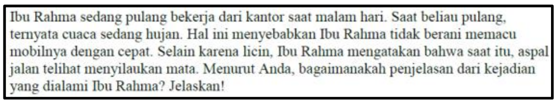
Butir soal nomor 1 ditunjukkan pada Gambar 3. Hasil validasi empiris menunjukkan bahwa 57,35% mahasiswa tidak dapat menjawab dengan benar soal nomor 1. Banyak mahasiswa yang menjawab bahwa kejadian tersebut rasional. Seperti yang diketahui bahwa cermin cekung hanya memiliki satu titik fokus saja. Jika ingin membakar seluruh armada dalam waktu yang bersamaan tidak mungkin dapat dilakukan. Hal ini analogi dengan prinsip kerja kompor surya. Banyak mahasiswa yang menjawab salah pada soal ini, tetapi dapat menjawab dengan benar pada soal yang lain. Hal inilah yang menyebabkan soal nomor 1 tidak valid.



Archimedes diceritakan telah membakar seluruh armada Roma di pelabuhan Syracuse dengan memfokuskan berkas sinar matahari dengan cermin cekung yang besar. Menurut pendapat Anda, apakah cerita tersebut rasional? Jelaskan!

Gambar 3. Soal Nomor 1

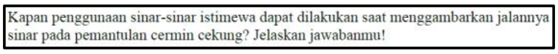
Butir soal nomor 2 ditunjukkan pada Gambar 4. Hasil validasi empiris menunjukkan bahwa 58,82% mahasiswa dapat menjawab dengan cukup benar soal nomor 2. Meskipun cukup benar, tetapi jawaban mahasiswa tersebut hampir sama sehingga mendapatkan nilai yang sama pada soal nomor 2. Selain itu, mahasiswa yang pada nomor lain tidak dapat menjawab dengan benar, dapat dengan mudah menjawab soal nomor 2. Hal ini menyebabkan butir soal nomor 2 tidak valid. Banyak mahasiswa yang menjawab bahwa keadaan silau yang dialami Ibu Rahma disebabkan karena pantulan dari lampu kendaraan lain oleh aspal yang basah sehingga mengenai mata Ibu Rahma.



Ibu Rahma sedang pulang bekerja dari kantor saat malam hari. Saat beliau pulang, ternyata cuaca sedang hujan. Hal ini menyebabkan Ibu Rahma tidak berani memacu mobilnya dengan cepat. Selain karena licin, Ibu Rahma mengatakan bahwa saat itu, aspal jalan telihat menyilaukan mata. Menurut Anda, bagaimanakah penjelasan dari kejadian yang dialami Ibu Rahma? Jelaskan!

Gambar 4. Soal Nomor 2

Butir soal nomor 11 ditunjukkan pada Gambar 5. Hasil validasi empiris menunjukkan bahwa 94,12% mahasiswa tidak dapat menjawab dengan benar soal nomor 11. Meskipun sebenarnya mahasiswa mengetahui macam-macam sinar istimewa, tetapi mereka tidak mengetahui syarat penggunaan sinar istimewa. Sinar-sinar istimewa dapat digunakan untuk melukiskan bayangan jika sinar tersebut berada di dekat sumbu utama. Karena hampir semua mahasiswa tidak dapat menjawab dengan benar soal tesebut, maka butir soal nomor 11 menjadi tidak valid.

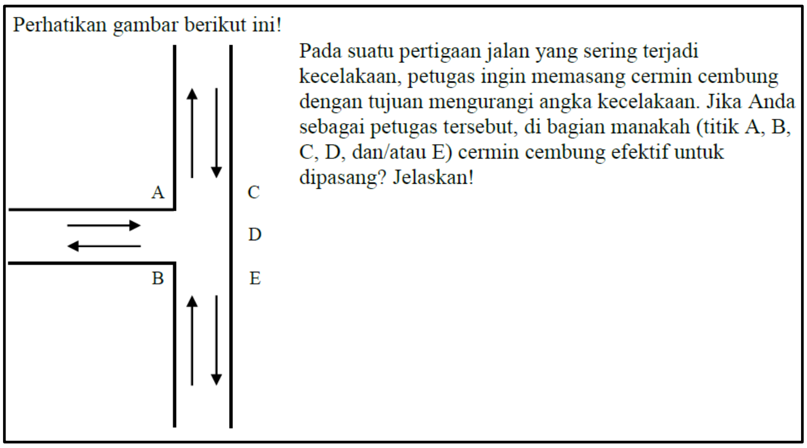


Kapan penggunaan sinar-sinar istimewa dapat dilakukan saat menggambarkan jalannya sinar pada pemantulan cermin cekung? Jelaskan jawabanmu!

Gambar 5. Soal Nomor 11

Butir soal nomor 13 ditunjukkan pada Gambar 6. Hasil validasi empiris menunjukkan bahwa 75,00% mahasiswa tidak dapat menjawab dengan benar soal nomor 13. Sebagian besar responden menjawab bahwa cermin cembung paling efektif diletakkan di posisi D. Padahal jawaban tersebut adalah jawaban yang kurang benar. Seharusnya posisi cermin cembung yang paling efektif adalah pada posisi B dan E. Selain disebabkan karena banyak mahasiswa yang tidak dapatmenjawab dengan benar, petunjuk arah hadap cermin cembung juga tidak dijelaskan pada soal. Hal inilah yang menyebabkan butir soal nomor 13 tidak valid.

Kesebelas soal yang dinyatakan valid telah mewakili lima aspek kemampuan berpikir kritis yang dijadikan patokan dalam penyusunan butir soal tes kemampuan berpikir kritis. Aspek pertama terdapat pada butir soal nomor 3 dan 4. Aspek kedua terdapat pada soal nomor 5, 6, dan 7. Aspek ketiga terdapat pada butir soal nomor 8, 9, dan 10. Aspek keempat terdapat pada butir soal nomor 12. Aspek kelima terdapat pada butir soal nomor 14 dan 15.



Perhatikan gambar berikut ini!

Pada suatu pertigaan jalan yang sering terjadi kecelakaan, petugas ingin memasang cermin cembung dengan tujuan mengurangi angka kecelakaan. Jika Anda sebagai petugas tersebut, di bagian manakah (titik A, B, C, D, dan/atau E) cermin cembung efektif untuk dipasang? Jelaskan!

Gambar 6. Soal Nomor 13

Setelah diketahui ada sebelas butir soal yang dinyatakan valid, maka kesebelas butir soal tersebut diuji reliabilitasnya untuk mengetahui tingkat keajegan saat digunakan untuk mengukur kemampuan berpikir kritis mahasiswa. Berdasarkan hasil perhitungan, didapatkan nilai koefisien reliabilitas Cronbach Alpha adalah $r_{ii} = 0{,}67$ yang berarti butir soal memiliki tingkat keajegan yang tinggi (Arikunto, 2012; Ghozali, 2007), sehingga dapat digunakan untuk mengukur kemampuan berpikir kritis mahasiswa. Validitas dan reliabilitas butir soal yang baik dipengaruhi oleh beberapa faktor. Menurut Istiyono, Mardapi, & Suparno (2014) terdapat empat faktor yang menyebabkan validitas dan reliabilitas baik, yaitu (a) butir soal dikembangkan sesuai dengan prosedur pengembangan, (b) butir soal dikembangkan dari acuan yang tepat, (c) butir soal melalui tahap validasi isi, dan (d) butir soal diuji empiris dengan responden yang mengerjakan dengan sungguh-sungguh dan diawasi dengan ketat. Semua faktor tersebut telah dilakukan dalam penelitian ini, sehingga buir soal dalam penelitian ini memiliki validitas dan reliabiltas yang baik.

Tes yang telah diketahui validitas dan reliabilitasnya kemudian digunakan untuk mengukur kemampuan berpikir kritis mahasiswa fisika Universitas Negeri Malang. Hal ini bertujuan untuk mengetahui bagaimana deskripsi kemampuan berpikir kritis mahasiswa fisika.

Hasil Implementasi Soal Tes

Implementasi butir soal dilakukan dengan subjek 109 mahasiswa fisika Universitas Negeri Malang dengan rincian 87 mahasiswa prodi S1 Pendidikan Fisika dan 22 mahasiswa prodi S1 Fisika. Hasilnya dapat dilihat pada Tabel 5.

Berdasarkan hasil pada Tabel 5, diketahui bahwa kemampuan berpikir kritis mahasiswa fisika Universitas Negeri Malang masih kurang. Hal ini dibuktikan dengan nilai rata-rata kemampuan berpikir kritis mahasiswa hanya pada nilai 27,20. Hasil ini sama dengan hasil penelitian Putra & Sudarti (2015) yang menunjukkan bahwa rata-rata kemampuan berpikir kritis mahasiswa pada nilai 37 dan hasil penelitian Pradana, Parno, & Handayanto (2016) yang menunjukkan bahwa rata-rata kemampuan berpikir kritis mahasiswa fisika adalah 24,29. Sejalan dengan penelitian tersebut, penelitian lain juga mendapatkan hasil nilai rata-rata kemampuan berpikir kritis calon guru fisika adalah 30 (Gunawan & Liliasari, 2012). Ini semakin memperkuat bahwa kemampuan berpikir kritis pebelajar, termasuk mahasiswa di Indonesia masih kurang. Hal inilah yang menyebabkan mahasiswa Indonesia kurang bisa bersaing dalam dunia internasional (Frijters et al, 2008).

Tabel 5. Hasil Penggunaan Tes Berpikir Kritis

Aspek	Nilai
Jumlah responden	109
Nilai rata-rata	27,20
Standar deviasi	11,66
Nilai tertinggi	71,05
Nilai terendah	2,63
Nilai maksimum	100,00

Pembahasan Impelemtasi Soal Tes

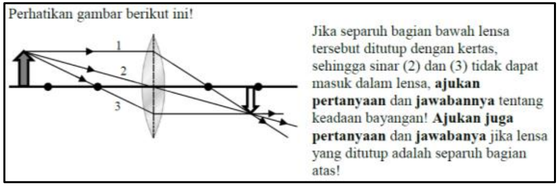
Aspek pertama: Memberikan penjelasan dasar

Aspek pertama diwakili oleh dua soal, yaitu soal nomor 3 dan 4. Rangkuman hasil penelitian untuk aspek pertama disajikan pada Tabel 6.

Tabel 6. Rangkuman Hasil Aspek Pertama

Aspek	Nilai
Nilai rata-rata	27,64
Nilai tertinggi	100,00
Nilai terendah	0,00
Nilai maksimum	100,00

Hal ini menunjukkan bahwa kemampuan berpikir kritis mahasiswa dalam memberikan penjelasan dasar masih kurang. Hal ini berbeda dengan hasil penelitian Dwijananti & Yulianti (2010) yang menunjukkan bahwa pada kemampuan berpikir kritis dalam memberikan penjelasan dasar memiliki nilai rata-rata yang tinggi, yaitu 79,83. Mahasiswa masih belum dapat menganalisis pertanyaan dalam soal yang disajikan. Selain itu, mahasiswa juga masih kesulitan dalam memahami maksud pertanyan dalam soal. Misalnya soal nomor 4 yang ditunjukkan pada Gambar 7.



Perhatikan gambar berikut ini!

Jika separuh bagian bawah lensa tersebut ditutup dengan kertas, sehingga sinar (2) dan (3) tidak dapat masuk dalam lensa, **ajukan pertanyaan** dan **jawabannya** tentang keadaan bayangan! **Ajukan juga pertanyaan** dan **jawabanya** jika lensa yang ditutup adalah separuh bagian atas!

Gambar 7. Soal Nomor 4

Pada soal tersebut terdapat perintah untuk mengajukan pertanyaan dan jawabannya yang disampaikan dengan jelas, tetapi banyak mahasiswa yang tidak memahami maksud soal tersebut. Ada mahasiswa yang hanya memberikan jawaban saja atau pertanyaan saja. Selain itu, tidak sedikit juga mahasiswa yang mengajukan pertanyaan dan jawaban tetapi tidak sesuai dengan ketentuan pada soal.

Penyebab hal tersebut dapat terjadi karena soal dengan tipe seperti soal nomor 4 masih jarang dihadapi oleh mahasiswa. Mahasiswa sering mengahadapi soal yang tidak diminta untuk mengajukan pertanyaan dan jawaban secara bersamaan. Selain jarang menghadapi soal dengan tipe tersebut, terdapat faktor lain yang mempengaruhi mahasiswa tidak dapat menjawab dengan baik soal nomor 3 dan 4, salah satunya adalah pemahaman terhadap materi optik geometri. Mahasiswa masih belum dapat menjelaskan dengan baik tentang peristiwa pembiasan (Aydin et al, 2012; Galili & Hazan, 2000) dan pembentukan bayangan pada lensa tipis (Chang et al., 2007; Parker, 2006; Galili & Hazan, 2000).

Aspek Kedua: Membangun Keterampilan Dasar

Tabel 7. Rangkuman Hasil Aspek Kedua

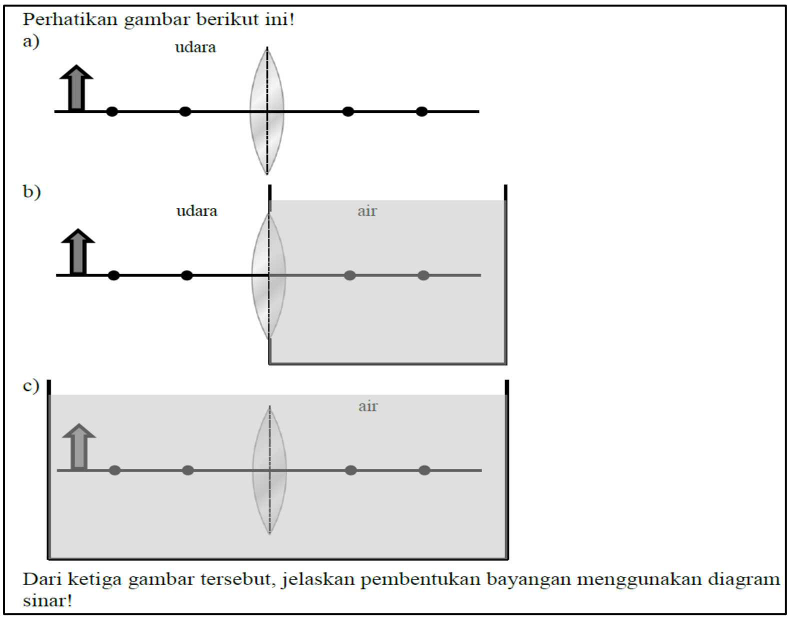
Aspek	Nilai
Nilai rata-rata	25,33
Nilai tertinggi	94,44
Nilai terendah	0,00
Nilai maksimum	100,00

Aspek kedua diwakili oleh tiga soal, yaitu soal nomor 5, 6, dan 7. Rangkuman hasil penelitian untuk aspek kedua disajikan pada Tabel 7.

Hasil tersebut menunjukkan bahwa kemampuan berpikir kritis mahasiswa dalam membangun keterampilan dasar masih kurang. Hal ini juga berbeda dengan hasil penelitian Wahyuni (2015) yang menunjukkan bahwa pada kemampuan berpikir kritis dalam membangun keterampilan dasar memiliki nilai rata-rata yang cukup tinggi, yaitu 67,11. Mahasiswa masih kesulitan saat diminta untuk mengilustrasikan suatu keadaan. Selain itu, mahasiswa juga masih belum bisa memberikan penjelasan dengan menggunakan gambar terkait dengan pengamatan terhadap suatu permasalahan jika ditinjau dari sudut pandang yang berbeda. Sebagai contoh adalah pertanyaan pada soal nomor 5 yang disajikan dalam Gambar 8.

Pada soal tersebut mahasiswa tidak diminta untuk menghitung jarak bayangan dari ketiga gambar, tetapi mahasiswa diminta untuk melukiskan jalannya sinar hingga terbentuk bayangan dari ketiga gambar tersebut. Hasilnya adalah sebagian besar mahasiswa tidak dapat melukiskan jalannya sinar hingga terbentuknya bayangan. Hal ini jelas menunjukkan bahwa kemampuan mahasiswa untuk mengilustrasikan suatu kasus masih kurang. Hal ini juga diperkuat dengan tanggapan mahasiswa saat mengerjakan soal. Mahasiswa mengatakan bahwa mereka merasa kesulitan dan tidak senang jika harus menggambarkan ilustrasi dari soal tes yang diberikan.

Penyebab mahasiswa kesulitan untuk mengilustrasikan adalah kurangnya latihan yang diberikan kepada mahasiswa dalam pembelajaran, terutama fisika. Mahasiswa fisika cenderung langsung dapat menyelesaikan permasalahan jika jelas hal diketahui dan rumus yang digunakan. Selain itu, faktor pemahaman materi juga tidak dapat terlepas dalam mempengaruhi kemampuan yang dimiliki mahasiswa. Pada kemampuan berpikir kritis 2 ini, hasil menunjukkan bahwa mahasiswa masih kesulitan untuk menggambarkan bayangan (Chang et al., 2007; Parker, 2006; Galili & Hazan, 2000) dari peristiwa pembiasan dan pemantulan cahaya (Aydin et al, 2012; Galili & Hazan, 2000).



Perhatikan gambar berikut ini!

Dari ketiga gambar tersebut, jelaskan pembentukan bayangan menggunakan diagram sinar!

Gambar 8. Soal Nomor 5

Aspek ketiga: Menyimpulkan

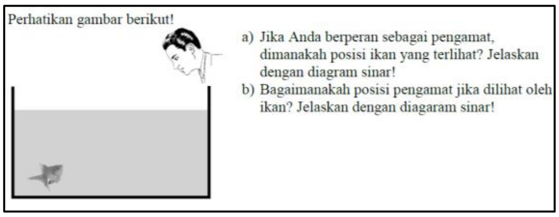
Aspek ketiga diwakili oleh tiga soal, yaitu soal nomor 8, 9, dan 10. Rangkuman hasil penelitian untuk aspek ketiga disajikan pada Tabel 8.

Tabel 8. Rangkuman Hasil Aspek Ketiga

Aspek	Nilai
Nilai rata-rata	12,44
Nilai tertinggi	86,67
Nilai terendah	0,00
Nilai maksimum	100,00

Hasil tersebut menunjukkan bahwa kemampuan berpikir kritis mahasiswa dalam menyimpulkan masih kurang. Hal ini berbeda dengan hasil penelitian Dwijananti & Yulianti (2010) yang menunjukkan bahwa pada kemampuan berpikir kritis dalam menyimpulkan memiliki nilai rata-rata yang cukup tinggi, yaitu 68,32. Sebagai contoh adalah soal nomor 9 yang ditunjukkan pada Gambar 9.

Pada soal tersebut mahasiswa diminta untuk menentukan letak bayangan saat pengamatan dilakukan oleh ikan. Mahasiswa berpendapat bahwa soal tersebut adalah soal yang aneh karena mereka masih belum terbiasa mendapatkan soal yang memerlukan analisis geometri. Soal ini merupakan soal yang tidak memerlukan hitungan, tetapi kemampuan dalam memberikan kesimpulan pengamatan yang dilakukan dari dua sudut pandang berbeda.



Perhatikan gambar berikut!

a) Jika Anda berperan sebagai pengamat, dimanakah posisi ikan yang terlihat? Jelaskan dengan diagram sinar!
b) Bagaimanakah posisi pengamat jika dilihat oleh ikan? Jelaskan dengan diagaram sinar!

Gambar 9. Soal Nomor 9

Penyebab mahasiswa kesulitan dalam memberikan kesimpulan adalah selain pembelajaran yang diberikan masih jarang mengajak mahasiswa untuk menyimpulkan, pemahaman materi juga memepengaruhi kemampuan mahasiswa. Pada kemampuan berpikir kritis 3, mahasiswa masih kesulitan dalam hal perambatan cahaya (Chu & Treagust, 2014; Aydin et al, 2012) dan penentuan letak bayangan (Chang et al., 2007; Parker, 2006; Galili & Hazan, 2000).

Aspek keempat: Memberikan penjelasan lanjut

Aspek keempat diwakili oleh satu soal, yaitu soal nomor 12. Rangkuman hasil penelitian untuk aspek keempat disajikan pada Tabel 9.

Tabel 9. Rangkuman Hasil Aspek Keempat

Aspek	Nilai
Nilai rata-rata	51,76
Nilai tertinggi	100,00
Nilai terendah	0,00
Nilai maksimum	100,00

Hasil ini merupakan nilai rata-rata tertinggi dalam penelitian ini. Hal ini sama dengan hasil penelitian Wahyuni (2015) yang menunjukkan bahwa pada kemampuan berpikir kritis dalam memberikan penjelasa lanjut memiliki nilai rata-rata yang tertinggi, yaitu 79,92. Pada soal nomor 12 ini mahasiswa diminta untuk menjelaskan benar atau tidaknya gambar yang disajikan. Gambar tersebut terdiri dari 6 gambar pembiasan cahaya pada dua medium yang berbeda. Mahasiswa diminta untuk menentukan benar atau tidaknya gambar tersebut serta dilengkapi dengan alasannya.

Nilai rata-rata tertinggi ini disebabkan karena mahasiswa dapat menentukan jawaban yang tepat dari soal tersebut. Meskipun demikian, masih banyak mahasiswa yang kurang dapat memberikan penjelasan lanjut dari jawaban mereka. Hal ini menyebabkan kemampuan memberikan penjelasan lanjut masih pada diri mahasiswa masih perlu dikembangkan lagi. Selain itu, bukti bahwa kemampuan memberikan penjelasan lanjut mahasiswa masih kurang adalah pernyataan mahasiswa saat mengerjakan soal. Mereka mengatakan bagaimana jika jawaban mereka tidak perlu diberikan alasan lebih lanjut. Selain itu masalah ini juga dipenagruhi oleh pemahaman materi yang kurang dari mahasiswa, terutama pada materi pembiasan cahaya (Aydin et al, 2012; Galili & Hazan, 2000) yang menjadi materi dari soal nomor 12 ini.

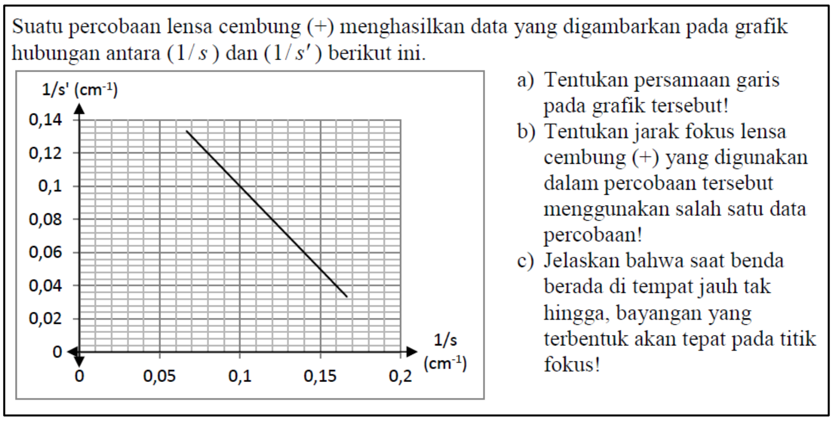
Aspek kelima: Strategi dan taktik

Aspek kelima diwakili oleh dua soal, yaitu soal nomor 14, dan 15. Rangkuman hasil penelitian untuk aspek kelima disajikan pada Tabel 10.

Tabel 10. Rangkuman Hasil Aspek Kelima

Aspek	Nilai
Nilai rata-rata	25,87
Nilai tertinggi	100,00
Nilai terendah	0,00
Nilai maksimum	100,00

Berdasarkan hasil pada Tabel 10, diketahui bahwa kemampuan mahasiswa dalam strategi dan taktik masih kurang. Hal ini sama dengan hasil penelitian Yuliati, Yulianti, & Khanafiyah, (2011) yang menunjukkan bahwa pada kemampuan berpikir kritis dalam strategi dan taktik memiliki nilai rata-rata yang rendah, yaitu 36,27. Mahasiswa masih kesulitan menentukan tindakan untuk menyelesaikan soal. Sebagai contohnya adalah soal nomor 15 yang ditunjukkan pada Gambar 10.



Suatu percobaan lensa cembung (+) menghasilkan data yang digambarkan pada grafik hubungan antara ($1/s$) dan ($1/s'$) berikut ini.

a) Tentukan persamaan garis pada grafik tersebut!
b) Tentukan jarak fokus lensa cembung (+) yang digunakan dalam percobaan tersebut menggunakan salah satu data percobaan!
c) Jelaskan bahwa saat benda berada di tempat jauh tak hingga, bayangan yang terbentuk akan tepat pada titik fokus!

Gambar 10. Gambar 10. Soal Nomor 15

Soal tersebut menyajikan grafik percobaan yang telah dilakukan. Soal tersebut memerlukan hitungan, tetapi tidaklah hitungan yang rumit. Akan tetapi banyak mahasiswa yang kesulitan untuk menyelesaikannya. Mereka kesulitan untuk menentukan persamaan garis, titik fokus lensa yang digunakan, serta menjelaskan dengan hitungan bayangan benda di jauh tak hingga akan tepat di titik fokus.

Faktor penyebab kesulitan mahasiswa adalah selain jarang diberikan permasalahan tersebut, mahasiswa juga kurang memahami maksud dari grafik. Tidak hanya pada meteri optik geometri saja, pada materi lain pun mahasiswa mengalami kesulitan saat diminta untuk membaca atau membuat grafik atau diagaram. Pembelajaran yang diberikan hendaknya melatih kemampuan mahasiswa dalam menyelesaikan permasalahan seperti pada soal nomor 15. Selain itu, mahasiswa juga harus dilatih untuk mengaitkan kejadian kehidupan nyata dengan ilmu pengetahuan yang mereka pelajari. Hal ini bertujuan agar mahasiswa dapat menganalisis dan menentukan suatu tindakan jika menghadapi permasalah di kehidupan nyata dengan menggunakan ilmu-ilmu yang telah mereka pelajari.

Simpulan

Berdasarkan hasil penelitian dan analisis data, diketahui bahwa, dari lima belas butir soal yang telah dikembangkan, terdapat sebelas soal yang memiliki kategori valid $(r_{it} > r_{tabel})$. Kesebelas soal yang dinyatakan valid memiliki tingkat reliabiltas Cronbach Alpha yaitu $r_{ii} = 0{,}67$. Hal ini menunjukkan bahwa kesebelas butir soal tersebut memiliki reliabilitas tinggi, sehingga dapat digunakan untuk mengukur kemampuan berpikir kritis mahasiswa secara valid dan reliabel. Selain itu, kesebelas butir soal tes tersebut telah mewakili kelima kemampuan berpikir kritis yang digunakan sebagai patokan penyusunan tes kemampuan berpikir kritis.

Tes yang telah dinyatakan valid dan reliabel kemudian digunakan untuk mengukur kemampuan berpikir kritis. Jumlah respondennya adalah 109 mahasiswa fisika Universitas Negeri Malang dengan rincian 87 mahasiswa prodi S1 Pendidikan Fisika dan 22 mahasiswa prodi S1 Fisika. Hasilnya adalah nilai rata-rata kemampuan berpikir kritis yang dicapai mahasiswa adalah 27,20. Nilai tertinggi yang dicapai mahasiswa adalah 71,05 dan nilai terendahnya adalah 2,63 $(S_D = 11{,}66)$. Hal ini menunjukkan bahwa kemampuan berpikir kritis mahasiswa masih kurang.

Daftar Pustaka

Amalia, N. F., & Susilaningsih, E. (2014). Pengembangan Instrumen Penilaian Keterampilan Berpikir Kritis Siswa SMA pada Materi Asam Basa. *Jurnal Inovasi Pendidikan Kimia*, Vol. 8, No. 2, pp. 1380-1389.

Amarila, R. S, Habibah, N. A., & Widiyatmoko, A. (2014). Pengembangan alat evaluasi kemampuan berpikir kritis siswa pada pembelajaran ipa terpadu model webbed tema lingkungan. *Unnes Science Education Journal, 3*(2). doi:10.15294/usej.v3i2.3449

Arikunto, S. (2012). *Dasar-dasar evaluasi pendidikan*. Jakarta: Bumi Aksara.

Aydin, S., Keleş, P. U., & Haşiloğlu, M. A. (2012). Establishment for Misconceptions that Science Teacher Candidates have about Geometric Optics. *The Online Journal of New Horizon in Education*, Vol. 2, No. 3, pp. 7-15.

Branch, R. M. (2009). *Instructional Design: The ADDIE Approach*. New York: Springer New York.

Chang, H., Chen, J., Guo, C., Chen, C., Chang, C., Lin, S., … Tseng, Y. (2007). Investigating primary and secondary students' learning of physics concepts in Taiwan. *International Journal of Science Education*, *29*(4), 465–482. https://doi.org/10.1080/09500690601073210

Chu, H.-E., & Treagust, D. F. (2014). Secondary Students' Stable and Unstable Optics Conceptions Using Contextualized Questions. *Journal of Science Education and Technology*, *23*(2), 238–251. https://doi.org/10.1007/s10956-013-9472-6

Djaali & Muljono, P. (2008). *Pengukuran dalam bidang pendidikan*. Jakarta: Grasindo.

Dwijananti, P. & Yulianti, D.. (2010). Pengembangan kemampuan berpikir kritis mahasiswa melalui pembelajaran problem based instruction pada mata kulkiah fisika lingkungan. *Jurnal Pendidikan Fisika Indonesia*, Vol. 6, pp. 108-114.

Ennis, R. H. (1987). *A taxonomy of critical thinking dispositions and abilities, in J. B. Baron & R. S. Sternberg (Eds.), Teaching thinking skills: Theory and practice*. New York: W. H. Freeman.

Ennis, R. H. (1993). Critical thinking assessment. *Theory Into Practice*, Vol. 32, No. 3, pp. 179-186.

Ennis, R. H. (1996). Critical thinking dispositions: their nature and assessability. *Informal Logic*, Vol. 18, No. 2 & 3, pp. 165-182.

Frijters, S., Dam, G., & Rijlaarsdam, G. (2008). Effects of dialogic on value-loaded critical thinking. *Learning and Instruction* (Vol. 18). https://doi.org/10.1016/j.learninstruc.2006.11.001

Galili, I., & Hazan, A. (2000). Learners' knowledge in optics: interpretation, structure and analysis. *International Journal of Science Education,* Vol. 22, No. 1, pp. 57-88.

Ghozali, I. (2007). *Aplikasi multivariate dengan program SPSS*. Semarang: Badan Penerbit Universitas Diponegoro.

Gunawan & Liliasari. (2012). Model virtual laboratory fisika modern untuk meningkatkan disposisi berpikir kritis calon guru. *Cakrawala Pendidikan*, Vol. 2, pp. 185-199.

Istiyono, E., Mardapi, D., & Suparno. (2014). Pengembangan tes kemampuan berpikir tingkat tinggi Fisika (PysTHOTS) peserta didik SMA. *Jurnal Penelitian dan Evaluasi Pendidikan*, Vol. 18, No. 1, pp. 1-12.

Jazuli, M & Wardani, S. (2015). Pengembangan Alat Evaluasi IPA Terpadu Topik Perubahan Materi Berbasis Kontekstual untuk Mengukur Kemampuan Berpikir Kritis Siswa. *Unnes Science Education Journal*, Vol. 4, No. 2, pp. 912-918.

Kartimi & Liliasari. (2012). Pengembangan Alat Ukur Berpikir Kritis pada Konsep Termokimia untuk Siswa SMA Peringkat Atas dan Menengah. *Jurnal Pendidikan IPA Indonesia*, Vol. 1, No. 2, pp. 21-26.

Luthvitasari, N, Putra, N. M. D., & Linuwih, S. (2012). Implementasi pembelajaran fisika berbasis proyek terhadap keterampilan berpikir kritis, berpikir kreatif, dan kemahiran generik sains. *Journal of Innovative Science Education*, Vol. 1, No. 2, pp. 92-97.

National Research Council. (2012). *A Framework for K-12 Science Education.* Washington, DC: The National Academies Press.

Parker, J. (2006). Exploring the Impact of Varying Degrees of Cognitive Conflict in the Generation of both Subject and Pedagogical Knowledge as Primary Trainee Teachers Learn about Shadow Formation. *International Journal of Science Education*. Vol. 28, No. 13, pp. 1545-1577.

Pradana, S. D. S., Parno, & Handayanto, S. K. (2016). *Kemampuan Berpikir Kritis Mahasiswa Tahun Pertama Jurusan Fisika Universitas Negeri Malang*. Makalah disajikan dalam Seminar Nasional Pendidikan IPA. Pascasarjana

Universitas Negeri Malang. Malang, 8 Oktober 2016.

Putra, P. D. A. & Sudarti. (2015). Pengembangan Sistem *E-Learning* untuk Meningkatkan Keterampilan Berpikir Kritis Mahasiswa. *Jurnal Fisika Indonesia*, Vol. 19, No. 55, pp. 45-48.

Quitadamo, I. J., Faiola, C. L, Johnson, J. E., & Kurtz, M. J. (2008). Community-based inquiry improves critical thinking in general education biology. *Life Sciences Education*, Vol. 7, pp. 327-337.

Wahyuni, Sri. (2015). *Pengembangan Bahan Ajar IPA untuk Meningkatkan Kemampuan Berpikir Kritis Siswa SMP.* Makalah disajikan dalam Seminar Nasional Fisika dan Pendidikan Fisika Ke-6. FKIP Universitas Sebelas Maret. Surakarta, 12 September 2015.

Yuliati, D. I., Yulianti, D. & Khanafiyah, S. (2011). Pembelajaran Fisika Berbasis *Hands On Activities* untuk Menumbuhkan Kemampuan Berpikir Kritis dan Meningkatkan Hasil Belajar Siswa SMP. *Jurnal Pendidikan Fisika Indonesia*, Vol. 7, pp. 23-27.

Jurnal Penelitian dan Evaluasi Pendidikan
Volume 21, No 1, June 2017 (65-79)
Online: http://journal.uny.ac.id/index.php/jpep

BIOLOGY LEARNING EVALUATION MODEL IN SENIOR HIGH SCHOOLS

Sri Utari [1]*, *Djukri* [2]
[1)] 8 State Senior High School, [2)] Universitas Negeri Yogyakarta
[1] Jl. Sidobali No.1, Muja Muju, Umbulharjo, Yogyakarta, DIY 55165, Indonesia
[2] Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
* Corresponding Author. Email: utedelayota@gmail.com

Abstract

The study was to develop a Biology learning evaluation model in senior high schools that referred to the research and development model by Borg & Gall and the logic model. The evaluation model included the components of input, activities, output and outcomes. The developing procedures involved a preliminary study in the form of observation and theoretical review regarding the Biology learning evaluation in senior high schools. The product development was carried out by designing an evaluation model, designing an instrument, performing instrument experiment and performing implementation. The instrument experiment involved teachers and Students from Grade XII in senior high schools located in the City of Yogyakarta. For the data gathering technique and instrument, the researchers implemented observation sheet, questionnaire and test. The questionnaire was applied in order to attain information regarding teacher performance, learning performance, classroom atmosphere and scientific attitude; on the other hand, test was applied in order to attain information regarding Biology concept mastery. Then, for the analysis of instrument construct, the researchers performed confirmatory factor analysis by means of Lisrel 0.80 software and the results of this analysis showed that the evaluation instrument valid and reliable. The construct validity was between 0.43-0.79 while the reliability of measurement model was between 0.88-0.94. Last but not the least, the model feasibility test showed that the theoretical model had been supported by the empirical data.

Keywords: *evaluation model, biology teaching, senior high schools*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v21i1.13697

Introduction

This article reviews the Biology learning evaluation model in senior high schools. The evaluation model that will be developed consists of the following components: input, activities, output and outcomes; this model itself refers to the logic model evaluation. Then, the components of input cover students ' facility and initial states; the components of activities cover teachers' performance, students' performance and classroom's atmosphere; the components of output refer to the Biology concept mastery; and the components of outcome refer to the scientific attitude. The evaluation is focused toward the activities or the learning process.

The condition of and the needs toward the Biology learning evaluation model in senior high school lever has become the background of the problems that underlie this review. Evaluation is an important stage in measuring the success of a program. Evaluation toward learning program is necessary in order to measure the learning effectiveness within the efforts of improving the learning quality. Evaluation might encourage the students to be more motivated in their learning programs continuously and might encourage the teachers to improve the quality of their learning process. For schools, evaluation might encourage the improvement toward facilities and school management quality. Up to date, learning evaluation has been prioritizing more on the learning results rather than the learning process. The evaluation of learning results has been based on the results of daily test, mid-semester test, final-semester test, final examination and even national examination without considering the process; whereas, learning process is a black box that might uncover the teachers' and then students ' activities in achieving the learning objectives. The improvement toward the learning process becomes an important step in improving the educational quality. Therefore, there should be a learning program evaluation that might describe the learning process effectiveness.

Biology learning has object characteristics and their problems and, as a result, this learning demands a specific evaluation model. Reinburg (2009, p. 29) states that an individual who have completed his or her Biology learning should understand the main concept of Biology science, the impact of human activities toward biosphere, the inquiry process and the history of biological development. The statement implies that students should understand the Biology concepts, their implementation for solving real-world problems and the scientific investigation process. Finally, a Biology science-literate individual should be able to think creatively, to formulate problems regarding the nature, to have logical and critical reasons, to use efficient technology and to take ethical personal decisions in relation to the biological issues. Therefore, Biology learning should provide facilities in order to achieve that objective.

There are two bases of scientific learning namely process and product. Amien (1987, pp. 16-17) states that the effectiveness of each instructional approach depends on the desired product and process. The final target of learning experience is the implementation of the experience in the future. The learning product might be implemented in the future through the transfer of scientific concepts and attitudes.

Biology learning evaluation in senior high schools should be in accordance to the Biology learning characteristics namely involving the dimension of process and the dimension of results. The dimension of process relates to the activities of designing, implementing and reporting the investigation along with the results (Ludwig & Reynold, 1998, p. 1). Scientific process might be defined as an approach in learning science that has been based on the observation toward what an individual has committed (Rezba, 2007, pp. 4-5; Holt, Rinehart & Winston, 1989, pp. 18-22). The scientific process covers the activities of observing, measuring, classifying, predicting, hypothesizing, investigating, drawing conclusion and communicating results. Then, the dimension of results includes the mastery of scientific concept and attitude.

Theoretical reviews and studies should be conducted in order to provide an alternative of Biology learning evaluation model in senior high schools. The relevant theoretical reviews and studies that have supported this study will be elaborated further. Then, the theoretical review itself includes the Biology learning evaluation characteristics evaluation model and the Biology learning evaluation model.

Program Evaluation Models

There are several models that have been frequently applied in evaluating a program namely CIPP, Kirkpatrick, Stake and logic model. Model selection is mainly based on the evaluation objective. Then, the concept of CIPP (Context, Input, Process and Product) evaluation model was introduced by Stufflebeam in 1965 with his opinion that the important objective of an evaluation has been to improve instead to prove (Madaus, Scriven & Stufflebeam, 1993, p. 118). There are four dimension in this evaluation model, namely context, input, process and product that might be the targets of an evaluation. The CIPP evaluation model might be implemented in any sectors such as education, management and company.

The Kirkpatrick evaluation model is mainly implemented in evaluating the training program that aims to develop human resources. The evaluation model that had been developed by Kirkpatrick is widely known in *Evaluating Training Programs: The Four Levels*. This model covers four level of an evaluation namely reaction, learning, behavior and result (Kirkpatrick, 1988, p. 20).

Logic model is model tool that describes the theory of change that underlies an intervention, a product or a policy (Frechtling, 2007, pp. 21-22). Logic model might be turned into a tool or an approach in order to describe the important elements of a program and to identify the focus of an evaluation. Logic model might be implemented in order to optimize a program through logical and directed planning and evaluation. There are four basic components in the logic model namely: (1) input, namely the human resources that a program has including the fund or the labor contribution; (2) activities, namely the actions that might be willingly conducted in order to achieve the desired objectives; (3) output, namely the direct results of an action that are stated in the form of number such as the number of service, event, document or participation; and (4) outcome, namely the aspect that displays the occurring change or achievement that leads to the final objective.

The review of this study applies the logic model evaluation because the focus of the evaluation is the learning process. Logic model is an evaluation model that has been based on performance and that described the causality logical path so that an individual might understand the components that influence the evaluation results. Logic model consists of four basic components that might be developed and be modified according to necessities.

Characteristics of Biology Learning

Biology has special characteristics that are different to other science in terms of object, problem and method. Biology has a clear scientific structure as having been developed by the Biological Science Curriculum Study (BSCS). According to the BSCS, Biology learning should not only be limited to the textual manner but also be followed by an observation that views any phenomena that might be objects or events. Carind & Sund (1989, pp. 4-5) state that scientific learning aims to train the children to master scientific methods in order to generate scientific products through investigating activities. Biology learning emphasizes scientific process or scientific method in order to generate scientific product.

Scientific process includes the activities of observation, hypothesis formulation, prediction, investigation, data interprettion, inferrence and result communication. Harlen (1992, pp. 83-93) elaborates the indicators of scientific process that might a basis for designing a research instrument.

The types of action that displays the activities of observation are as follows: (1) paying attaention to surrounding objects in detail; (2) identifying similarities and differences; (3) defining sequence of occuring events; and (4) operating assisting tools in order to learn objects in details.

Then, several types of action that might be categorized into the activities of hypothesis formulation are as follows: (1) displaying consistent explanation along with evidence; (2) displayiong consistent explanation along with scientific principles or concepts; (3) implementing knowledge that has been attained previously in providing consistent explanation; (4) realizing that there are more than one possible explanation that might be given to one phenomenon; and (5) realizing that consistent explanation is temporary.

Next, several types of action that might be categorized into the activities of prediction are as follows: (1) using evidence from previous observation or present observation in order to state what might happen; (2) using evidence from previous observation or present observation for extrapolation and interpolation; (3) stating what has happened in relation to the past evidence or experience; (4) recalling the implementation of a pattern that does not have evidence in stating assumption; and (5) differentiating prediction and guess.

Furthermore, the indicators of investigation skills include: (1) deciding independent and dependent variables; (2) manipulating variables in order to carry out appropriate investigation; (3) identifying variables that will be measured or that will be compared; (4) measuring and comparing dependent variables by using the right instrument; and (5) working under appropriate stages.

The indicators of data interpretation and inferrence skills are apparent in the following actions: (1) using information in order to state several meaningful statements; (2) finding certain patterns in observation or investigation results; (3) identifying the relationship between one variable and others; and (4) ensuring that a pattern has been in accordance to data.

Last but not the least, the behaviors of communication results are apparent in the following actions: (1) talking, listening or writing opinions in order to filter ideas or clarifying meanings; (2) taking notes on observation within the investigation process; (3) using graphics and tables in order to deliver information; (4) selecting appropriate communication media so that the informationn might be understood by other people; and (5) using secondary information source.

Biology learning activities that habituate students to implement scientific process will create scientific attitude. Scientific attitude is the attitude toward science that should be developed within the scientific learning (Harlen, 1992, p. 39). This attitude includes curiosity, critical reflection, creativity and discovery, environmental care and cooperation.

Definitions and indicators of scientific attitude has been elaborated by Harlen (1992, pp. 39-44). Curiosity is the attitude of being interested in all matters within the surrounding environment (Harlen, 1992, p. 41). The indicators of curiosity are as follows: (1) being interested in new matters; (2) displaying interest in performing detailed observations; (3) asking many questions in order to find answers; and (4) using multiple information sources2 in order to find new matters.

Respect for evidence is the attitude of not easily trusting opinions and/or information (Harlen, 1992, p.42). This attitude is apparent from the following actions: (1) providing reports according to facts although the facts might be different than expectations; (2) investigating evidence that is different than the pattern that has been found previously; and (3) using conclusions as part of further study.

Next, critical reflection is the analysis or the efforts of reviewing knowledge, understanding and belief that has been understood previously (Harlen, 1992, p. 43). The indicators of critical reflection are as follows:

(1) having desires to review what actions that have been taken; (2) considering alternative procedures that have been implemented; (3) identifying results that deny or support the results of previous studies; and (4) carrying out critical reflection on the results of previous studies in planning and implementing a study.

Learning objectives influence the way teachers teach Biology. The skills of scientific process and attitudes will not be attained if teachers teach Biology genetically. Multiple learning methods might be implemented by teachers in order to achieve expected competencies within the Biology learning process. The scientific learning experts emphasize the importance of scientific and inquiry method. Harlen (2007) recommends the inquiry method within the science learning. Learning by means of inquiry method does not only engage students into learning science but also into learning itself. Both of these aspects are the important results for future of science-literate society. The interaction between students and learning objects and between students and teachers are necessary in the inquiry-based learning process. In such learning,the skills of performing observation, raising questions, clarifying information sources, designing plan, performing experiment, gathering data, analyzing data, interpreting data, performing prediction and communicating results are necessary.

Biology Learning Evaluation in Senior High Schools

Learning program evaluation is an integral part of learning plan (William, 2012, p.2). As the integral part of a program, evaluation will answer the following questions: What will be evaluated? Who will use the evaluation results? What are the criteria that will be applied within the evaluation? Evaluation will help clarifying values, identifying needs and considering alternative manners in order to meet the needs of learning design conceptualization, to conduct the learning process and to improve self-evaluation. In learning evaluatio, there are two most components namely teacher and learning participant; then, the third component is learning supports (librarian, laboratory operator and alike).

Biology learning process evaluation is a process of providing learning activities information in order to determine the learning effectiveness. Teaching, according to the context of standard educational process, is not a mere transfer of knowledge from educators to students . In general there are two concepts of learning process,namely learning as a process of delivering learning materials and learning as a process of regulating environment (Department of National Education, 2008, p.4). The first concept emphasizes learning materials mastery, while the second concept emphasizes process in which students might change their behaviors. This statement implies that teaching-learning process should be turned into the center of the activities. Therefore, teachers should be skillful in mastering learning materials, learning management, evaluation and conducive learning environment.

According to Doran (2009, pp.9-15), based on the focus, evaluation might be categorized into learning evaluation that focuses to students , to teachers, to classrooms and to curricula. The focus and the criteria that will be used in the evaluation influence the type of data that should be gathered. Recommendations from multiple society elements such as educators, parents, students and other interested parties are very necessary. Each group may have different emphasis. The method for attaining these recommendations might be various and might come from multiple questionnaires, checklists and written forms.

The Biology learning evaluation that will be developed in the study will emphasize the teacher performance, the learning participant performance and the classroom atmosphere in accordance to the Biology learning characteristics. The evaluation is focused on uncovering information on the inquiry-based Biology learning activities. The information source is attained from the teachers and the students through the questionnaire and the interview.

Harlen (1992, p.130) explains the guidelines of teacher and learning participant performance according to the objective of scientific learning. The objective of scientific learning is to provide an opportunity for the students to: (1) carry out investigation; (2) develop scientific skills; (3) develop scientific concepts according to the curriculum; (4) develop scientific attitude; (5) develop interests in performing activities; (6) develop learning activities in the daily life; and (7) associate scientific learning to other lessons.

Students' activities in a scientific classroom includes: (1) completing assignment according to the learning objective; (2) enjoying the learning activities; (3) displaying curiosity, concept understanding and discipline; (4) associating the learning activities to the daily life; (5) discussing assignments with peers or teachers; (6) displaying shared cooperation and decision-making activities; (7) displaying critical attitude and open mind; (8) conducting study in order to find answers from the problems provided by teacher; (9) conducting study in order to develop interests; (10) predicting the results that will be attained; (11) observing and recording observation results systematically; (12) classifying and looking for the pattern of observation results; (13) drawing conclusions; (14) selecting accurate and appropriate measurement tools; (15) planning and designing experiment in order to display the concept understanding; (16) completing multiple aspects in a study independently in order to conduct the study thoroughly; (17) checking the results that have not been in accordance to the expectation and repeating the measurement; and (18) applying the results of the study in order to conduct another study so that more convincing results might be attained.

Flick & Lederman (2006, pp. 16-167) state a teacher's enormous role in supporting and developing the students ' thinking capacity. From the psychological perspective, teenage students are often inconsistent in performing their actions. Therefore, there should be an appropriate learning method that might support their metacognitive capacity. Metacognitive capacity is not instantly ready for use in the classroom or in the daily life; instead, this capacity should be demanded, be conducted and be trained. Teachers have a central role in developing this metacognitive capacity through the learning environment that supports the understanding and the implementation of scientific study. Teachers also have a role in creating opportunities for the students to develop the skills that they will use in conducting a study. From these statements, teachers have an important role in designing assignments, selecting learning methods and creating learning environments that support the development of students ' capacity.

Laboratory has another important role in the Biology learning. According to the Committee of National Research Council (Bybee et al., 2006), the learning objectives that should be met as the results of laboratory learning are as follows: (1) improving the mastery of learning materials; (2) developing scientific reasoning; (3) understanding the complexity of empirical works; (4) developing practical skills; (5) understanding the nature of science; (6) growing interest in science and in learning science; and (7) developing group cooperation capacity.

Laboratory learning is frequently conducted by Biology teachers. Several principles in conducting the laboratory learning are namely (Department of National Education , 2008, pp.34-35): (1) learning to do or to practice; (2) curiosity, being encouraged by students ' curiosity; (3) performing scientific thinking, developing the scientific thinking capacity. According to these principles, laboratory therefore is being used to perform experiment and demonstration.

Laboratory and laboratory learning become a general strategy among Biology teachers in order to develop the scientific capacity. Many laboratory activities that should be facilitated and be managed by the teachers in order that the students ' skills may appear; the appropriate use of knowledge might trigger curiosity so that the students will be motivated to perform investigation. Flick & Lederman (2006, pp. 161-162) reviewed a study about laboratory and found several drawbacks. Most of the researchers paid less attention toward students' background, teachers' behaviors, learning environment within classrooms or interaction between teachers and students . As a result, these researchers did not attain a complete description with regards to what actually happened in the class or in the laboratory.

Studies regarding scientific learning process display a description that has been in accordance to the psychological teenage study. Students will be involved in a scientific inquiry if: (1) they are supported by appropriate learning process; (2) they are demanded to work with reflective- and critical-thinking people; and (3) their teachers have necessary knowledge for performing scientific inquiry. If the teachers have low knowledge, do not emphasize critical-thinking activities and propose lower cognitive demands then the students will be less focused toward their inquiry assignments. Learning science through inquiry manner places teachers as secondary information sources within the teacher-learning participant relationship. The teacher-learning participant interaction becomes important in assessing the learning effectiveness.

The success of learning process is influenced by many factors namely teachers, students, supporting facilities and classroom environment. Classroom environment refers to the classroom condition in relation to the learning process that has been marked by a pattern of inter-classroom member interaction or communication. Nasution (2003, pp. 119-120) states that there are three types of classroom situations based on the teachers' attitude. The first type is authoritarian situation where a teacher exerts his or her authority in order to achieve the learning objective and this includes the use of threat and penalty. The second type is permissive situation where a teacher lets the students to develop under emotional freedom. The third type is real situation where the students have their freedom but they are still under control.

Levin & Nolan (1996, p. 147) mention that there are two most important variables that influence the classroom environment namely the physical environment and the classroom regulation, which has been influenced by the culture of the teacher and the students. Teacher has an important role in managing the learning activities so that these activities will be effective. Classroom and school are part of culture. If the school's culture and the students ' culture are appropriate from one to another, then both cultures might improve the positive behaviors and the positive relationship between the teacher and the students. However, if the teachers and the students have different values, norms, behaviors and expectations, then there will be misunderstanding, conflicts and disbelief improvement.

Based on the theoretical review and the results of discussion among the Biology teachers in the senior high schools located in the City of Yogyakarta, the researchers will design a Biology learning evaluation model for senior high schools. The evaluation model will emphasize the learning process that might be apparent from the teacher performance, the learning participant performance and the classroom atmosphere. The evaluation model that will be developed is displayed in the Figure 1.

Method

The study was a research and development research that had been adopted from the Borg & Gall (1983) model. The procedures of research and development activities in the study covered four stages namely: (1) information study and gathering; (2) product development; (3) product testing and model revision; and (4) implementation.

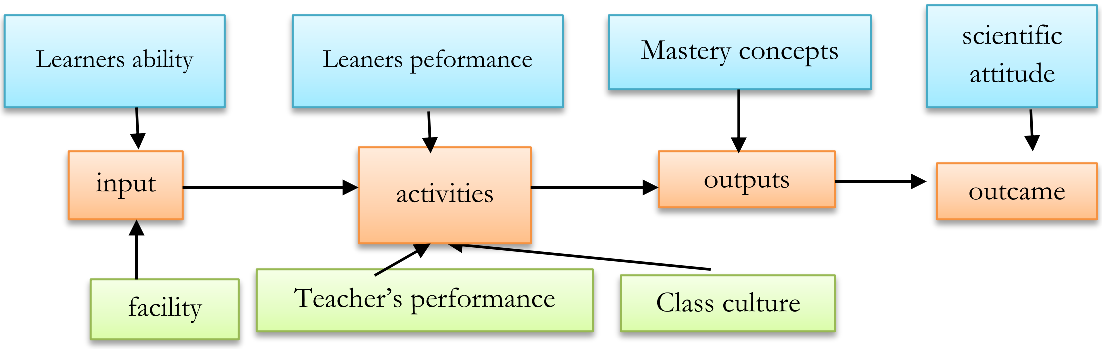


Figure 1. Biology Learning Evaluation Model for Senior High Schools

The subjects in the study were the Biology teachers and the XII Grade students from the City of Yogyakarta. In conducting the instrument readability test, the researchers involved 12 teachers and 32 students. Then, in conducting the instrument experiment, the researchers involved 189 subjects who consisted of 183 students and 6 Biology teachers. In the implementation stage, the test respondents consisted of 16 Biology teachers and 250 students from 16 senior high schools located in the City of Yogyakarta. The selection of the schools as the sample in this study was based on the score of Biology National Examination from the previous year and the composition of these schools was as follows: 4 A-category schools, 4 B-category schools, 4 C-category schools and 4 D-category schools.

The instruments that would be used in the study were questionnaire, observation sheet and interview guideline. The questionnaire would the main instrument in the model development, whereas the observation sheet and the interview guideline would be the supporting instrument in order that the researchers might gather in-depth information regarding the Biology learning program evaluation.

The data would be analyzed by means of Exploratory Factor Analysis and Confirmatory Factor Analysis. The data that had been gathered in the instrument experiment would be analyzed by means of Exploratory Factor Analysis with the assistance of SPSS 17.0 software. The EFA analysis would be conducted in order to identify the presence of variable relationship and to reduce the data so that the researchers might attain new variables or factors. The data that had been attained in the implementation stage would be analyzed quantitatively and would be tested by means of Confirmatory Factor Analysis in order to attain information regarding instrument validity and instrument reliability as well as model fitness. The CFA analysis would be conducted with the assistance of LISREL 8.71 program (Wijayanto, 2008, p. 146).

Results and Discussions

The results of this studying an evaluation model complete with the instrument that might be applied in order to identify the learning process effectiveness. The elaboration of these results covered the evaluation instrument test and the evaluation process. The tests of instrument construct validity and reliability and the measurement model fitness had been an important of this study.

The results of the preliminary study showing that the evaluation that the teachers had conducted had only been in the learning results assessment level. The other type of evaluation was the teacher performance assessment that had been conducted by the principals or the learning supervision that had been conducted by the school supervisors. The program evaluation that associated the learning process and the learning results through an integrated manner had not been implemented in senior high schools all over the City of Yogyakarta. The results of a review toward previous studies showed that there had not been any Biology learning evaluation models that considered the learning process and the learning results in the same place. On the other hand, the existing studies only partly reviewed the learning process and the learning results.

Evaluation Instrument

The instrument designing activities were initiated by theoretical review, review of previous studies and laws that regulated the learning process in senior high schools. These reviews resulted in an initial draft of the instrument in the form of teacher questionnaire and learning participant questionnaire. The teacher questionnaire entailed questions regarding the teacher performance, while the learning participant instrument entailed questions regarding the teacher performance, the learning participant performance, the classroom atmosphere and the scientific attitude.

After conducting the review, the researchers performed a validity test against the evaluation instrument. The evaluation instrument was discussed by the experts of Biology education, measurement and linguistics in order to attain the expert judgment. The experts who had been involved in the study consisted of two Biology learning experts, two measurement experts and one linguistics expert. This stage resulted the following changes: (1) the instrument should be clarified so that the instrument would be appropriate to the Biology learning in senior high schools; (2) the directions of instrument completion should be clarified; (3) the items on the learning participant performance should be added so that this instrument would be more specific; and (4) the language aspects should be improved in order that the instrument would be easier to understand.

The instrument readability test was conducted by the teachers and the students in order to gather feedback qualitatively. The instrument readability test by the teachers involved 11 respondents and this test gathered feedback of improvement on the aspects of language and teacher performance. On the other hand, the instrument readability test by students involved the students of Sports Classroom in the State 4 State Senior High School Yogyakarta and this test gathered feedback of improvement on the direction of questionnaire completion and the aspects of language. This readability test then became the basis of improving the writing mechanics. The revisions turned into a matter of reference in conducting the next stage namely the instrument experiment.

The instrument experiment was conducted in order to attain the valid and reliable instrument. The instrument experiment was conducted in three senior high schools that represented the high, the mauderate and the low senior high school in the City of Yogyakarta and these high schools were State 3 State Senior High School Yogyakarta, State 5 State Senior High School Yogyakarta and State 11 State Senior High School Yogyakarta. The number of returned learning participant questionnaire was 183 bundles but the number of learning participant questionnaire that might be processed was 120 bundles.

The data that had been attained were analyzed in order to identify the size of the reliability coefficient by using the Cronbach's Alpha formula. The Exploratory Factor Analysis (EFA) by means of SPSS 17.0 was conducted in order to explain the dimensions that had been measured. The criteria of reliability coefficient were as follows: if the reliability coefficient was closed to 1.000, then the coefficient would be better; if the reliability coefficient was lower than 0.600, then the coefficient would be inferior; if the reliability coefficient was around 0.700, then the coefficient would be acceptable; and if the reliability coefficient was higher than 0.800, then the coefficient would be good.

The reliability coefficient was apparent in the Table 1.

Table 1. Instrument Reliability

No	Aspect	Reliability Coefficient
1	Teacher Peformance	0.908
2	Students peformance	0.903
3	Class Culture	0.856
4	Scientific attitude	0.805

Based on the variable reliability, the researchers found that the reliability coefficient for all of the variables that had been tested had been higher than 0.800. The reliability score of teacher performance had been equal to 0.908, the reliability score of learning participant performance had been equal to 0.903, the reliability score of classroom atmosphere had been equal to 0.856 and the reliability score of scientific attitude had been equal to 0.805. From these scores, the researchers found that the questionnaire that would be administered in the experiment had high reliability level.

Furthermore, the EFA analysis was conducted in order to identify the presence of inter-variable relationship and to reduce the data so that the researchers would attain new variables and factors which had been simpler. From the EFA analysis, the researchers would like to expect that they would attain the factors that had influenced the learning process and the learning results. The results of EFA analysis might be viewed in the Table 2.

Table 2. Results of EFA Analysis

No	Aspect	EFA Results	
		MSA	Communalities
1	Teacher Performance	0.81 – 0.87	0.57 – 0.86
2	Learning Participant Performance	0.77 – 0.90	0.53 – 0.78
3	Classroom Atmosphere	0.72 – 0.90	0.52 – 0.73
4	Scientific Attitude	0.54 – 0.80	0.51 – 0.80

The MSA scores showed that the variables might still be predicted and might be analyzed further because the score of each variable had been higher than 0.50. The further analysis was conducted in order to decide whether these variables might be grouped into one or several factors. The communalities score showed that the fixed factors might be determined because the average had been higher than 50.00%. From the total score of variance explained, component matrix and rotate component matrix, the researchers determined the categorization of variable input into certain factors based on the size of the correlation between the variables and the factors. The results of the product testing then were used in revising the product.

The EFA from the product experiment resulted in several changes on the number of factors and the items of the questionnaire. The determination on the number of the factors was based on the score of total variance explained, component matrix and rotate component matrix. The categorization of variable input into certain factors was based on the size of correlation between the factors and the variables.

After performing the instrument validation, the readability test and the product experiment, the researchers performed several revisions. The five factors of teacher performance variable now should be reduced into three factors namely the learning management capacity, the students ’ characteristics understanding and the learning evaluation conducting capacity. The learning participant performance still had two factors namely the classroom performance and the laboratory performance. The classroom atmosphere was divided into two factors namely the class support and the self motivation. Last but not the least, the seven factors of scientific attitude now should be reduced into three factors namely the curiosity, the discovery/creativity and the sensitiveness toward surrounding environment.

The instrument that had been revised was administered toward 16 senior high schools located in the City of Yogyakarta. The results of CFA toward this evaluation instrument showed that this instrument that had been administered in the Biology learning evaluation model for senior high schools had good validity and reliability. The validity and the reliability score of this instrument was apparent from t-value score and CR (construct reliability) score in the following Table 3.

Table 3. The Instrument's Validity and Reliability

No	Variable	t-value	CR
1	Learning Participant Performance	4.82 – 7.09	0.88
2	Teacher Performance	5.83 – 10.53	0.94
3	Classroom Atmosphere	7.56 – 10.24	0.89
4	Scientific Attitude	4.91 – 9.08	0.92

The fitness index between the model and the data might be viewed from several GOF (Goodness of Fit) scores namely the score of Chi-Square, RMSEA, GFI and AGFI. These scores measured the fitness of measurement model that had been designed in the evaluation instrument. The objective was to attain empirical evidence regarding the existing factors and indicators in the measurement model of evaluation instrument. The model fitness index of this instrument can be be viewed from the Table 4 until the Table 7.

In several GOF measures that had been apparent from the Table 4 until the Table 7, the researchers found that the measurement model of teacher performance, learning participant performance, classroom atmosphere and scientific attitude had been apparent to the criteria of good fitness. The latent variables of teacher performance were designed by the learning conduct capacity (KG 1, first variable of teacher performance), the learning participant understanding (KG 2, second variable of teacher performance) and the evaluation conduct capacity (KG 3, third variable of teacher performance). The latent variables of learning participant performance were designed by the classroom performance (KPD 1, first variable of learning participant performance) and the laboratory performance (KPD 2, second variable of learning participant performance). The latent variables of classroom atmosphere were designed by the classroom support (IK 1, the first variable of classroom atmosphere) and the self motivation (IK 2, the second variable of classroom atmosphere). The latent variables of scientific attitude were designed by the curiosity (SI 1, the first variable of scientific attitude), the discovery/creativity (SI 2, the second variable of scientific attitude) and the sensitiveness toward surrounding environment (SI 3, the third variable of scientific attitude).

Table 4. The Fitness Index of CFA for the Teacher Performance

	Limit	Fitness Index
Chi-Square	≥ 0.05	χ^2 = 227.70; df=195; p-value = 0.054
RMSEA	≤ 0.08	0.026
GFI	≥ 0.90	0.920
AGFI	≥ 0.90	0.900

Table 5. The Fitness Index of CFA for the Learning Participant Performance

	Limit	Fitness Index
Chi-Square	≥ 0,05	χ^2 = 119.81; df=98.00; p-value = 0.067
RMSEA	≤ 0,08	0.03
GFI	≥ 0,90	0.94
AGFI	≥ 0,90	0.92

Table 6. The Fitness Index of CFA for the Classroom Atmosphere

	Limit	Fitness Index
Chi-Square	≥ 0,05	χ^2 = 32.94 ; df=23.00; p-value = 0.08
RMSEA	≤ 0,08	0.042
GFI	≥ 0,90	0.970
AGFI	≥ 0,90	0.940

Table 7. The Fitness Index of CFA for the Scientific Attitude

	Limit	Fitness Index
Chi-Square	≥ 0,05	χ^2 = 250.45; df= 222.00; p-value = 0.09
RMSEA	≤ 0,08	0,023
GFI	≥ 0,90	0,92
AGFI	≥ 0,90	0,90

The results of Biology learning program evaluation model development for senior high schools results in an instrument that had met the requirements of validity and reliability. This instrument then was implemented in order to identify the level of learning effectiveness. The effective criteria were differentiated into very good (mean score > 4.20), good (mean score ranging around 3.40 – 4.20), moderate (mean score ranging around 2.60 – 3.40) and vey low (mean score < 1.80).

Evaluation Results

The evaluation model consisted of the following components: input, activities, output and outcome. The evaluation profile was presented in order according to the evaluation components namely the input evaluation, the activities evaluation, the output evaluation and the outcome evaluation. The graphic was presented in the sample group A, B, C and D. The input evaluation included students ' initial capacity and supporting learning facilities. The activities evaluation included teacher performance, learning participant performance and class-room atmosphere. The output evaluation included the national examination score and the outcome evaluation included the scien-tific attitude score.

Input Evaluation

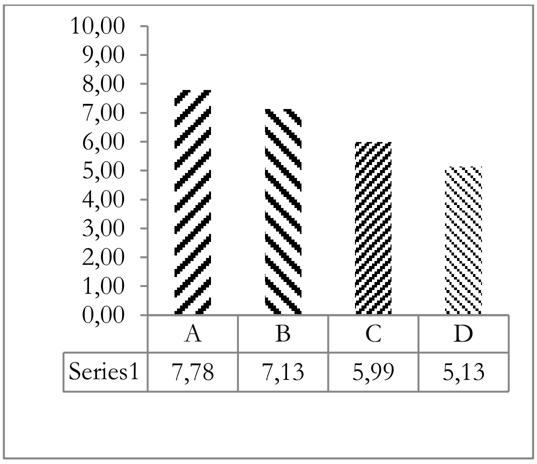


Figure 2. The Graphic of Biology National Examination Average Score in 2013/2014 Academic Year

The school classification based on the national examination scores in the City of Yogyakarta had been relatively stable over the years. The reason was that the senior high school Biology National Examination scores as the learning output indicator had become one of the students ' reference in selecting the school where they would like continue their study. The students with high national examination scores were inclined to select senior high schools that displayed high national examination score. In other words, the average national examination score of a senior high school might describe the students ' input. In the Figure 1, the researchers displayed the differences of Biology National Examination scores from the sampled schools, which had been ranging from 5.13 – 7.78. The average score gap between the A-classified school and D-classified school was 2.65, whereas the average score gap between the B-classified school and C-classified school was 1.14.

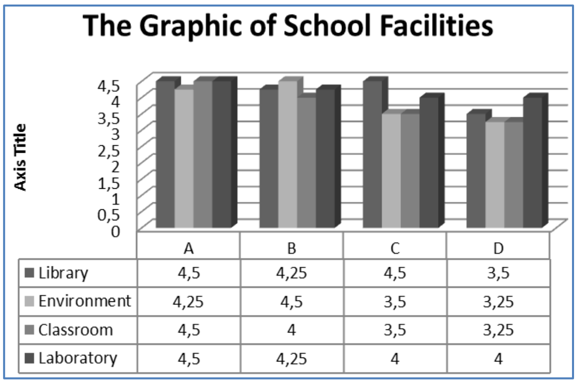


Figure 3. The Graphic of School Facilities

In terms of school facilities, all of the sampled schools had good library, school environment, classrooms and laboratory. On the other hand, in terms of completeness and quality, the A-classified school turned out to be the best.

Activities Evaluation

Learning activities had been the focus of this program evaluation. In the model logic evaluation, the activities were conducted in order to achieve the output. Output referred to the direct results that might be measured from a program. Learning activities would be apparent from the teacher performance, the learning participant performance and the interaction between the teacher and the learning participant in order to create classroom atmosphere that might be helpful for the learning process.

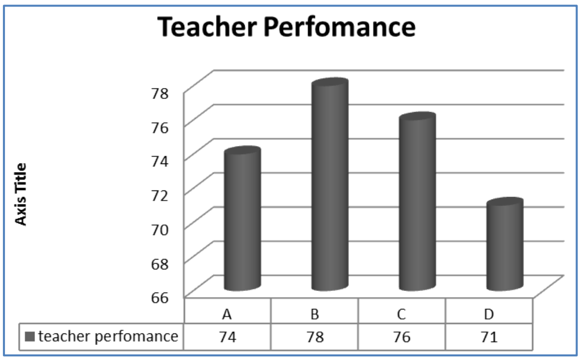


Figure 4. The Graphic of Teacher Performance Achievement Level

The teacher performance achievement level was ranging from 71.00% until 78.00%. The teachers assessed themselves higher than the students. The distribution of teacher quality was quite moderate; in terms average, the rank of teacher performance achievement level from the highest to the lowest was B-classified school, C-classified school, A-classified school and D-classified school. From these data, it was apparent that the highest teacher performance had been found in the B-classified school. Theoretically, teacher performance should influence learning results so that the researchers expected that the highest teacher performance might be found in the A-classified school. This difference showed that there had been other factors that influenced the learning results. The distribution of civil servant-status teachers in the City of Yogyakarta had been determined by the head of Education Office in order to meet the meeting the 24 teaching hours-obligation. Therefore, the schools could not choose the desired the teachers.

The learning participant performance achievement level was ranging from 73.00% until 77.00%; in other words, this achievement level belonged to the good category. The learning activities were determined by the lesson plans that the teachers had prepared and were supported by the students ’ motivation. The lesson plans prepared by the teachers determined the type of the activities, whereas the students ’ motivation determined the quality.

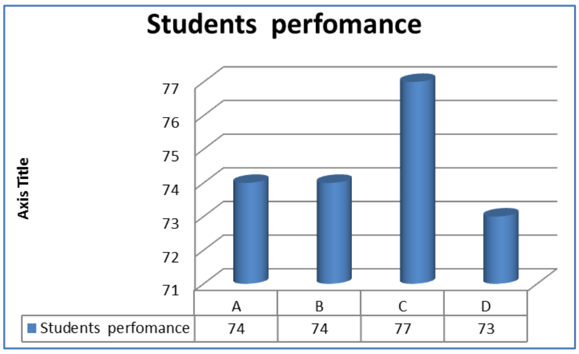


Figure 5. The Graphic of Students Performance Achievement Level

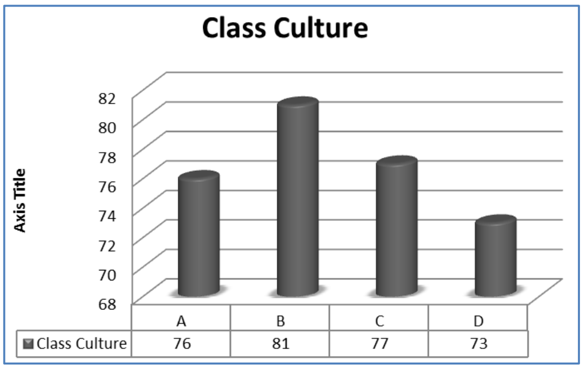


Figure 6. The Graphic of Classroom Atmosphere Achievement Level

The classroom atmosphere effectiveness was ranging between 73.00% until 81.00%. The most effective classroom was found in the B-classified school, followed by the C-classified school, the A-classified school and the D-classified school. The classroom atmosphere was established by the components of self motivation and classroom support. The high self-motivation was supported by the good classroom culture and the good physical environment; in turn, the good classroom culture and the good physical environment established the classroom atmosphere that would be conducive for the learning process.

Output Evaluation

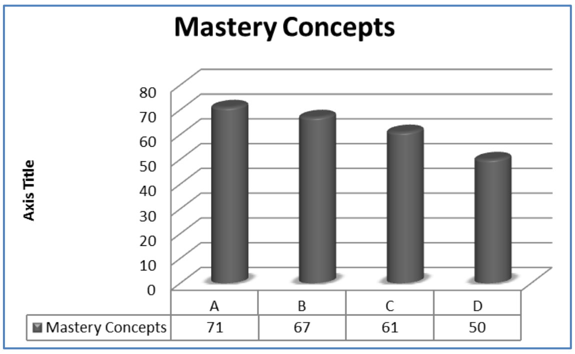


Figure 7. The Graphic of Students ’ Concept Mastery Level

According to the essence of Biology learning, the Biology learning results were in the form of Biology concept mastery and scientific attitude. The Biology concept mastery as the direct result of the learning process might be viewed as the output in the logic model evaluation. The Biology concept mastery was quite various and was categorized into VG (very good), G (good), M(moderate) and P(poor). The categorization referred to the following criteria: 85.00 < SB ≤ 100.00, 70 < B ≤ 85.00, 55.00 < C ≤ 70.00 and 0 < K ≤ 55.00. The average national examination scores that described the students ’ concept mastery might be viewed in the Table 7.

The students ’ concept mastery was ranging from 50.00 until 71.00. These results were in line with the input that had been categorized based on the Biology National Examination scores in the previous year. The order of the concept mastery score did not change from the initial categorization namely the A-classified school that had the highest score and the D-classified school that had the lowest score. This matter implied that students ’ input had a dominant role in determining the concept mastery output. The high input had a tendency to generate the high output, while the low input had a tendency to generate the low output.

Outcome Evaluation

Outcome referred to the indirect results of learning process that had been attained in the long term. Scientific attitude, therefore, might be viewed as the outcome of Biology learning that had been established from a sequence of learning activities that benefitted scientific process. The score of students ’ scientific attitude was ranging from 73.00% until 81.00%. There was also a tendency that the high concept mastery would be followed by the high scientific attitude.

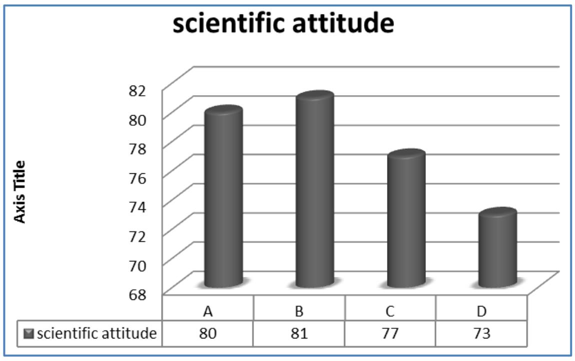


Figure 8. The Graphic of Students ’ Scientific Attitude

The scientific attitude scores that had been almost similar in all schools were related to the learning activities. Flick & Lederman (2006, pp. 161-167) stated the enormous role of a teacher in supporting and in developing the students ’ thinking capacity. Teachers have a central role in developing cognitive capacity through a learning environment that supported the students ’ understanding and scientific study conduct. Several studies showed that teachers often decreased the cognitive demand and directed the students to the intended answers and, as a result, the students had decreasing motivation in accomplishing their investigation assignments.

Conclusions

The Biology learning evaluation model development for senior high schools has results in good evaluation model and instrument for evaluating the teacher performance, the students performance, the classroom atmosphere and the scientific attitude. The results of the significance test have implied that the classroom atmosphere, the teacher performance and the learning participant performance are parts of the learning process. The profile of evaluation results for the Biology learning in the City of Yogyakarta shows that the learning process has been 100% good. On the other hand, the profile of these evaluation results also shows that 31.25% of the learning process has been good, 43.75% of the learning process has been moderate and 25.00% of the learning process has been low.

References

Amien, Moh. (1987). *Mengajarkan ilmu pengetahuan alam (IPA), dengan menggunakan metode "discovery" dan "Inquiry*. Jakarta: Departemen Pendidikan dan Kebudayaan Direktorat Jendral Pendidikan Tinggi.

Borg, W. R. & Gall, M. D. (1983). *Educational research*. New York: Pearson Education.

Carind, A. A., & Sund.R. B. (1989). *Teaching science through discovery*. London: Merril Publishing Company.

Department of National Education. Peraturan Menteri Pendidikan Nasional Republik Indonesia Nomor 3 Tahun 2008. (2008). Jakarta

Doran, R. L. (2009). *Basic measurement and evaluation of science instruction*. Retrieved February 18, 2013, from http://physicsed.buffalostate.edu/pubs/pdf.

Flick, L. B., & Lederman, N. G. (2006). *Scientific inquiry and nature of science*. Netherlands: Springer.

Frechtling, J. A. (2007). *Logic modeling methods in program evaluation*. USA: John Wiley & Sons.

Harlen, W. (1992). *The teaching of science*, London: David Fulton Publisher.

Herlen,W. (2007). *Assesment of learning* . Singapore: Sage.

Holt, et.al. (1989). *Modern Biology*, United State of America: Holt, Rinehart, and Winston Inc.

Kirkpatrick, D. L. (1998). *Evaluating training programs, The four levels.* (2nd ed.). San Fransisco: Barrett-Koehler Publisher, Inc.

Levin, J., & Nolan, J. F. (1996). *Principles of classroom management*. Boston: Allyn and Bacon.

Madaus, G. F., Scriven, M., & Stufflebeam, D. L. (1993). *Evaluation models, viewpoints on educational and human service evaluation*. Boston: Kluwer-Nijhoff Publishing.

Nasution. (2003). *Berbagai pendekatan dalam proses belajar & mengajar*. Jakarta: PT. Bumi Aksara.

Reinburg, C. (2009). *Theacher's handbook*. (4th ed.). Virginia: NSTA Press.

Rezba. (2007). *Learning and assessing Scieence Process Skills*. United States of America: Kendall/Hunt Publishing Company.

Wijayanto, S. H. (2008). *Sructural equation modeling dengan lisrel 8.8*. Jakarta: Graha Ilmu.

Jurnal
Penelitian dan Evaluasi Pendidikan

Jurnal Penelitian dan Evaluasi Pendidikan
Volume 21, No 1, June 2017 (80-92)
Online: http://journal.uny.ac.id/index.php/jpep

AN EMOTION ASSESSMENT MODEL FOR ELEMENTARY SCHOOL STUDENTS

Herwin [1]*, *Djemari Mardapi* [2]
[1]Universitas Cokroaminoto Palopo, [2]Universitas Negeri Yogyakarta
[1]Jl. Latammacelling No.19, Tompotika, Wara, Palopo, Sulawesi Selatan 91911, Indonesia
[2]Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
* Corresponding Author. Email: winunm@gmail.com

Abstract

This study aims to produce an emotion assessment model for elementary school students, identify the characteristics of the quality of the emotion assessment instrument, and obtain information about the results of emotion assessment. The study employed the design and development (D&D) approach. The study was conducted at 9 elementary schools. The data were collected through questionnaires, observations, and interviews. The data analysis techniques were Cohen's Kappa Inter-Rater analysis and Goodness of Fit analysis using Mokken Scalability Analysis. The results of the study show that the emotion assessment model for students consists of six aspects of emotion, i.e.: fear, anger, sadness, boredom, joy, and curiosity. The model consists of 16 indicators and 60 observed items. The emotion assessment model consists of instrument grids, a user's guide, a scoring rubric, and a guide for result interpretation. The emotion assessment model is valid and reliable based on the in inter-rater testing through Cohen's Kappa statistics with an average Kappa coefficient of 0.82 (almost perfect). The results of emotion assessment by teachers are: the fear of elementary school students ranging from high, medium and low category. The anger of students ranges from high, medium and low. The sadness of students ranges from moderate to low category. Boredom of students ranges from medium to low category. The joy of the students is in the high category. The curiosity of students ranges from high and medium category.

Keywords: *assessment model, emotion, elementary school students*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v21i1.14504

Introduction

Pedagogic competence requires that teachers have the ability to understand the emotions of learners and the ability to conduct assessment and evaluation. Tottenham, Hare, & Casey (2011, p. 6) explains that emotion is a very important aspect in the process of child development in general. But the phenomenon that occurred in Soppeng District seems still not in line with expectations. Evaluation activities by teachers so far in Soppeng District tend to focus only on aspects of learning achievement alone. The implementation of judgments on other aspects such as emotions that characterize learners seem rare and even less likely to be applied.

The main problem faced by teachers in Soppeng district in doing emotion assessment of learners is that teachers do not know the instruments that can be used to conduct emotion assessment of learners when the emotions of learners is an aspect that is very important to be understood by the teacher.

Based on the background description of the problem, the formulation of this research problem is (1) How is the model of emotion assessment of learners in elementary school?, (2) What are the characteristics of the quality of the emotion assessment instrument of elementary school students ?, (3) How is the result of the participant's emotion assessment Educated elementary school based on the application of teachers?

This study aims to derive an emotion assessment model of elementary school students, identify the characteristics of the quality of the emotion assessment instruments of elementary school learners and obtain information on the results of emotion assessment of elementary school students based on teacher application.

This research is expected to be useful as a supporting pedagogic competence of teachers in understanding the psychological characteristics of learners. In addition, the results of this study are also expected as a developmental method of assessing the development of elementary school students.

Research Method

This study uses *Design and Development* (D&D). The development procedure consists of six phases: problem identification, goal setting, design and model development, model testing, evaluation of model test results and model deployment. This research was conducted in 9 elementary schools in Soppeng Regency, South Sulawesi Province. The data were collected through questionnaires, observations, and interviews. Data analysis techniques used were Inter-Rater Kappa Cohen analysis and Goodness of Fit analysis using Mokken Scalability Analysis (MSA).

Finding and Discussion

Design Model

The first emotion aspect to be developed is fear. Conceptually, the fear referred to in this study is the emotion state that arises in the learner because of the threat or perceived risk perceived as measured by the indicators of dodging and being quiet. In the initial design the aspect of fear was measured by 2 indicators and 8 observation items. The avoidance indicator is measured by 4 observation items. The items are: Item 1, Item 2, Item 3, and Item 4. The second indicator on the emotion aspect of fear is to be quiet. The indicator of being reticent was measured by four observational items. The items are: Item 5, Item 6, Item 7 and Item 8.

The second aspect of emotion developed is anger. Conceptually, the anger referred to in this study is the emotion state that arises because of the pressure that affects the actions affecting others to follow, obey, and act in accordance with what is desired through threats measured through indicators of scolding, Desire to hit and alienate. In the initial design the aspect of anger was measured by 3 indicators and 10 observation items. The indictment indicator (berated) is measured by 3 observation items. The items are: Item 9, Item 10, and Item 11. The second indicator on the emotion aspect of anger is showing the desire to hit. The indicator is measured by 4 items of observation. The items are: Item 12, Item 13, Item 14 and Item 15. The third indicator on the emotion aspect of anger is seclusion. The alienation indicator is measured by three observational items. The items are: Item 16, Item 17 and Item 18.

The third emotion aspect to be developed is sadness. Conceptually, the sadness referred to in this study is the emotion state of learners that arise when experiencing a loss of importance in the learner as measured by an indicator of silence and crying. In the initial design the aspect of sadness was measured by 2 indicators and 9 observation items. The silence indicator is measured by 5 observation items. The items are: Item 19, Item 20, Item 21, Item 22 and Item 23. The second indicator on the emotion aspect of grief is crying. The indicator of crying is measured by 4 items of observation. The items are: Item 24, Item 25, Item 26 and Item 27.

The fourth emotion aspect developed is boredom. Conceptually, the boredom referred to in this study is the emotion state of learners arising from lack of passion, or encouragement of activities in schools that affect the tendency to not be interested, stop and do not do any more activities that are considered boring as measured through indicators indicate the attitude of not Interested and showing the desire to quit. In the initial design the aspect of boredom was measured by 2 indicators and 8 observation items. The indicator shows the uninterested attitude measured by the 4 items observed. The items are: Item 28, Item 29, Item 30 and Item 31. The second indicator on the emotion aspect of boredom is showing the desire to stop. The indicator is measured by 4 items of observation. The items are: Item 32, Item 33, Item 34 and Item 35.

The fifth aspect of emotion developed is joy. Conceptually, the excitement referred to in this study is the emotion state of learners arising from the achievement of the goal or the existence of something good is happening or experienced as measured through indicator showing a smile or laughter, saying words cheerful words (eg: yes, ok, wah , Hurray, etc.) and shows cheerful behavior (eg jumping, screaming, hugging, etc.). In the initial design the aspect of excitement was measured by 3 indicators and 12 observation items. The indicator shows a smile or laughter measured by 4 observation items. The items are: Item 36, Item 37, Item 38 and Item 39. The second indicator on the aspect of emotion excitement is saying cheerful words of exclamation. The indicator is measured by 4 items of observation. The items are: Item 40, Item 41, Item 42 and Item 43. The third indicator on the aspect of emotion excitement is showing cheerful behavior. The indicator shows that the cheerful behavior is measured by four observational items. The items are: Item 44, Item 45, Item 46 and Item 47.

The sixth aspect of emotion developed is curiosity. Conceptually, the curiosity referred to in this study is the desire of learners to find new information through their activities and experiences in schools as measured by indicators: taking note, taking notes, asking/asking, comparing. In the initial design the aspect of excitement was measured by 4 indicators and 16 observation items. The attention indicator is measured by 4 observation items. The items are: Item 48, Item 49, Item 50 and Item 51. The second indicator on the emotion aspect of curiosity is recorded. The record indicator is measured by 4 observational items. The items are: Item 52, Item 53, Item 54 and Item 55. The third indicator on the emotion aspect of curiosity is asking. The indicator in question is measured by 4 items of observation. The items are: Item 56, Item 57, Item 58 and Item 59. The fourth indicator on the emotion aspect of curiosity is compare. The compare indicator is measured by 4 observational items. The items are: Item 60, Item 61, Item 62 and Item 63.

Model Testing

The first stage of testing is the expert judgment stage. This stage is intended to obtain relevant information relvansi or compatibility between aspects of emotions, indicators and observation items that have been developed in the initial design model. In this research, purposively based on the consideration of expertise, three experts were awarded trust to give an assessment of the initial design model that has been developed.

Expert assessment results are grouped into three groups. The first group is a group of items received without repairs. The second group is a group of items received with improvement. The last group or third group is the item group suggested by the expert to be aborted or excluded. The result of expert assessment related to the grouping is presented in Table 1 as follows.

Table 1. Expert Assessment Results on Emotion Rating Instrument

Expert Judge	Items	Tot
Accepted without repairs	1,4,5,6,7,8,9,10,11,19,20,21,22,23,24,25,26,27,28, 29,30,31,33,35,37,38,39, 40,41,42,43,44,45,46,47, 48,49,50,52,53,54,55,56, 57,58,59,60,61,62,63	50
Accepted with repairs	2,12,13,14,15,16,17,18, 32,34,36	11
Rejected	3,51	2

Source: Expert Rating Results

Based on the data presented in Table 1 it can be explained that the expert assessment results concluded that 50 items assessed by experts can be accepted without improvement, 11 items assessed must be improved and as many as 2 items to be rejected. The items declared acceptable without direct improvement are included in the Revision 1 Model group which will continue on further model testing.

The second model testing stage is a field trial. Field trials were conducted to test the quality of the empirical assessment model. The model of assessment in question is Revision I Model which has been assessed feasible by experts. Instruments judged worthy by the experts is as much as 61 items. The field trial procedure in this phase is to provide a model of assessment to the teacher to be piloted on the schools that have been selected as the study sites. The teacher referred to in this case becomes an assessor or in this research is termed as rater. The results of empirical testing through inter-rater analysis with Kappa Cohen Statistics show the results of the analysis in Table 2 as follows.

Table 2. Summary of Coefficient Analysis Results Kappa

Schools	Object	Koef. *Kappa*	Category
SDN 4 Kalenrunge	PD 1	0.84	*Almost Perfect*
	PD 2	0.80	*Substantial*
SDN 2 Masewali	PD 1	0.77	*Substantial*
	PD 2	0.83	*Almost Perfect*
SDN 168 Kessing	PD 1	0.84	*Almost Perfect*
	PD 2	0.82	*Almost Perfect*
SDN 250 Bulu	PD 1	0.87	*Almost Perfect*
	PD 2	0.85	*Almost Perfect*
MIS Asadiyah	PD 1	0.83	*Almost Perfect*
	PD 2	0.89	*Almost Perfect*
SDN 100 Dare Bunga	PD 1	0.79	*Substantial*
	PD 2	0.85	*Almost Perfect*
SDN 276 Latappere	PD 1	0.75	*Substantial*
	PD 2	0.81	*Almost Perfect*
SDN 97 Ungae	PD 1	0.74	*Substantial*
	PD 2	0.78	*Substantial*
SDN 256 Benteng Jati	PD 1	0.88	*Almost Perfect*
	PD 2	0.84	*Almost Perfect*

Based on the results presented in Table 2, it can be explained that the result of instrument testing in the form of Kappa coefficients revolves around the substantial and almost perfect categories. The coefficient has shown a good agreement index for an instrument to be declared eligible to use. In addition, all Kappa coefficients obtained from the instrument test results have been categorized as reliably. So that the tested instrument has fulfilled the element of reliability.

Emotion scoring models have met the criteria of reliability through empirical testing and testing, but at the stage of empirical trials there are findings that need to be given attention to the refinement of the assessment model being designed. The findings stem from teacher or rater responses when applying the model. At the time of the trial there were items that the teacher found difficult to observe. The item in question is Item 16. The item reads "Learners are silent / unfriendly when lied to by their friends". According to some rater it is very difficult to observe the situation of learners lied to by his friend. Although this happens, it is difficult to observe by the teacher.

Upon the findings, the researchers conducted an evaluation to consider the Item 16. Based on the evaluation result through consideration of suggestion from several rater then Item 16 items are decided to be removed from the instrument. Another consideration on which the item is based is that in some cases in the test this item has a very small situation occurring and the item often experiences different views of the rater. In addition, the issue of Item 16 also does not invalidate the observation indicator that would be measured, since the observer indicator still has other observation items other than Item 16.

Based on the evaluation results of model testing, it is obtained that there are 60 items of observation that are considered feasible to be included in the standard instru-ment of trial result. 60 items are divided into 6 aspects of emotions and 16 indicators. These results are labeled in this study as the Revision II Emotion Assessment Model. The revised Emotion Appraisal II model has basically been standardized through theoretical testing process based on expert analysis or empirically through field trials.

The tests of Goodness of Fit Model using Mokken Scalability Analysis classify the results of the analysis into three groups: scalability testing of pair between items *(H_{ij})*, item sakalability testing *(H_i)*, and scalability testing of all items *(H)*. The following table 3 summarizes the results of scalability testing of pairs between items.

Table 3 concludes that the result of scalability testing of pairs between items indicating the result of all pairs of items (Hij) has been tested measuring one indicator of the same. The result is seen after the whole pair coefficient between items in each indicator shows Hij> 0.3. Another thing that is tested is the sakalabilitas item (Hi). Table 4 below summarizes the results of scalability test items (Hi).

Table 3. Summary of Scalability Tests Pairs Antar Item (Hij)

Indicators	Items	H_{ij}
Eschew	1,2,3	> 0.3
Being introverted	4,5,6,7	> 0.3
Hurl insults (berate)	8,9,10	> 0.3
Hit	11,12,13, 14	> 0.3
Not friendly at the target rage	15,16	> 0.3
Silence	17,18,19, 20, 21	> 0.3
Crying	22,23,24, 25	> 0.3
Attitude shows no interest	26,27,28, 29	> 0.3
Indicated a desire to stop	30,31,32, 33	> 0.3
Show smile or laugh	34,35,36, 37	> 0.3
Cheerful cry uttered	38,39,40, 41	> 0.3
Shows cheerful behavior	42,43,44, 45	> 0.3
Pay attention	46,47,48	> 0.3
Record	49,50,51, 52	> 0.3
Ask	53,54,55, 56	> 0.3
Compare	57,58,59, 60	> 0.3

Table 4. Summary of Scalability Tests Item (Hi)

H_i	Items
≥ 0.3	1,2,3,4,5,6,7,8,9,10,11,12,13,14,15, 16,17,18,19,20,21,22,23,24,25,26,27, 28,29,30,31,32,33,34,35,36,37,38,39, 40,41,42,43,44,45,46,47,48,49,50,51, 52,53,54,55,56,57,58,59,60

Table 4 presents the scalability test items (Hi). The results show that all items (60 items) have good and acceptable power. It is seen after the item sakalabilitas coefficient (Hi)> 0.3. Another test result is the scalability of all items (H). Table 5 below summarizes the results of scalability testing of all items (H).

Table 5. Summary of Scalability Tests All Items (H)

Emotion Aspect	Indicators	*H*
Fear	Eschew	> 0.3
	Being introverted	> 0.3
Anger	Hurl insults (berate)	> 0.3
	Hit	> 0.3
	Not friendly at the target rage	> 0.3
Sadness	Silence	> 0.3
	Crying	> 0.3
Boredom	Attitude shows no interest	> 0.3
	Indicated a desire to stop	> 0.3
Joy	Show smile or laugh	> 0.3
	Cheerful cry uttered	> 0.3
	Shows cheerful behavior	> 0.3
Curiosity	Pay attention	> 0.3
	Record	> 0.3
	Ask	> 0.3
	Compare	> 0.3

Table 5 presents the results of scalability testing of all items. These results indicate that all observed indicators have been fit with the data. For that the model has been tested or has fulfilled the element of Goodness of Fit.

Results of Emotion Assessment

Emotion scoring models that have been tested both theoretically and empirically are given back to the teacher for implementation. This step is called model deployment. This is done to obtain information related to the results of emotion assessment. The results of the emotion assessment in question is the emotion picture of elementary school students. Based on the results of emotion assessment by the teacher obtained the result that the general emotions of elementary school students in Soppeng District vary. The results of the assessment for the emotion aspects of fear are presented in Figure 1 below.

The results of the assessment on the aspect of these fears indicate that the emotions of the students' fears vary from low, medium, and high. In general, the emotions of fear of learners tend to be more dominant in the low category. If it is reviewed based on the gender of the learners then it is informed that female learners tend to have greater fear than male learners. Furthermore, the results of the assessment on anger aspects of emotion are presented in Figure 2 as follows.

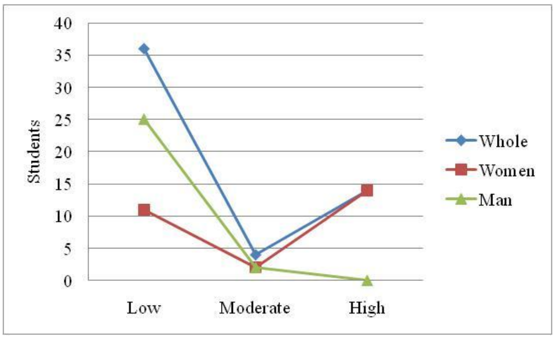


Figure 1. Emotion Aspect Appraisal Results Fear

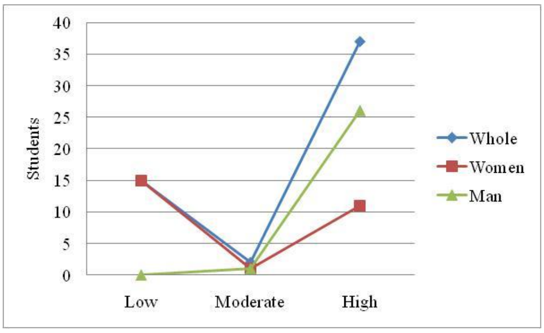


Figure 2. Result of Aspect Emotion Aspect Assessment Angry

The result of the assessment on the anger aspect shows that the anger emotions of learners vary from low, medium, and high. In general, the emotions of anger learners tend to be more dominant in the high category. When viewed on the basis of the gender of learners, it is found that male learners tend to have a greater sense of anger than female learners. Furthermore, the assessment results on the emotion aspects of sadness are presented in Figure 3 below.

The results of the assessment on the aspect of the sadness indicate that the emotions of learners' sadness vary from low, medium, and high. In general, the emotions of student sadness tend to be more dominant in the low category. When viewed based on the gender of learners, it is found that women learners tend to have greater sadness than male learners. The next emotion is boredom. The results of the assessment on the emotion aspects of boredom are presented in Figure 4.

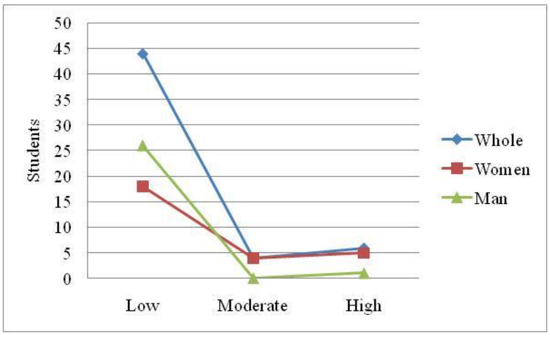


Figure 3. Emotion Aspect Appraisal Results Sadness

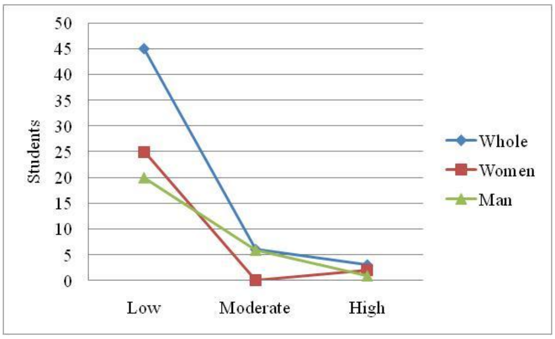


Figure 4. Emotion Aspect Evaluation Results Boredom

The results of the assessment on the aspect of boredom shows that the emotions of boredom learners vary from low, medium, and high. In general, the emotions of boredom learners tend to be more dominant in the low category. When viewed on the basis of the sex of learners then obtained information that male learners tend to have more boredom higher than the female students, but the difference is not so great. The next emotion is excitement. The results of the assessment on the emotion aspects of excitement are presented in Figure 5 below.

The results of the evaluation showed that the joy of emotion excitement of students varies from low, medium and high. In general emotion excitement learners tend to be more dominant in the high category. If reviewed by sex learners says that female students tend to have more joy than male students, but the difference is not so great. The next emotion is curiosity. Results of votes on the emotion aspects of curiosity presented in Figure 6.

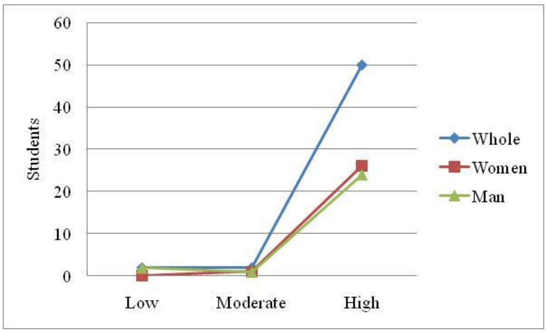


Figure 5. Assessment of Emotion Joy

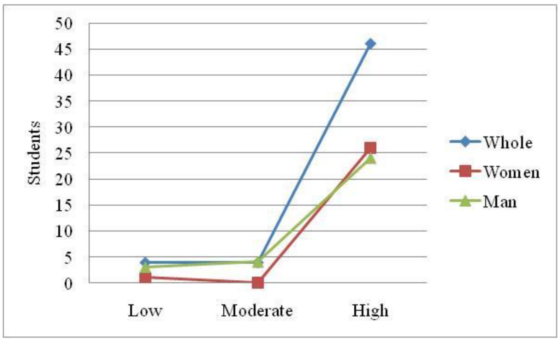


Figure 6. Emotion Aspect Appraisal Results Curiosity

The results of the assessment on the aspect of curiosity shows that the curiosity emotions of learners vary from low, medium, and high. In general, the emotions curiosity of learners tend to be more dominant in the high category. When viewed on the basis of the gender of learners, it is found that female learners tend to have more curiosity compared with male learners, but the difference is not so great.

Discussion of Emotion Rating Model

The model of emotion assessment of learners developed in this study consists of 6 aspects of emotion namely: fear, anger, sadness, boredom, joy, and curiosity. The emotion aspect that has been developed is supported by the results of research conducted by Boehner, DePaula, Dourish & Sengers (2007, p. 289) which states that basically emotion patterns that are generally dominant in childhood are: joy, sadness, fear, Anger, and curiosity of the child. The results of this study indicate conformity with the results of this study or the product developed in this study.

The first emotion aspect developed in the study was fear. The attachment referred to in this assessment model is the emotion state that arises in the learner because of the threat or perceived risk. The fear aspect of this assessment model is measured by the dodge indicator and being quiet. The results of the development correspond to the view that Lerner & Keltner (2001, p. 146) explains that fear is a feeling of risk estimation of something a person will face. Fear is associated with risk aversion. Fear directs a person to avoid risk.

The results of other studies that have been compatible with this study are the results of research from Hansen & Zambo, (2007, p. 274) which explains that fear is the emotion that a person uses to “survival”. When the emotion of fear arises in the child, the child becomes aware of the environment and raises a caution on the child. The results of this study support the emotion assessment model of elementary school students in this study, especially on the emotion aspects of fear that develop indicators to avoid and become quiet.

The second emotion aspect developed in this study is anger. The anger referred to in this study is the emotion state that arises because of the pressure that affects the actions affecting others to follow, obey, and act in accordance with what is desired through threats. In this study anger is measured by indicators throwing insults, hitting, and silencing/unfriendly to angry targets. The results of these developments are supported by Lee & Lang (2009, p. 153) which suggests that anger is generally described as a state of intense emotion in which the desire to attack a reproach object. The results of this study are in conformity with the results of the model of emotion assessment in this study.

This fear aspect of the emotion appraisal model is also supported by Hurlock's (1984) assertion that the child can also show his anger by alienating or remaining silent as a form of deep disappointment in the child. In addition, the fear aspect of this research is also supported by Renshaw & Kiddie, (2012, p. 222) which explains that anger is the basic emotion that often arises when one interprets situations such as hostility. Based on this it makes the basis for researchers to develop indicators of berating, hitting, and unfriendly/silent on the emotion aspects of this study.

The third emotion aspect developed in this study is sadness. The sadness referred to in this study is the emotion state of the learner who appears when experiencing a loss of importance in himself. The aspect of sustainability developed in this study is measured by an indicator of silence and crying. The results of the development supported by Bonanno, Goorin, & Coifman (2008, p. 4) make it clear that the emotions of sadness within a person serve as a form of personal reflection of the sense of loss that can not be prevented. Soreness is generally shown by crying and silence to show sorrow.

The fourth emotion aspect that has been developed in this research is boredom. Boredom is meant in this study is the emotion state of learners that arise due to lack of passion, or encouragement of activities in schools that impact on the tendency to not interested, stop and do not do any more activities that are considered boring. The boredom in this emotion appraisal model is measured by indicators showing a disinterested attitude and showing a desire to quit the activity. The results of the development are supported by research results from Perkun, Goetz, Daniels, Stupnisky, Perry (2010, p. 532) which suggests that boredom is seen as part of an emotion consisting of feelings of unfeeling, lack of stimulation, and low one's passion for something. People who experience boredom have a tendency to run away, get out or not participate from situations that cause boredom. These findings support the outcome of developing an emotion assessment model in this study which concludes that boredom can be measured by observational indicators indicating disinterest and showing a desire to quit the activities.

The fifth emotion aspect developed in this study is excitement. The excitement referred to in this study is the emotion state of learners arising from the achievement of goals or the existence of something good that is happening or experienced. The excitement in this emotion assessment model is measured by indicators showing a smile or laughter, saying cheerful words of cheer and showing cheerful behavior. This is supported by Hurlock (1984) which explains that joy is a pleasant emotion known for joy and happiness or happiness. Each child has a different intensity of excitement and expresses it to some extent. There are various expressions of joy that range from silence, calm, complacency, to an overwhelming in great joy. At the age of school excitement in children is always accompanied by a smile and laughter. The excitement of school-aged children is largely due to the success of children in achieving the goals they expect. Hurlock's opinion indicates the relevance of the results of developing models of emotion assessment of learners in this study.

The last emotion aspect that has been developed in this research is curiosity. Curiosity referred to in this study is the desire of learners to find new information through activities and experiences in school. Emotion curiosity in this study is measured by indicators of paying attention to, taking note, comparing, comparing. The results of this study are supported by the results of research Kashdan, Rose & Fincham, (2004: p.291) that describes the tendency of someone who has a strong curiosity is actively looking for varied sources of new things and new challenges as well as indicate the liveliness of seeking depth of knowledge and experience as a stimulus In him.

Support from other research results from Litman, (2005, p. 793) which suggests that curiosity can be defined as the desire to know, see, or experience that motivates individual behavior and directs it to find new information. It is the basis that the curiosity is a desire learners to find new information through attention activities, record, ask and compare.

The model of emotion assessment of elementary school students developed in this study was designed by Direct Observation Method. The direct observation assessment method that has been developed in this study contains observational situations outlined from the observed indicators to be measured. The selection of this method of emotion assessment is supported by Merrell (2003, p. 51) explaining that the Direct Behavioral Observation or the so-called direct observation method is an emotion assessment method in which the observer as an assessor develops an operational definition of targeted observational behavior, then conducts observations, Recording systematically based on observational subject behavior. This is the basis for researchers to develop the method of direct observation, because the method can reveal the behavior of observation subject directly and systematically.

Discussion of the Characteristics of the Emotion Appraisal Instrument

In this sub-chapter described the related characteristics of the students' emotion assessment instruments. Characteristics of the student's emotion assessment instrument in this case is the feasibility of the instrument in the form of validity and reliability. The validity is content validity through the assessment by experts or experts in the field, while the reliability here is a consistency between rater commonly known as inter-rater technique to assess consistency, closeness and appraisal agreement or rater in doing emotion assessment of learners.

Based on the results of this study obtained the results that the model of emotion assessment of elementary school students have been declared valid in content based on the assessment of some experts or experts who are trusted to provide an assessment. If related to theoretical review as stated by Haynes, Richard, & Kubany (1995, p. 239) which states that the validity of the content can basically be interpreted as evidence of the extent to which elements of the valuation instrument are relevant or are representative of the targeted constructs in an assessment instrument. This view shows that the students 'emotion assessment instruments developed in this study are relevant or representative of the targeted constructs of the learners' emotions.

Another point put forward by Gillespie, Watson, Emery, Lee, & Murchie (2011, p. 2) that the content validity is a description of how far the sample items included in the instrument can measure the content. If associated with the results of this study indicates that the emotion assessment instrument of elementary school learners in this study has measured its content. The intended content is an observational indicator and an emotion aspect. This means that the sample items observed in the emotion assessment instrument have measured the observed indicator and the emotion aspect to be measured.

Another result obtained in this study is reliability. Based on the results of the study showed that the students' emotion assessment instruments have been reliable. If the result is associated with the statement put forward by Ziegler & Detje (2012; p. 3) which explains that reliability describes the overall consistency of the measurements though given several times. Measurements with high reliability are said to be reliable measurements. Reliability itself has other names such as reliability, reliability, stability, stability, consistency, and so forth. However, the central idea embodied in the concept of reliability is the extent to which a measurement is reliable.

In the results of this study the instrument of emotion assessment of elementary school students has been reliably through inter-rater reliability techniques. On the other hand Graham, Milanowski, Milner, & Westat (2012, p. 4) explains that inter-rater reliability basically shows that different observers tend to provide relatively similar assessments on the same observational object so that it indicates that Instruments have been reliable, consistent and reliable. This view shows that the emotion assessment of elementary school learners that have been developed in this study has been reliable, consistent and credible even if applied by different observers or teachers.

The next result in this study is the results of statistical tests Kappa showed the average coefficient of 0.82. If the results are compared with Landis & Koch's (1977, p. 165) view that the coefficients are in the almost perfect category. Other experts who gave explanations related to the coefficient of Kappa namely Bonagamba, Coelho and Anamaria (2010, p. 435), then the coefficient of 0.82 the results of the study has been on the category exellent. Based on some of the views of these experts it can be collected that the emotion assessment instrument of elementary school learners that have been developed in this study has been feasible to use because it already has consistency, reliability and the results obtained from such instruments can be trusted.

Discussion on the Application of Emotion Appraisal Model

The result of applying the emotion appraisal model referred to in this research is divided into two: the results of teacher assessment and teacher perceptions on the model of emotion assessment of learners. Based on the results of emotion assessment of learners from teachers, then obtained the result that for the emotion fears of learners ranged from low, medium and high. Viewed from the aspect of gender then the results of research that for the fear aspect of female students have a tendency to fear higher than the male students. The results of this study are supported by Hurlock (1984) which states that girls show more fear than boys and the daughters fear is socially acceptable.

In the anger aspect, it is found that the emotions of learners range from low. If viewed from the aspect of gender then the results of this study indicate that male students have a tendency of a higher sense of anger compared with female learners. This is supported by the opinion of Aldrich &

Tenenbaum (2006, p. 776) explains that if viewed based on the emotion of anger, then the boy is more emotion than the girl. This means that men more easily feel angry than girls. Anger is interpreted as a masculine emotion. On the other hand Renati, Cavioni, & Zanetti, (2011, p. 49) argued that anger is an emotion that is not easy to manage especially for elementary school age learners. For elementary school children in the classroom, anger can be caused by a dispute or seizure of possession of an object, a mismatch between peers, physical stress, rejection or neglect and when a child is forced to do something he does not like.

In the emotion aspects of grief, the results obtained that the emotions of learners ranged from low to moderate. Viewed from the aspect of gender then the results of this study indicate that female students have a tendency of higher sadness compared with male learners. The results of the study were supported by Aldrich & Tenenbaum (2006, p. 776) explaining that if viewed on the basis of emotion sadness, then girls are more emotion than boys. This means that women are more easily to feel sorrow than boys. Sorrow is interpreted as a feminine emotion.

For emotion aspects of boredom assessment results obtained that the boredom of learners ranged from low to moderate. If viewed from the side of the sex then there is no difference in boredom between female learners with male students. Based on the theoretical review of Perkun et al, (2010, p. 532) suggests that boredom is seen as part of an emotion consisting of unpleasant feelings, lack of stimulation, and low one's passion for something. People who experience boredom have a tendency to run away, get out or not participate from situations that cause boredom. In addition, boredom can also be expressed with monotonous activities such as daydreaming. Perkun, et al, (2010, p. 545) concluded in his research that boredom indicates in children such as lack of concentration in learning, lack of attention to the lessons followed, and consequently, boredom has a negative effect on academic achievement.

The results of this study did not find any high boredom in elementary school students. Indications of boredom described by various experts are not found in this study. It is a common hope that elementary school students especially in schools can learn without being haunted by boredom, because the boredom can be bad for students in school.

For the emotion aspect of the joy of elementary school students in Soppeng Regency has a uniform result that is high category. Lee & Lang (2009, p. 151) who argued that joy as a very pleasant emotion when one is in the context of progressing toward the desired goal. A sense of excitement arises when a person's goal is achieved, either the expected objective or the abrupt goal that arouses one's passion. When feelings of happiness arise then the motivation moves or increases.

If the results of this study compared to the view of Lee & Lang (2009) is then certainly obtained a positive result related to the emotions of joy in elementary school students in Soppeng district. It shows that learners are in the context of getting progress toward the desired goals. But the excitement must also continue to be controlled by the teacher, as Scherer, (2005, p. 723) explains that joy is a positive emotion dimension that requires high control for those who feel it. This suggests that teachers should continue to monitor the joy of their high-end learners. Do not let it is not controlled so it will also be less good for learners.

The emotion aspect of curiosity is the result that the curiosity of learners ranges from moderate to high. The results of this study are supported by Hurlock (1984) explaining that for the curiosity aspect of the child has a tendency to react positively to new elements they find. This suggests that the child's age is the time when a person has a tendency to react positively to new elements they find. The results of this study are in line with the theoretical concepts described by Hurlock (1984). In general, elementary school children have a strong curiosity tendency, but the strong curiosity should always be controlled by teachers at school, as well as parents at home.

Conclusion and Recommendation

The results of the study show that the emotion assessment model for students consists of six aspects of emotion, i.e.: fear, anger, sadness, boredom, joy, and curiosity. The model consists of 16 indicators and 60 observed items. The emotion assessment model consists of instrument grids, a user's guide, a scoring rubric, and a guide for result interpretation. The emotion assessment model is valid and reliable based on the in inter-rater testing through Cohen's Kappa statistics with an average Kappa coefficient of 0.82 (almost perfect). The results of emotion assessment by teachers are: the fear of elementary school students ranging from high, medium and low category. The anger of students ranges from high, medium and low. The sadness of students ranges from moderate to low category. Boredom of students ranges from medium to low category. The joy of the students is in the high category. The curiosity of students ranges from high and medium category.

Based on the conclusions obtained in this study, it is suggested the following matters. (1) this model of emotion assessment of elementary school students is suggested to be applied by teachers to understand the characteristics of learners that make the basis for teachers both to plan the learning process in accordance with the characteristics of learners, as well as provide an approach in social interaction with students in the school; (2) prior to implementation in schools, teachers are advised to attend training in advance in order to obtain information on how to apply the model of emotion assessment properly; (3) the model of emotion assessment of elementary school students developed in this research has been tested both theoretically and empirically, so it is suggested that the model of emotion assessment can be applied continuously in Soppeng District or in other places or areas. (4) to other researchers interested in similar topics to develop other aspects of emotion assessment, because it is very useful for teachers and learners in particular and the development of the world of education in general.

Limitation

This study aims to obtain a model of emotional assessment of primary school students. Based on the consideration that emotions have a very wide scope aspect, but in this research only developed six aspects of emotion, among others: fear, anger, sadness, boredom, joy, and curiosity. This is felt to be a limitation in this study. To that end, the development and expansion of other aspects of emotion is necessary to continue both by researchers themselves and other researchers who are interested in similar topics.

This emotional appraisal model is only designed and used in high school students. Nevertheless, researchers continue to realize that developing a model of emotional assessment for low-grade primary school learners is also important. Therefore, the development of a further model of emotional assessment for primary school students in lower classes is considered necessary for future research.

References

Aldrich, N. J., & Tenenbaum, H. R. (2006). Sadness, anger, and frustration: Gendered patterns in early adolecents' and their parents' emotion talk. *Sex Roles Journal*, *55*, 775-785.

Boehner, K., DePaula, R., Dourish, P., & Sengers, P. (2007). How emotion is made and measured. *International Journal of Human Computer Studeis*, *65*, 275-291.

Bonagamba, G. H., Coelho, D. M., & Anamaria, O. (2010). Inter and Intra-Rater Reliability of Scoliomi. *Rev Bras Fisioter*, *14*(5).

Bonanno, G. A., Goorin, L., Coifman, K. G. (2008). Sadness and grief. *The* *handbook of emotion.* (3rd Ed.). New York: Guilford.

Gillespie, H. S., Watson, T., Emery, J. D., Lee, A, J. Murchie, P. (2011). A questionnaire to measure melanoma risk, knowledge and protective behaviour: Assessing content validity in a convenience sample of Scots and Australians. *Medical Research Methodology*, *11*, 1-9.

Graham, M., Milanowski, A., Milner, J., & Westat. (2012). *Measuring and promoting inter-rater agreement of teacher and principal performance ratings*. U.S: CECR.

Hansen, C. C., & Zambo, D. (2007). Loving and learning with wimberly and david. foresting emotional development in early childhood education. *Early Childhood Education Journal, 178*(3), 259-272.

Haynes, S. N., Richard, D. C. S., & Kubany, E. S. (1995). Content Validity in Psychological Assessment: A Functional Approach to Concepts and Methods. *Psychological Assessment, 7*(3), 238-247.

Hurlock, E.B. (1984). *Child development.* (6th Ed.). Singapore: McGraw-Hill International Book Company.

Kashdan, T. B., Rose, P., & Fincham, F. D. (2004). Curiosity and exploration: Facilitating positive subjective experiences and personal growth opportunities. *Journal of Personality Assessment*, *82,* 291-305.

Landis, J. K., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics, 33*(1), 159-174.

Lee, S., & Lang, A. (2009). Discrete emotion and motivation: Relative activation in the appetitive and aversive motivational systems as a function of anger, sadness, fear, and joy during televised information compaigns. *Media Psychology*, *12,* 148-170.

Lerner, J. S., & Keltner, D. (2001). Fear, anger, and risk. *Journal of Personality and Social Psychology, 81*(1), 146-159.

Litman, J. A. (2005). Curiosuty and the pleasures of learning: Waiting and lingking new information. *Psychology Press, 19*(6), 193-814.

Merrell, K. W. (2003). *Behavioral, social, and emotional assessment of children and adolescents.* New Jersey: Lawrence Erlbaum Associates.

Perkun, R., Goetz, T., Daniels, L. M., Stupnisky, R. H., Perry, R. P. (2010). Boredom in achievement settings: Exploring control–value antecedents and performance outcomes of a neglected emotion. *Journal of Educational Psychology, 102*(3), 531-549.

Renati, R., Carvoni, V., Zanetti, M. A. (2011). 'Miss, i got mad today!' the anger diary, a tool to promote emotion regulation. *The International Journal of Emotion Education,* 3(1), 48-69.

Scherer, K. R. (2005). What are emotions? and how can they be measured?. *Sage Publications, 44,* 695-729.

Tottenham, N., Hare, T., & Casey, B. J. (2011). Behavioral assessment of emotion discrimination, emotion regulation, and cognitive control in childhood, adolescence, and adulthood. *Frontiers in Psychology, 2,* 1-9.

Ziegler, J & Detje, F. (2012). Aplication of Empirical Methodology to Evaluate Information Fusion Aproaches. *International Journal Methodology, 20,* 327-337.

Jurnal
Penelitian dan Evaluasi Pendidikan

Jurnal Penelitian dan Evaluasi Pendidikan
Volume 21, No 1, June 2017 (93-103)
Online: http://journal.uny.ac.id/index.php/jpep

DEVELOPMENT AND VALIDITY OF MATHEMATICAL LEARNING ASSESSMENT INSTRUMENTS BASED ON MULTIPLE INTELLIGENCE

Helmiah Suryani [1]*, *Badrun Kartowagiran* [2], *Jailani* [2]
[1]SMAN 8 Samarinda Kalimantan Timur, [2]Universitas Negeri Yogyakarta
[1]Jl. Untung Suropati, Karang Asam Ulu, Sungai Kunjang, Samarinda, Kalimantan Timur 75243
[2]Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
* Corresponding Author. Email: emysuryani29@yahoo.com

Abstract

This study was aimed to develop and produce an assessment instrument of mathematical learning results based on multiple intelligence. The methods in this study used Borg & Gall-Research and Development approach (Research & Development). The subject of research was 289 students. The results of research: (1) Result of Aiken Analysis showed 58 valid items were between 0,714 to 0,952. (2) Result of the Exploratory on factor analysis indicated the instrument consist of three factors i.e. mathematical logical intelligence-spatial intelligence-and linguistic intelligence. KMO value was 0.661 df 0.780 sig. 0.000 with valid category. This research succeeded to developing the assessment instrument of mathematical learning results based on multiple intelligence of second grade in elementary school with characteristics of logical intelligence of mathematics, spatial intelligence, and linguistic intelligence.

Keywords: *multiple intelligence, logical intelligence mathematics, spatial intelligence, and linguistic intelligence, assessment, mathematics*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v21i1.15286

Introduction

The results of field study show that in the three surveyed schools, they have implemented learning process with multiple intelligence approach, but in assessment of learning results, they still use conventional modeling examination. There is a gap between learning and assessment process. The results of field study in detail as follows. (1) Teacher prepares Learning Implementation Plan (RPP) with Teaching and Learning Activities (KBM) using multiple intelligence approach; (2) Teachers teach using multiple-intelligence approach; (3) Students receive learning materials through multiple-intelligence approach; (4) assessment of learning results uses traditional-assessment instrument; (5) learning results reach minimum passing grade (KKM) at least (7) there is a gap between the learning process and the instrument used to assess learning results. (8) Teachers have not been introduced to instruments which are based on multiple intelligences specifically; (9) from the Education Authority, it requires instrument use made by the Local Education Authority with using traditional instrument which is consisting only of sentences and numbers (10) the issue from students is when they have to work on traditional instruments that are not relevant to the learning process. (11) Teachers argue: traditional test has weakness if applied to the learning process with multiple-intelligence approach. Some weaknesses include (a) the instrument does not accommodate the multiple intelligence that is in accordance with the learning process, (b) The form of examination is in sentences and numbers, without any elements of multiple intelligences such as pictures, puzzle, riddle box. (C) on the instrument there is no color therefore it is less attractive.

To overcome these problems, it is offered a solution in which a model of learning-result assessment that considers multiple intelligences. This model of assessment will accommodate the characteristics of multiple intelligences that teachers use as an approach in the learning process.

In the development of assessment instrument, the mathematics lesson for second grade of elementary school is selected. Some of the reasons are mathematics is a subject that is closely related to life, mathematics is the basis to study other subjects. In certain classes students assume that mathematics lesson is tiresome.

Douglas, Smith, & Reese (2008) carried out research on impacts of multiple-intelligences learning model towards the achievement of eight-grade students on mathematics subject in Turkey. Samples consisted of two classes, one class as experimental class receiving mathematics with teaching-activity model, one class as control class to receive mathematics with Direct Instruction (DI) model. The learning results of both classes showed the difference. The experimental class had average point at 25.48, while the control class at 17.25 point. Based on this study, it is concluded that learning model with multiple intelligences increase learning result of students.

The problems in this research are: (1) What is the characteristic of assessment instrument on intelligence-based mathematics learning results? (2) What are the criteria of the assessment instrument on intelligence-based mathematics learning results? The purpose of this study is to describe: (1). Characteristics of the assessment instrument of multiple intelligence-based mathematics learning results. (2). Criteria of quality in instrument assessment of multiple intelligence-based mathematics learning results.

The purposes of the study are: The development of assessment instrument is expected to ease the teacher in implementing assessment towards the learning process with multiple intelligence approach. Teachers are expected to broaden their insights to be able to teach with high creativity whose orientation resides to the needs of students. Teachers should view students as individuals whose intelligence can develop according to the theory of multiple intelligences.

The result of the research on the development of assessment instrument of multiple intelligence is helpful for the students because the assessment is designed by the teacher that considers the multiple intelligences of the students. The assessment instrument of multiple intelligences can create a fun learning atmosphere, and students can absorb the learning materials more easily.

Students feel happy, because the content of the examination contains material related to real life. The assessment instrument of multiple intelligences is directly related to the environment and the experience of students. The assessment instrument of multiple intelligence assessment also creates a pleasant atmosphere for students, because the examination has form in colorful images, graphics, riddles, and puzzle. The way to answer is also varied, not only in multiple choices, but students can answer by coloring the images, or filling in the blank boxes in the riddle.

Armstrong (2003, pp. 2-4) outlined eight of Gardner's theoretical intelligences as follows. (1) Linguistic Intelligence: The ability to use words effectively, Logical Intelligence: the ability to use numbers well. Spatial intelligence: The ability to perceive the spatial-visual world accurately. (2) Kinesthetic Intelligence: The skill to use whole body in expressing ideas. (3) Musical Intelligence: Ability to handle musical forms, (4) Interpersonal Intelligence: Ability to perceive and differentiate mood. (5) Intra-personal Intelligence: Self-understanding (6) Naturalist Intelligence: Skill to recognize and categorize species.

(7) Logical-Mathematical Intelligence: Campbell, Campbell & Dickinson (2002, p.: 41) described some conditions that enable one's logical-mathematical intelligence to thrive well as follows: feeling their goals and functions within their environment, recognizing concepts with properties of quantity, time, and causal relationships; using abstract symbols to demonstrate in concrete way, both objects and also concepts; showing logical problem solving skills.

Armstrong (2003, p. 26) explains that a person with high mathematical logical intelligence expresses the following features: being able to calculate numbers off the head easily, being fond of mathematics, enjoying games or solving puzzles that require logical reasoning, being eager to look for patterns, regularities, or logical sequences, believing everything with rational explanation.

Someone with strong mathematical logical intelligence tends to be fond of following activities such as science, mathematics, accounting, detective work, law and computer programming. One who makes use of mathematical logical intelligence in daily life will be easier to apply mathematical concepts. When dealing with problems and in assuming or arguing, he or she will use mathematical calculations frequently. Professions that can be developed from logical mathematical intelligence include accountants, statisticians, computer programmers, scientists, and researchers.

Spatial Intelligence (Visual-Spatial)

Space intelligence (visual-spatial) is an intelligence that can be developed for students. Students with spatial intelligence have several features, including as the following: being able to read maps easily, charts, graphs, being eager to work on puzzles, can build three dimensional constructions, and being more easily to learn through images than text. Armstrong (2003, p. 48) describes a person who is strong in spatial intelligence or space usually shows an interest in color, photo or video camera, images, reading materials that have many illustrations.

In line with those disclosed by Campbell et al (2002) on visual-spatial intelligence, it is revealed that a person with spatial intelligence exhibits several observable features such as: being fond of learning by seeing and observing, being fond of thinking in pictures, reading graphs, maps, diagrams or in visual method; enjoying three-dimensional shapes, origami, composing patterns, being fond of art, cards, pictured stories, being fond of drawing and painting.

Linguistic Intelligence

Gardner (1983) revealed that language is the most important example of human intelligence that is indispensable to society. Gardner (1983) explains the important meaning of language-rhetoric aspect, or the ability to convince others from the series of actions, the potential for language recalling, or the ability to use language (Campbell et al, 2002, p. 10).

The ability possessed by a person in utilizing language intelligence can bring a student's confidence in learning to maintain a position in a forum and discussion. Utilization the ability in understanding the language in the lesson brings the opportunity to be able in discussion or teaching friends with what has been learned. In the scope of learning, a teacher must provide an opportunity for students to convey their arguments and provide opportunities for students to learn together with their friends.

A person who possesses linguistic intelligence exhibits characteristics as described (Amstrong, 2003) being fond of reading, writing, telling stories; being able to remember name, date; being fond of the word-guessing game; being fond of reading poetry, rhymes; being able to communicate well. Campbell et al (2002) describes the characteristics of linguistic intelligence such as: learning through listening, reading, writing, and discussion, effectively listening, understanding, deciphering and remembering the spoken words; being effective in writing, understanding language rules, spelling, being fond of learning other languages.

Assessment of Mathematics Learning

Assessment is the process of collecting and processing information to measure the achievement of student's learning results. Assessments that is made by teachers to students can be interpreted as a process of collecting various informations that can provide a true picture of student's learning progress. It means, if there are signs of students experiencing barriers in learning, teachers can take the right steps immediately. The taken steps in the process of handling students in the learning process can provide an overview towards the progress of learning. The learning progress of students is required throughout the learning process through assessment effort. Assessment is not only implemented at the end of the period (semester) in the learning process but during the learning process as well as the formative assessment (Stiggins & Chappuis 2011, p. 15). "The assessment is one of the main tasks of teachers ..." (Kartowagiran, 2012). Each teacher is required to possess assessment techniques to support the main task. According to Mardapi, (2007, p. 5) ".... assessment includes all the means used to assess individual performance ...". Performance appraisal allows students to demonstrate skills and attitudes that they have in addition to knowledge. Assessment in the learning process serves to determine the condition of students.

Assessment can be used as a method to motivate students in learning, not as a threat to students, in accordance with the theory presented by Nitko & Brookhart, (2007, p. 11) that:

> *Assessment may also motivate student to study. unfortunately, some teacher use this form of accountability as a weapon rather than as a constructive force. Teachers may hope that using an assessment as a possible threat will encourage their student to take studying seriously. Sometime teacher use the surprise quiz or pop quiz in this manner to encourage more frequent studying and less cramming (Nitko & Brookhart, 2007, p. 11)*

Teachers can use assessment as a way to motivate students in learning. Assessment is not intended to make students depressed, fearful or tense. Assessment if implemented properly can improve the quality of learning, an assessment supported by opinion of Nitko & Brookhart (2007):

> *How making your own assessments improve your teaching: (1) knowing how to choose or to craft quality assessments increases the quality of your teaching decisions; (2) what and how you assess communicated in a powerful way what you really value in your students learning; (3) when you carefully define assessment tasks, you are clarifying what you want*

> *students to learn; (4) you use your knowledge of how to craft quality assessment tasks when you evaluate assessment materials available from other source; (5) learning to craft assessment tasks increases your freedom to design lesson; 6) you will improve the validity of your interpretations and uses of assessment result. (Nitko. & Brookhart, 2007, p. 107)*

Theory of Nitko & Brookhart explains that assessment can improve the quality of teacher in teaching. Assessment can explain what students need. Assessment also functions in designing the next lesson. Assessment results can improve the validity of teacher interpretation towards the students. The preparation of examination items for assessment does require knowledge and high creativity, there is an influence of the assessment on improving the quality of learning.

Mardapi (2007, p. 6) described a principle to be considered in the assessment that: The essential principles of assessment are accurate, economical, and encourage the improvement on the quality of learning. Therefore, the assessment system used in each educational institution should be able to: (1) provide accurate information, (2) encourage students to learn, (3) motivate teachers, (4) improve institutional performance, and (5) improve education quality

A teacher must be able to design an assessment that fulfills the function in the learning process. The result of the assessment is expected to be helpful, both for the students and for the teachers themselves. Assessment on learning results of students at primary and secondary education levels is based on the following principles according to the Regulation of Education and Culture Minister of Indonesia No 23 Year 2016: Art 12 (2), assessment procedures undertaken by educators: Assessment of knowledge aspects is carried out through stages of: (a) preparing assessment plan ; (b) developing assessment instruments; (c) carrying out an assessment; (d) making use of the assessment results; and (e) reporting the assessment results in the form of numbers on a scale of 0-100 and description. Art 13 (1) while the process of assessing the learning process and results by educators is carried out by the following rules: (a) establishing an assessment objective with reference to the RPP that has been prepared; (b) developing an assessment points; (c) establishing an assessment instrument its guidelines; (d) conducting quality analysis of the instrument; (e) preparing assessment; (f) processing, analyzing, and interpreting the results of the assessment; (g) reporting the results of the assessment; and (h) utilizing the assessment report.

In Regulation of Education and Culture Minister of Indonesia Number 23 Year of 2016, it was clearly described how the assessment procedures should be done by the educators. An explanation of assessment both regarding assessment in the process and also the assessment in the learning results for the knowledge aspect has been detailed. Teachers only follow and carry out them. The steps to be taken to assess the knowledge aspect, from the plan to the report on the results of the assessment.

Model of Assessment in Learning Mathematics

In preparing assessment examination in mathematics, it is necessary to pay attention towards several matters related to the material, as the following opinion Schoenfeld (2002, p. 9) :

> *The "interwoven and interdependent" components of mathematics proficiency advanced by the NRC Committee are: Understanding: Comprehending mathematical concepts,...; Computing: Carrying out mathematical procedures,...; Applying: Being able to formulate problems mathematically ...; Reasoning: Using logic to explain and justify a solution to a problem ...; Engaging: Seeing mathematics as sensible, useful and doable....*

Schoenfeld explaines that the skill components in mathematics include: understanding, implementing procedures, formulating math problems, reasoning, using mathematics as a logical thing. Assessment should be designed in such a way to meet the purpose of the assessment, the mathematical assessment is able to explore what is mastered and what has not been mastered by the students from the teaching materials, therefore the teacher can plan the next lesson. According to Schoenfeld (2002, p. 94) states:

> *Designing and developing good assessment tasks, which have meaning to students and demand mathematics that is important for them. The tasks must enable students to show what they know, understand and can do without the help from teachers that classroom activities can provide. Task design is usually subject to too-tight constraints of time and form. Starting with a good mathematics problem is necessary..*

Schoenfeld's opinion lays out that teachers should be able to design good, detailed, clear assessments therefore students are able to carry it out independently. A good assessment is designed to meet the actual assessment function in education.

Learning Assessment Model with Multiple Intelligence Approach

In carrying out the assessment of mathematics subject for elementary school with Multiple Intelligence approach will be able to implement with the examination. The research that will be carried out is to developing a focus towards the three development intelligences i.e. logical, mathematical, spatial, and linguistic intelligence. Assessment instruments used in research are conducted in the form of riddle, guess words, graphics, puzzle, drawings, stories, poems. Each test item is developed by associating three multiple intelligences i.e. logical mathematics-intelligence, spatial intelligence, and linguistics intelligence. Instrument of assessment that is conducted in the study was in accordance with the model of learning pursued by students, therefore between the process and the assessment there is sustainability in approach.

The tests in the development of instruments of assessment, it is used to assess mathematics learning for elementary school. The developed instrument aims to provide a sense of pleasure in students as the tests run, reducing students' anxiety during the test, while being in test students feel happy, and being able to relate test materials with real life.

Mathematical appraisal with mathematical logical approach, spatial intelligence and linguistic intelligence will be carried out by presenting the spatial drawings to the students in which in them, there are various images, such as ball, can, triangle, rectangle etc. Students are required to observe the pictures to answer the questions. The test items are in the form of puzzle inside, the students look for the empty puzzle pair. Furthermore, the students are presented with natural picture, in which there are animals, plants in it, students are asked to count the number of animals, plants present in the picture. Further, students are presented with pictures of unit-hundred-thousand blocks. Students are required to count the number. The test is in the form of story, students are required to observe the number of goods, and the price of goods.

The result of the Ellis's (2011) research "highlight generalization as a dynamic, socially situated process that can evolve through collaborative acts". The conclusion of the research is that the dynamic learning process and social process can be improved through collaborative action. Amy's research illustrates that the role of the learning process in mathematics affects the learning outcomes greatly. It is necessary for teachers to develop innovative and creative learning. Learning can be created by the approach of multiple intelligences.

According to the research of Duskri, Kumaidi, & Suryanto (2014) that the learning process can be effective and successful when individual differences get attention. The difference will affect the level of understanding of students. The teacher must know the individual differences that has form in students' difficulties in understanding the subject matter, the factors that cause difficulties and other factors. As a result, the diagnostic test is a solution therefore the teacher can design the learning process in accordance with the needs of students.

Method of Research

The research that was used is the research and development, According to Borg & Gall (1983, pp. 771-794) research & development was conducted in ten stages as follows: (1) Research and Information Collection. The first step relates with preliminary study towards the product development plan. (2) Planning. The second step, after the preliminary study, it is carried out planning preparation. In the planning of this study, it includes: identification of competency standards and basic competencies of mathematics for elementary school of grade V, demanding the prerequisite materials (learning continum), preparing concept maps, preparing material clues, defining the objectives, determining the steps of the development activity, determining the place, time, research sample and required funding, determining the experts that are involved in the FGD or expert judgment, and determining the product trial samples on small scale.

(3) Developing Preliminary Form of Product. The third step is to developing the initial shape of the product. (4) Preliminary Field Testing. In this fourth step it is carried out to test the product design on a limited basis. (5) Main Product Revision. The main product revision step aims to improve product design based on limited trials. Based on information and inputs from experts on the initial test, improvement was made towards the developed product. (6) Main Field Testing. Products that have been refined, tested more widely to potential users. Based on the second test, it will be obtained empirical information, whether the developed product has met the empirical validity or not, both in terms of substance and also of effectiveness of the product. (7) Operational Product Revision. Based on field trials more broadly and based on empirical results obtained, an improvement to the product developed. Product improvement is a second improvement. Unqualified items must be discarded or repaired.

(8) Operational Field Testing. Based on the second improvement of the developed product, it is followed with the feasibility test on the user more extensively than the previous field test. From the feasibility test, it will be obtained information both in terms of substance and also methodology towards the product design developed to be applied in the field. (9) Final Product Revision. Based on the information obtained during the feasibility trial, it is continued with a revision to complete the resulting product. (10) Dissemination and Implementation. After implementing improvements to the resulting product, the final step is dissemination and introduce product results either through workshops, scientific meetings or in the form of scientific journals. The findings of the product can be used or implemented by both the teachers in the field and for the concerned parties to advance the education world.

In conducting the research and development of the assessment instrument of multiple intelligence-learning results of mathematics was started from February 2014 to March 2017. The places of study were: SD Mutiara Ilmu pandaan Pasuruan East Java, SD YIMMI Gresik East Java, SD Muhammadiyah I Samarinda, SD Muhammadiyah IV Samarinda East Kalimantan. Preliminary study for Grade-2 students was 11 classes, teachers for grade 2 were 17 people for Small Trial: Grade-2 students were 89 people with teachers of 7 people. The trial was expanded: Grade-2 student were 200 people, teachers amounted to 9 people.

Preliminary research aimed to deepen about the learning process by applying the approach of multiple intelligences. In Indonesia, the application of multiple intelligences in learning was driven by Chatib (2009). Schools that have conducted the learning process with multiple intelligences of Howard Garner are in SD Plus Mutiara Ilmu in Pandaan Pasuruan East Java. Gresik Sekolah YIMI (Malik Ibrahim Islamic Foundation) that is located on Jalan JA Suprapto and SD Muhammadiyah 1 Samarinda.

Model Development: (1) Preparation of characteristics of multiple intelligence (2) Preparation of Curriculum-2013 clues (3)

Writing examination items. (4) Instrument seminar (5) FGD to obtain assessment of experts towards the examination items that have been already in the form of test equipment. (6) User Validation.

Small-scale trial with subject of 89 grade-2 students of SD Muhammadiyah IV Samarinda. Based on the results of experiments, it was conducted item analysis using computer assistance such as EFA with SPSS, then the next item would be revised in accordance with the results of computer analysis. After the test items were revised, the second test would be conducted with larger number of sample. The second test subject was also the same as the first trial, SD Muhammadiyah I Samarinda with total of 200 students.

From the result of second test, then it was reanalyzed, further revision was made to the test items that needed to be revised. The next step of the implementation phase. In the implementation phase of the instrument product, it could be applied to the actual situation. The results of the implementation of this instrument were analyzed to identify the achievement of the learning process for one semester with multiple intelligences approach. Revised point: from the results of analysis of instrument implementation, then the final revision was made.

For the quantitative data in the form of mathematics-learning achievement can be reviewed from the developed assessment instruments. For qualitative data, it was used interview/observation/questionnaire. Descriptive ana-lysis in research is used to describe stages of development and application of assessment instruments and the results as well. Descriptive analysis to illustrate the quality of assessment models ranging from early prototypes, seminar of prototype 1, FGDs, small-scale trials, expanded trials, to model trials. Quantitative analysis was performed on the data to determine the validity, model test and reliability of the multiple intelligence assessment model based on the empirical data obtained in the field.

Content validity is a test to examine validity of instrument items. Content validity test aims to determine if the item has already included the material to be measured. If the item is in accordance with the prepared indicators. Validity test is carried out by experts (expert judgement), as well as by the user (teacher). Results of validity test of expert and user were analyzed using aiken analysis with formula:

$$V = \frac{\sum s}{n(c-1)}$$

The value criteria of V aiken less than 0.600 is included in less good category, between 0.600 - 0.88 is included in good category, while greater V than 0.800 is included in very good category.

Exploratory Factor Analysis: factor analysis is carried out to determine the variable-forming factors. A common criterion of EFA is the KMO MSA value> of 0.5. The sig. value i.e. <0.5. Loading factor >0.3, eigen value >1. Qualitative analysis, qualitative analysis is carried out on the readability. Assessment regards to the use of language, writing techniques, the use of punctuation, the use of fonts, the use of pictures, the length of the sentence. Assessment technique with questionnaire instrument, containing statement with four answer choices, 4 for excellent, 3 for good, 2 for less good, 1 for not good.

Results of Research and Development

Instrument Development

Instruments in the form of initial drafts should be validated. The purpose of validation is to obtaining feedback, criticism, suggestions on model improvement according to the area of expertise of each validator. Expert validation aims to provide an assessment to the items in the instrument.

Assessment relates to the point suitability towards indicator, the suitability of point to the psychology of primary school children, the suitability of point towards mathematics for elementary school, graphs, suitability of choice answers. Other things are assessed to the type and size of the letters, the number of words.

Table 1. Recapitulation on Results of Aiken Instrument Validity Instrument by Expert

Validator (initial)	Area of Expertise	Valid item	V Aiken
FH	Psychology		0.714–0.952
DM	Measurement		
BK	Evaluation		
YA	Psychology		
JA	Mathematics		
FAS	Psychology		
MAR	Phil. of maths		

Table 1 shows that the analysis results of instruments provided by the experts indicate that all items are eligibly valid. The resulting V aiken values are between 0.714 and 0.952.

Validity of Instrument Item of Assessment of Multiple Intelligence

Instrument items were validated by 7 experts as well as by 9 teachers, the data was analyzed with the V Aiken formula. Here is the result of Aiken's validity on each developed multiple intelligence.

Validity of Logical Mathematical Intelligence Item

Table 2. Results of Validation of LM Intelligence Item

Instrument Item	V Aiken (Expert)	V Aiken (Teacher)
1	0,857	0,833
2	0,762	0,866
3	0,905	0,833
4	0,810	0,733
5	0,762	0,833
6	0,762	0,866
7	0,714	0,866
8	0,810	0,766
9	0,762	0,866
10	0,857	0,766
11	0,952	0,833

Table 2 shows that the validation results of experts as well as of teachers towards the items of logical mathematical intelligence can be concluded that the whole instrument items of mathematical logical intelligence has good-level validity. 13 items that represent logical intelligence of mathematics have expert-validity results ranging from 0.619 to 0.952. The validation results by teachers as practitioners ranging between 0.733 to 0.866. Based on the achievement of V Aiken result value, all items are declared valid and are in medium and good category.

Validity of Spatial Intelligence Items

From Table 3 shows that 12 points for spatial intelligence have good validity, both validation results of expert and also teacher. This conclusion is based on the value of Aiken Validity, in which the results of V aiken by experts are between 0.714 to 0.952. The validation results of teacher are between 0.733 and 0.90

Table 3. Validation Results of Spatial Intelligence Items

Instrument Item	V Aiken (Expert)	V Aiken (Teacher)
1	0,810	0,733
2	0,857	0,833
3	0,952	0,900
4	0,857	0,833
5	0,762	0,800
6	0,810	0,866
7	0,810	0,833
8	0,714	0,800
9	0,905	0,833
10	0,667	0,833
11	0,714	0,866
12	0,857	0,833
13	0,857	0,833
14	0,857	0,833

Source: Excel, Analysis of V Aiken

Table 3 shows that the validation results of experts as well as of teachers towards the items of logical mathematical intelligence can be concluded that the whole instrument items of mathematical logical intelligence has good-level validity. 14 items that represent logical intelligence of mathematics have expert-validity results ranging from 0.714 to 0.952. The validation results by teachers as practitioners ranging between 0.733 to 0.900. Based on the achievement of V Aiken result value, all items are declared valid and are in medium and good category.

Validity of Items of Linguistic Intelligence

Table 4. Validation Results of Linguistic Intelligence Items

Instrument Item	V Aiken (Expert)	V Aiken (Teacher)
1	0,762	0,733
2	0,857	0,833
3	0,714	0,900
4	0,762	0,833
5	0,857	0,833
6	0,667	0,866
7	0,714	0,833
8	0,762	0,733
9	0,810	0,800
10	0,762	0,900
11	0,619	0,866
12	0,762	0,733

Source: Excel, Analysis of V Aiken

From Table 4 it shows the validation results of expert as well as teacher after being analyzed with V Aiken, it can be concluded that the whole items of linguistic intelligence have good validity, therefore it is worthy to use. This conclusion is supported by Aiken Validation result data by experts between 0.667 to 0.810 while validation by teachers were between 0.733 to 0.90.

Small-Scale Trial

The results of small-scale trials were analyzed by EFA (exploratory factor analysis.) The purpose of EFA was to investigate the factors contained in the observational variables. All measurable variables were associated with each factor in an estimation of loading factor. In this EFA analysis, we want to learn the items that are included in the developed factors i.e. factors of mathematical logical intelligence, spatial intelligence, and linguistic intelligence.

Table 5. KMO and Bartlett's Tes

KMO	0.661
Df	0,780
Sig.	0,000

Source: Output of SPSS

Based on the analysis of SPSS, it shows that KMO MSA (Kaiser-Meyer Olkin Measure of Adecuasy sampling) has a value of 0.661. The value of KMO MSA is included in good category for further analysis, since it is greater than 0.5.

Table 6. Analysis of Total Variance Explained

Componen	Initial Eigenvalues		
	Total	% of Variance	Cumulative %
1	13.342	33.356	33.356
2	10.300	25.749	59.105
3	8.103	20.257	79.362

Source: Output SPSS Total Variance Explained

Table 6 shows that 40 variables that were analyzed consist of three factors. This is observed from the value of eigenvalue that is located ≥ 1with a cumulative value of 79.362%. Three factors are formed in accordance to the theory developed in the assessment instrument of multiple intelligences.

Conclusions

From the results of data analysis can be concluded as follows. Fisrt, Instrument of assessment of learning results in mathematics is based on multiple intelligences has forms of: a. each item of instrument accommodate the characteristics of multiple intelligences. b. Instrument has form of drawings, graphics, number squares, puzzles, poems, short stories, tables, and color maps. c. The answer options are on the number square, on the animal image, on the fruit drawing, and on the puzzle image. Second, The multiple-intelligences based criteria of assessment instrument of mathematics has good validity. It has a moderate and good difficulty level of items.

Suggestion

Some suggestions need to be delivered for further refinement and development towards the results of research and development of multiple intelligence-based mathematics learning assessment instrument: (1) Teachers may develop a multiple-intelligence-based learning appraisal instrument on other types of multiple intelligences, or other subjects; (2) Researchers and teachers can develop this multiple intelligence assessment instrument into a computer-based test assessment with various softwares; (3) The education authority can develop an assessment research of multiple intelligence in junior or senior high school level

References

Ellis, A. B. (2011) Generalizing promoting actions: how classroom collaborations an support students' mathematical generalizations. *Journal for research in Mathematics Education (JRME), 42*(4).

Armstrong, T. (2003). *Sekolah para juara: menerapkan multiple intelligences di dunia.* Bandung: Kaifa.

Borg, W. R., & Gall, M. D. (1983). *Educational research: An introduction.* London: Longman Publishing.

Campbell, L. & Campbell, B., & Dickinson, D. (2002*). Teaching & learning through multiple Intelligences (metode terbaru melesatkan kecerdasan).* Bandung: Inisiasi Press.

Chatib, M. (2009). *Sekolahnya manusia.* Bandung: Kaifa.

Duskri, Kumaidi, & Suryanto. (2014) Pengembangan tes diagnostik kesulitan Belajar matematika SD. *Jurnal Penelitian dan Evaluasi Pendidikan, 18*(1).

Douglas, O., Smith, K., & Reese, N. (2008). The effects of the multiple intelligences teaching strategy on the academic achievement of eighth grade math students. *Journal of Intructional Psychology*, *35*(2), 182.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Kartowagiran, B. (2012) *Penulisan butir soal.* Yogyakarta: UNY

Mardapi, D. (2007). *Teknik penyusunan instrumen tes dan non tes.* Yogyakarta: Mitra Cendikia Press.

Menteri Pendidikan dan Kebudayaan. Peraturan Menteri Pendidikan dan Kebudayaan Republik Indonesia Nomor 23 Tahun 2016 tentang Standar Penilaian Pendidikan (2016).

Nitko, A. J., & Brookhart, S. M. (2007). *Educational assessment of student* (6th ed.). New York: Pearson Merrill Prentice Hall.

Schoenfeld, A. H. (2002). Making mathematics work for all children: Issues of standards, testing and equity. *Educational Researcher*, *31*(1).

Stiggins, R.J., & Chappuis, J. (2011). *An Introduction to student –involved assessment for learning.* California: Pearson.

Jurnal Penelitian dan Evaluasi Pendidikan
Volume 21, No 1, June 2017 (104-115)
Online: http://journal.uny.ac.id/index.php/jpep

THE EVALUATION OF AUTHENTIC ASSESSMENT IMPLEMENTATION OF CURRICULUM 2013 IN ELEMENTARY SCHOOL

Muhammad Nur Wangid[1]*, *Ali Mustadi*[1], *Anwar Senen*[1], *Nur Luthfi Rizqa Herianingtyas*[1]
[1]Universitas Negeri Yogyakarta
[1]Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
* Corresponding Author. Email: m_nurwangid@uny.ac.id

Abstract

This research was aimed to evaluate the implementation of authentic assessment of elementary school in Province of Yogyakarta and also to know the obstacles of its implementation. This was an evaluative research by a Stake's evaluation model approach. The results of observation in compare to the standard of assessment should be criteria to determine the succeed. This research subjects were elementary teachers in Province of Yogyakarta. Observation, interview, and documentation were used to gather data. The research showed that: (1) Planning (antecedents) stage or understanding towards authentic assessment planning has not been fulfilled the standard to be categorized as Good with percentage of 68.75%; (2) Process (transaction) or implementation stage that obtained 63.41% in percentage was classified in Good category; (3) Outcome stage or authentic assessment report showed 68.48% in percentage and should be categorized Good. The implementation of authentic assessment in Province of Yogyakarta elementary schools have not 100% met the standard. Therefore, results from this research finding were expected to be tools to improve performance from all stakeholders.

Keywords: *authentic assessment evaluation, Curriculum of 2013*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v21i1.15779

Introduction

Curriculum of 2013 comes with a new color in the world of education in Indonesia. Its dynamic and revolutionarily character make a fundamental transformation in the practices, for instance is an emergence of the paradigm of authentic assessment as the basis of assessment in teaching practice. The assignment is based on the reflection of the humanist that represents the result of human learning is not only in terms of cognition, but also there are some sides which involved in it; which are affective and psychomotor, therefore, in regards to see learning outcomes, student cannot be seen only by single variable, thus education should be able to become a place as well as a tool that can reflect the ability of each individual to learn from various authentic sides.

Authentic assessment in the Curriculum of 2013 refers to regulation of Ministry of Culture and Education No. 66 Yeaer 2013 on the Standards for Educational Assessment (Mendikbud RI, 2013), which is a process of collecting, reporting and use of information about student learning outcomes by applying the principles of assessment, implementation of sustainable, authentic evidences, accurate, and consistent as public accountability. The aim is to plan an assessment that suitable with the achieveable competence and based on the principles of assessment, professional, open, educative, effective, efficient, and in line with the social and cultural context; and reported the results in an objective, accountable, and informative way.

The term authentic derived from synonyms of original, real or true so that authentic assessment is often described as an assessment of the fact of the development of learners that viewed from different sides and competences, which in this case includes three domains of learning process which are manner, skills, and knowledge. Hein (1991, p. 116) gives a perspective that, "through authentic assessment we can assess students in a variety of ways: we can observe what they do, listen to what they say, read what they write, and analyze what they produce. Any behavior that can be perceived can be adapted for assessment". A teacher can find out the ability of students through a variety of assessment strategies as well as reflects the learning process that has been implemented. Fook & Sidhu (2010, p. 154) imply that "authentic assessment emphasizes the practical application of tasks in real-world settings". It means that an authentic assessment is not only measure students' theoretical ability but rather the application of their skill and manner. Thus the student achievement is represented by the ability to practice not only memorizing study materials. Mueller (2005, p. 5) defines, "authentic assessments as direct measure of students' acquired knowledge and skills throughformal education to perform authentic tasks. Therealistic contexts can make problems more engaging forstudents and help the teachers evaluate whether astudent who can solve a problem in one context cantransfer the skills to a similar setting." Authentic assessment is able to directly measure how far the knowledge, skills and manner can be applied by students in solving practical problems in their daily life. Furthermore, Palm (2008) describes that "authentic assessment is what claimed in or by the task or assessment is really true. The fact that something is supposed to be true, however, gives the concept different meanings depending on the chosen frame of reference. The meaning of the word authentic makes the choice of focus an open question, and different foci have also been applied in the literature. Two main issues are of interest here: what it is that is supposed to be real or true, and what it is that it is supposed to be true to." It can be interpreted that authentic assessment is the correct assessment and in accordance with the facts. The meaning of word authentic lead us to the open, relevant, and right questions. Wiggins (1989, p. 41) describes four characteristics of authentic assessment among others: (1) The task should be representative of performance in the field, (2) Attention should be paid to teaching and learning the criteria for assessment, (3) Self-assessment should play a great role, (4) When possible, students should present their work publicly and defend it. Thus, representatively this kind of assessment assesses the process and outcome of students learning process. Additionally, in a separate article, Wiggins (1989a, p. 711) emphasized the importance of contextual student homework as a part of the authentic assessment such as creating a report paper to enforce student to collaborate with their friends.

Ministry of Culture and Education (2014) in this context mentioned that authentic assessment is a significant meaningful measurement to the student's result of study in terms of manner, skills, and knowledge. It is known that authentic assessment may reflect aspects of affective, psychomotor, and cognitive that found in the students' learning activities. Thus, authentic assessment requires students to show manner, using the knowledge and skills gained from learning in the conduct of the actual situation. Authentic assessment carried out com-prehensively to assess the input, process, and output of learning, including manner, knowledge, and skills. Authentic assessment is aimed to measure student's knowledge, skills and attitudes in a valid and concrete way. To assess those three aspects, there are various types of assessment that can be done, Ministry of Culture and Education (2014) mentioned the types of authentic assessment as follows: (1) performance evaluation consist of: log and learning journal; structured assignment; task performance; long-term projects; portfolio; demonstration; experiment; presentation; and simulation, (2) project assessment, (3) portfolio (4) written assessment, and (5) attitude. Forms of authentic assessment to assess student's manner competence, such as: (1) oservation; (2) self-assessment; (3) peer-assessment; and (4) journal assessment, meanwhile the aspects of assessment for knowledge competencies include: (1) written test; (2) discussion observation—question and answer, and conversation; and (3) assignments. The form of assessment for skill competence are: (1) performance assessment; (2) project assessment; (3) product assessment; and (4) portfolio assessment.

As an integral part of learning process, it is important to design and apply an assessment systematically, this refers to the necessity of teachers to know the mechanisms, procedures and instruments of student learning outcomes assessment in accordance with the competencies to be measured. Competence as measured through authentic assessment describes the demands that exist in the Standard of Competence or Core Competence, and the Basic Competencies. The focus of assessment in Curriculum of 2013 was the success of students in achieving competency standards as specified, including manner, skills, and knowledge. Therefore, teachers must pay attention to the attainment of students to ensure they are measured empirically by the standard and achieve the purpose of learning. Authentic assessment is also need to give a picture about student's development right after the learning process, this refers to the necessity of teachers to do perpetual and comprehend assessment, it means that teachers should understand the development of their student everyday, authentic assessment is not only seeing the result but also the process behind that to be a consideration in the assessment process itself.

Teachers in this case play a role as an evaluator of learning, have the ideal capacity and are able to understand and implement authentic assessment in a professional manner in accordance with what required in Curriculum of 2013. But in fact, since second year of Curriculum 2013, the implementation of authentic assessment still be one of the main evaluation issue due to the obstacles in its implementation, it implies by the results of a questionnaire given to 100 teachers grade 5 in Yogyakarta, the result shows that 100% of the respondents think there are obstacles in implementing authentic assessment. In addition, a survey conducted by Khilmiyah, Sumarno, & Zuchdi (2015, p. 2) found that 70% of intelligence and judgment that has been developed in elementary school around Yogyakarta is through cognitive assessment. Teacher's attention and understanding to the importance of development of emotional, social, and spiritual intelligence is still low in the learning and assessment process. Assessment of the student's character is only based on teacher's observations during school day. It means that authentic assessment that includes cognitive, affective, and psychomotor aspects has not been applied optimally. In regards to that problem, to know more about the obstacles and problems we need to evaluate the implementation of authentic assessment, especially to elementary teachers in Province of Yogyakarta. Arikunto (2013, p. 325) argues that the evaluation of a program is a series of activities and it is purposes to see how far the program is succeed. According to Sukmadinata (2009, p. 172), evaluation and curriculum have a causal relationship where a change in the curriculum will influence the curriculum evaluation and curriculum evaluation would otherwise change the aspects of curriculum implementation. The evaluation is purposes to find out the lacks so that it can be a consideration in the future implementation of education system. The results of evaluation curriculum can be used by teachers, principals and other educational stakeholders. In line with the aforementioned opinion, Doll (1964, p. 22) says that "Acknowledge presence of value and valuing, orientation to goal, comprehensiveness, continuity, diagnostic worth and validity and integration." An evaluation and assessment should contain value and assessment, has a clear purpose or objective, comprehensive and perpetual, has a function as diagnostic tool, and integrated. Hence, the evaluation was not carried out randomly, but systematic, detailed and use procedures that have been tested thoroughly. Furthermore, Alkin (2011, p. 10) also states that "a definition of evaluation based on its goal. Evaluation is the favored term when we talk of judging a program," it implies that evaluation is an activity to gather information about how the program runs, which is used to determine future action to be done.

Evaluation model used in this study is a stake model, which is one of the educational system of evaluation models. Stake emphasizes on implementation of the two main aspects, namely (1) description and (2) consideration (judgments), and distinguishes three stages in the evaluation of the program, which are (1) antecendents/context, (2) transaction/process, and (3) output-outcomes. According to that, this study aimed to find out the implementation of authentic assessment in elementary school in Province of Yogyakarta and also to know about the obstacles of its implementation.

Methods

This is an evaluative research that aimed to evaluate the implementation of authentic assessment of Curriculum 2013 in elementary school of Province of Yogyakarta. The result of evaluation is expected to be a base or foundation to measure the achievement of curriculum assessment in 2013 and also to give judgement or recommendation in regards to improve the quality of its implementation. Evaluation approach that be used in this study is Stake's Countenance Model that measure the implementation of authentic assessment with the standard of that should be a criteria to determine the succeed. Stake model design's evaluation emphasized two major aspects which are description and judgement. Those two major aspects is distinguished into three stages of evaluation, as follow: (1) input and planning (antecedent), (2) process (transaction), (3) outcomes. Stake model is a systematic method to evaluate the overall implementation of authentic assessment process that includes planning, implementation, and assessment. We choose Stake model of evaluation with the consideration that we would like to conduct and focus on evaluation of the implementation of authentic assessment. Evaluation design with this model has several stages, such as: (a) antecedent phase that describes teacher's experience on the design of authentic assessment (knowledge, skills, and manner), (b) transaction phase, describes implementation of authentic assessment, (c) output phase, describes assessment's result management. Furthermore, the authors made the judgement pertaining to stakeholder's understanding on authentic assessment implementation in elementary school in Province of Yogyakarta. Our decision were based on two thing: (1) absolute standard, which explain on the existing process, and (2) relative standard that put the base on standard or criteria which is in line with authentic assessment program. The authors will relate its relevance to the congruence between what is intended and observed. We observed 63 elementary schools that has already applied Curriculum of 2013 in Province of Yogyakarta since August 2016 and we finish this study in October 2016. The study's populations are teachers. To determine the sample we use an equation from Issac and Michael (Sukardi, 2011, p. 55) as follow:

$$S = \frac{X^2.N.P\,(1-P)}{d^2\,(N-1) + X^2\,P\,(1-P)}$$

Information:
S = Number of sample
N = Total Populations
P = The proportion of population as a basic assumption of table creation. This price is taken at P = 0.50.
d = Degree of accuracy reflected by an error that can be tolerated in the P sample fluctuation, d is generally taken at 0.05.
X2 = Value of chi-square table for one degree of relative liberation desired confidence level. X2 = 3.841 confidence level of 0.95.

The results of sample enumeration from 60 fourth grade teacher with a error sampling of 5% obtained a sample of 18 fourth grade of elementary school teachers. However, in this study we used a sample of 20 teachers. This is a descriptive evaluation study which aims to provide an overview of reality in the implementation of authentic assessment in Curriculum 2013 by having a theoretical concept that has been developed towards aspects that will be evaluated. All the data were analyzed using percentage analysis techniques then described and drawn conclusions about each of the components on the basis of predetermined criteria. The amount of the percentage in each category shows the information disclosed directly and the position of each aspect of whole or parts of the problems studied can be seen. Data were analyzed using an inter-active model of (Miles, Huberman, & Saldana (2014, p. 12), which consists of data collection, display, and conclusion. While quantitative data on the results of observations were analyzed using percentage analysis.

Results and Discussion

Result

Study of evaluation in implementation of authentic assessment in elementary school in Province of Yogyakarta conducted based on description and judgement principles. Both are obtained through the depiction of the preliminary stage (antecedent), stage of the process (transaction), and the results (outcomes). Data obtained in this study can be divided into three parts, namely the input and planning (Antecedent phase), process/implementation (Transaction phase), and results/evaluation (Outcomes phase). At each stages will be compliance (horizontally) between planning (intents) and the data obtained from the results of the implementation of the observations (observations). If there is a discrepancy, then there will be consideration/suggestions/feedback on the implementation of authentic assessment that suitable with the real condition. Furthermore, we will monitor if there is discrepancy at each stages. Further analysis to look for suitability (congruence) between the expected implementation of authentic assessment (intended) in accordance with the standards observed at each stages. The results are shown in Table 1.

Table 1. Data Evaluation Authentic Assessment

Stage	Aspect	Percentage	Category	Judgment Matrix	
				Standard	Judgments
Input (Atencedents)	Manner	63.99%	B	100%	Yet need to be given appropriate consideration
	Skills	71.06%	B	100%	Yet need to be given appropriate consideration
	Knowledge	71.25%	B	100%	Yet need to be given appropriate consideration
Process (*Transaction)*	Manner	43.42%	C	100%	Yet need to be given appropriate consideration
	Skills	61.88%	B	100%	Yet need to be given appropriate consideration
	Knowledge	84.96%	B	100%	Yet need to be given appropriate consideration
result *utcome)*	Manner	54.76%	C	100%	Yet need to be given appropriate consideration
	Skills	70.97%	B	100%	Yet need to be given appropriate consideration
	Knowledge	80.08%	B	100%	Yet need to be given appropriate consideration

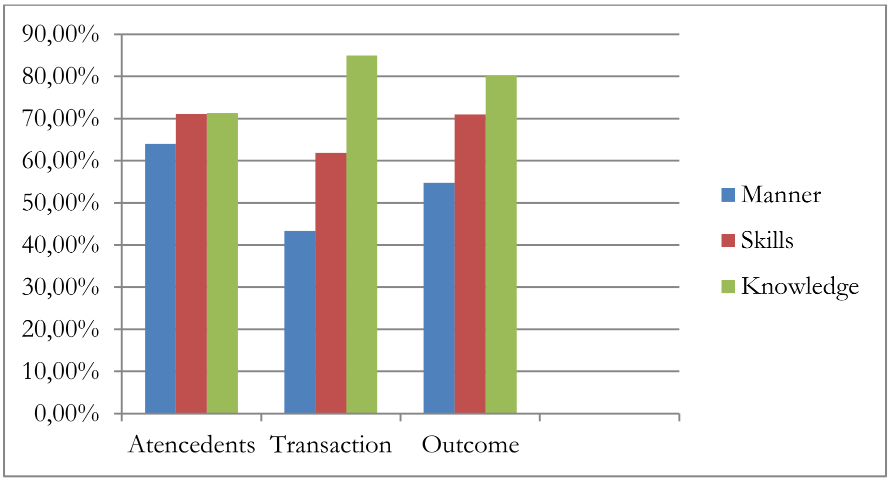


Figure 1. Implementation of Authentic Assessment at Every Stages

At the input or planning stages, known that assesses on the manner are still not optimal in the amount of 63.99%, which is only categorized in Good level, the results imply that the assessing plan on the manner do not meet with the standard required. Parameter of skills is also do not implemented optimally with 71.06% of percentage and placed in Good category. Meanwhile the knowledge parameter have 71.25% in percentage and do also categorized as Good and still need some considerations. At this stage of the process or implementation, the lowest percentage is 43.42% and categorized as Enough, it indicates that most of teachers do not apply proper manner assessment in learning process, in addition to the implementation of the skills assessment is also only reached 61.88% of percentage and still need to get feedbacks and improvements. Aspects of knowledge get the highest percentage compared with other aspects which reached 84.96%, it means the teacher has been able to carry out an assessment of knowledge well although not optimal. In the output stage, known that manner assessment is still very low which is only at the 54.76% of percentage, skills in 70.97%, and knowledge assessment is in 80.08%. Those aspects still need further considerations.

Implementation of authentic assessment can be seen from diagram in Figure 1.

Furthermore, the results of these evaluations are systematically processed with Stake model evaluation (Stake's Countenance Model), which measures the enforceability of authentic assessment on the Curriculum 2013. Stake's model evaluation design emphasizes the implementation of the two major things which are description and consideration of the decisions (judgments). The following chart is presented according to the Stake's Countenance Model in the Table 2.

From that chart we can analyze vertically the antecedent stage (planning), transaction (process), and outcomes (results). At the expected conditions, the third stage has a percentage of 100%, means that there are no gaps between the three stages. But in actual conditions (observed), there is a gap between the percentage of the third stage. There is a gap between antecedent (preliminary) and transaction (process). Antecedent transaction amounted to 68.75% and amounted to 63.41%. From these two stages, there was a decrease of 5.34%. That is at the planning stage teachers already planed well although there are still shortcomings in the implementation, meaning that not all of the plan can be implemented by teachers appropriately. Gaps also occured between transaction (process) and outcomes (results). Transaction (process) have 63.41% of percentage, while outcomes (results) are in 68.48%. The gap at second stage is an increase of 4.07%. This means that although the process of implementation of authentic assessment is not optimal but teachers are already understand how to make a report on authentic assessment. However, it still need a lot of feedbacks and considerations.

Table 2. Results of Evaluation Model Stake

The Expected Conditions			Actual state (observed)
Antecedent Understanding of the design of authentic assessment	100%	Conformity ⟷	68.75%
Discrepancy			Discrepancy
Transaction Implementation of authentic assessment	100%	Conformity ⟷	63.41%
Discrepancy			Discrepancy
Outcomes Reporting the results of an authentic assessment	100%	Conformity ⟷	68.48%

Discussion

According to the Table 1 and 2 we can see that overall the assessment has been well conducted but there is still need for some improvement in terms of its implementation. In refers to Regulation of Ministry of Culture and Education No 66 Year 2013 and The Model of Assessment in Student Competencies Achievement, the standard of authentic assessment are as follow: (1) the framework of assessment in syllabus, consists of assessment techniques and the range period for each main study materials; (2) the framework of assessment in lesson plan, consists of assessment technique, instrument, and materials; (3) developing the indicator of competency, manner, skills, and knowledge achievement; (4) manner indicators are refers to major competencies 3; (5) the indicators are formulated using the operational verbs; (6) the indicators are customized according to the related major competencies; (7) determine various assessment technique according to the manner competencies (journal, peer-assessment, self-assessment, observation), skills competencies (practice, project, and portfolio) and knowledge competencies (written test, oral test, and assignment); (8) create assessment instrument according to the assessment technique that contain achievement indicator, assessment rubric, and assessment criteria; (9) determine the scoring guidelines that contain how to do scoring and process the scores into the final score; (10) determine assessment rubric that contain guidelines/description on the scale assessment; (11) determine assessment criteria that contain achievement in the form of predicate. From those standard of assessment we can conclude that teachers have not optimally implement this assessment. In fact they are not capable enough to create optimal assessment since there are problems, such as: (1) Cannot formulate good operational verbs assessment indicators; and (2) Cannot understand how to formulate achievement indicators, assessment rubric, scoring, and assessment criteria.

Based on these constraints, some inputs that can be given to teachers are: (1) Teachers should be able to improve their knowledge to understand the use of the operational verb, formulate indicators of achievement, scoring, make assessment criteria, as well as more creative in preparing the assessment rubric; (2) Teachers should design assessment rubric at beginning of each semester, and ensure that every main competencies are already have assessment rubric; (3) Teacher should be able to do an effective assessment based on the time allocation and learning objectives; (4) Teacher should learn from other teachers to get feedbacks for their assessment plan. In regards to that assessment aspects, Jones (2005, p. 7) argued that “Planning is an essential part of a teacher’s workload. Teachers need to plan and create opportunities within each session for both the learner and the teacher to obtain information about a learner’s progress towards the learning goals defined by the teacher at the start of the session. It is crucial that the learning goals are communicated to the learner, and of equal importance is that the teacher checks to ensure that the learner not only understands the learning goals, but also appreciates the assessment criteria which will be used to assess the work.” Assessment criteria is an important part of authentic assessment so the assessment of the student can be valid according to their capability to learn, besides assessment criteria can also reduce the subjectivity in terms of scoring, as what Wiggins (1989a, p. 711) said that “Scoring must be complex and authentic test cannot be scored on a curve, but instead are criterion-referenced, based on standards. As with formative assessment, self-assessment is central.” Therefore as a teacher they need to create assessment criteria as a guidelines to assess student’s achievement and process of learning.

Evaluation in this planning phase are oriented to teacher’s capability to plan and prepare the manner, skills, and knowledge assessment that will be conducted at the learning process, since the authentic assessment need to be done programmatically and systematically, the implementation need to be prepared with clear and precise measurement. In authentic assessment plan, assessment criterion and the way we process the score should be thoroughly understood by the teachers. Furthermore, according to Regulation of Ministry of Culture and Education Year 2013 and The Model of Assessment in Student Competencies Achievement, the standard implementations of authentic assessment are as follow: (1) informing manner, skills, and knowledge competency that will be assess-ed; (2) informing techniques that will be applied; (3) informing assessment rubric and criteria; (4) conducting integrated manner, skills, knowledge assessment; (5) Using techniques and instruments that has been planned; (6) Conducting conducive, quite, and comfortable assessment.

One of the authentic assessment principles is to measure the competence of students in various ways and sources. That can be a source of assessment process and products Suarta, Hardika, Sanjaya, & Arjana (2015, p. 48). Therefore, the implementation of authentic assessment in this regard depends on the ability of teachers to implement assessment methods and tools to assess the manner, skills, and knowledge of students in accordance with the competence to be achieved. The precision of the use of appropriate methods and instrument will be able to describe the actual student's competence. From that implementation standard, teachers are known have not apply the instrument properly, it can be seen from the obstacles experienced by teachers. In fact, most of teachers does not: (1) Inform their student about assessment techniques; (2) Inform assessment rubric and criteria; (3) Use planned technique and instrument, for instance teachers do not use assessment rubric as what cited in lesson plan; (4) Using all of instrument as cited in lesson plan; and (5) Most of teachers are assessing knowledge rather than other competencies.

From those problems we can suggest some feedbacks, such as: (1) Teachers should be committed to carry out the evaluation as planned and the assessment carried out consistently; (2) Teacher should carry out a thorough assessment by seeking involvement of social aspects of spiritual, attitudes, skills and knowledge, as well as the implementation of the continuous assessment to determine the development of students' abilities and carry out the follow-up of their development; (3) Teacher should inform students that assessment will be implemented in lesson; (4) Teacher should carry out an assessment on know-ledge, skills, and attitudes competencies in a balanced, coherent and comprehensive ways. The findings coincide with what presented by Sadler (2005) that "the success of any assessment is depending on the effective selection and use of appropriate procedures as well as on the proper interpretation of student's performance. Thus, assessment procedures also help in evaluation the suitability and effectiveness of the curriculum, the teaching methodology and the instructional materials." Thus the success of assessment depends on the effectiveness in the selection of valuation techniques and the implementation of the assessment procedure that is able to interpret valid students' ability. Therefore, in its implementation required a competent teacher who commit to implement professional and procedural assessment. Puckett & Black (2008) defines the scope of authentic assessment, namely: "The "*four P's*" of authentic assessment. The four words starting with P that are listed as characteristics of authenticity, however, seem to describe a valid performance-based assessments in general, not what most of advocates would argue are the crucial dimensions of authentic assessment: Process, Performance, Products and Portfolios." Therefore, in the implementation phase, authentic assessment includes an assessment of how the development process of student learning, how the student's performance, and how the products produced by the students.

Besides the planning and process phases, according to Regulation Ministry of Culture and Education No 66 Year 2013 and The Model of Assessment in Student Competencies Achievement, standard of outcome assessment are as follow: (1) process assessment result based on scoring criteria and guidelines for each assessed competencies; (2) determine the value with certain calculation and formulation; (3) compare the result of knowledge assessment with the minimum score, meanwhile the result of skills and manner assessment are determined by the achieved score using assessment rubric base; (4) create written report with numbers and described description; (5) assessment result is analyzed to know student's obstacles and development, that result is given back to the students with constructive comments; (6) write the result of assessment in numbers and/or competencies category as the documentation; (7) Conduct remedial program for those who have not met the minimum score; (8) conduct knowledge enrichment program for those who have met the minimum score. From that outcome standard, we can see that teachers are not apply this kind of assessment optimally, it implies from the obstacles and problems experienced by teachers. In fact, teachers do not do a result documentation systematically.

We are strongly recommend for teacher to have peer-discussion to enrich their knowledge in terms of the application of this authentic assessment program, so that there will be positive feedback for the students as well as their parents. Therefore, communication with their parents is really important, hence the parents will also know the developing of their children. The involvement of parents in this authentic assessment is really important, as what Frey & Schmitt (2007, p. 11) said that "the involvement of families in the assessments parallels the role of the students in authentic assessment for school-aged children."

The results of the assessment (outcome) should be valid, comprehensive, and representative so that it can truly describe the development of student learning process. This authentic output can be used as a self-reflection for both teachers and students to keep doing improvements, so students should always know their learning development. Paris & Ayres (1994) affirmed that "authentic assessment in terms suggesting that authenticity requires that the assessments be formative. They join some who argue that authentic assessment, Because it is formative, creates reflective students and teachers", it means that the authentic assessment is an assessment which can be used as a reflection towards a better achievement. Thus the assessment activities not only provide feedback to the students but also to teachers to continuously improve the quality of learning. It is also expressed by Black, Harrison, Lee, Marshall, & Wiliam (2003) that "An assessment activity can help learning if It provides information to be used as feedback by teachers and their pupils in assessing Themselves and each other, to modify the teaching and learning activities in the which they are engaged. Such assessment Becomes ormative assessment when the evidence is actually used to adapt the teaching to meet learning needs." Assessment activity can help learning process if there is information to be used as feedback by teachers and their students in assessing themselves and each other so that the implementation and learning process of assessment will always carried out through the right way. That argument is supported by Hattie & Timperley (2007) who wrote about the importance of feedback given to students, "Generally, the feedback has to be given as soon as possible after the completion of the learning task. Also students need to see that the feedforward comments can be incorporated into subsequent performance and overall influence the quality of Reviews their learning in positive ways. At the same time, in some instances, Temporarily withholding feedback is needed to allow the students to internalize and process the demands of the task." Generally, the feedback should be given as soon as possible after the completion of the students' task to make students realize and reflect advantages and disadvantages as well as to strive better in the future.

From some of the results of an evaluation, it is known that the overall teacher has not been able to implement an authentic assessment of learning optimally since there are obstacles faced by teachers, therefore, should be efforts to provide additional information that is more specific and practical toward teachers in terms of authentic assessment and its implementation in the learning activities, hence teachers would really understand the terms and implement the assessment program professionally. Professionalism of teachers in assessing students is important, success assessments by teachers can affect student success in learning and achievement of learning goals because the result can motivate students to see their reflection in learning process. Related to these findings, Jones (2005) support by saying "Teachers of make professional judgments on learners' performance in every teaching and learning sessions undertaken from, whether consciously or subconsciously. Using professional Reviews These judgments and translating them into a feedback on the quality of individuals' work is the focus of Assessment for Learning. Successful Assessment for Learning strategies result in improved learner progress on a continual basis. The principal characteristic of Assessment for Learning is effective feedback provided by teachers to learners on their progress." Teachers should carry out a professional assessment in order to describe the capabilities and quality of students in valid way, also able to provide feedback for both students and the undertaken learning.

Conclusion

From the result, we can conclude that: The planning stage (antecendents) or understanding of the authentic assessment design is not fully fulfil the standard, it can only categorized as Good with 68.75% of percentage. Process stage (transaction) or execution gets percentage of 63.41% and also categorized as Good. Results phase (outcomes) shows 68.48%, so that it also classified in Good category. Although the overall result has been running well, but of the standard that should be 100%, there are several obstacles in each stages and caused the implementation cannot be run well. Therefore we need some feedbacks toward for the sake of improvement, we need some efforts to provide additional information that is more specific and practical to teachers about authentic assessment and the need for commitment and professionalism of teachers in carrying out the assessment in accordance with the terms and systematics specified on the Curriculum 2013's standard of assessment. The results of this evaluation conducted in Province of Yogyakarta shows that not all aspects met the standards of 100%. Therefore, the findings of this study are expected to be used as a reflection to improve the performance of all stakeholders, especially teachers, principals, and parents to continue work together for the sake of improvement of the quality of authentic assessment in students. With coordination among all stakeholders, implementation of authentic assessment will be run more optimally.

Refferences

Alkin, M. C. (2011). *Essential evaluation*. New York: The Guilford Press.

Arikunto, S. (2013). *Dasar-dasar evaluasi pendidikan* (2nd ed.). Jakarta: Earth Literacy.

Black, P., Harrison, C., Lee, C., Marshall, B., & Wiliam, D. (2003). *Assessment for learning: putting it into practice*. Buckingham: Open University Press.

Doll, R. C. (1964). *Curriculum improvement: decision-making and process*. Boston: Allyn and Bacon.

Fook, C. Y., & Sidhu, G. K. (2010). Authentic assessment and pedagogical strategies in higher education. *Journal of Social Sciences*, *6*(2), 153–161. https://doi.org/10.3844/jssp.2010.153.161

Frey, B. B., & Schmitt, V. L. (2007).

Coming to terms with classroom assessment. *Journal of Advanced Academics*, *18*(3), 402–423. https://doi.org/10.4219/jaa-2007-495

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hein, G. E. (1991). *Active assessment for active science*. Alexandria: Association for Supervision and Curriculum Development.

Jones. (2005). *Assessment for learning*. London: Learning and Skills Development Agency.

Kementerian Pendidikan dan Kebudayaan. (2014). *Materi pelatihan guru implementasi kurikulum 2013 Tahun 2014 SD Kelas IV*. Jakarta: Kementerian Pendidikan dan Kebudayaan.

Khilmiyah, A., Sumarno, S., & Zuchdi, D. (2015). Pengembangan model penilaian keterampilan intrapribadi dan antarpribadi dalam pendidikan karakter di sekolah dasar. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *19*(1), 1–12. https://doi.org/10.21831/pep.v19i1.4550

Mendikbud RI. Peraturan Menteri Pendidikan dan Kebudayaan Nomor 66 Tahun 2013 tentang Standar Penilaian (2013).

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative data analysis: a methods sourcebook* (3rd ed.). New York: SAGE Publications, Inc.

Mueller, J. (2005). The authentic assessment toolbox: Enhancing student learning through online faculty development. *Journal of Online Learning and Teaching*, *1*(1). Retrieved from http://jolt.merlot.org/documents/vol1_no1_mueller_001.pdf

Palm, T. (2008). Authentic assessment and performance assessment: A conceptual analysis of the literature. *Practical Assessment, Research & Evaluation*, *13*(4). Retrieved from http://pareonline.net/getvn.asp?v=13%26n=4

Paris, S. G., & Ayres, L. R. (1994). *Becoming reflective students and teachers : with portfolios and authentic assessment*. Washington, D.C: American Psychological Association. Retrieved from http://www.apa.org/pubs/books/4316450.aspx?tab=1

Puckett, M. B., & Black, J. K. (2008). *Authentic assessment of the young child: celebrating development and learning* (2nd ed.). Des Moines, IA: Prentice-Hall Inc.

Sadler, D. R. (2005). Interpretations of criteria-based assessment and grading in higher education. *Assessment & Evaluation in Higher Education*, *30*(2), 175–194. https://doi.org/10.1080/0260293042000264262

Suarta, I. M., Hardika, N. S., Sanjaya, I. G. N., & Arjana, I. W. B. (2015). Model authentic self-assessment dalam pengembangan employability skills mahasiswa pendidikan tinggi vokasi. *Jurnal Penelitian Dan Evaluasi Pendidikan*, *19*(1), 46–57. https://doi.org/10.21831/pep.v19i1.4555

Sukardi. (2011). *Metodologi penelitian pendidikan: kompetensi dan praktiknya*. Jakarta: PT. Bumi Aksara.

Sukmadinata, N. S. (2009). *Pengembangan kurikulum teori dan praktik*. Bandung: PT Remaja Rosdakarya.

Wiggins, G. (1989a). A true test: toward more authentic and equitable assessment. *Phi Delta Kappan*, *70*(9), 703–713.

Wiggins, G. (1989b). Teaching to the (authentic) test. *Educational Leadership*, *46*(7), 41–47.

Jurnal Penelitian dan Evaluasi Pendidikan
Volume 21, No 1, June 2017 (116-125)
Online: http://journal.uny.ac.id/index.php/jpep

PENGEMBANGAN TES KETERAMPILAN DASAR OLAHRAGA BOLA TANGAN BAGI MAHASISWA

Ermawan Susanto
Fakultas Ilmu Keolahragaan Universitas Negeri Yogyakarta
Jl. Colombo No. 1, Depok, Sleman 55281, Yogyakarta, Indonesia
Email: ermawan_s@yahoo.com

Abstrak

Artikel ini bertujuan untuk: (1) menyusun butir-butir tes keterampilan dasar olahraga bola tangan, dan (2) mengetahui validitas dan mengestimasi reliabilitas tes keterampilan dasar olahraga bola tangan. Metode penelitian yang digunakan adalah *research and development*. Subjek penelitian 30 mahasiswa. Instrumen pengumpulan data berupa lembar observasi dan kuesioner. Jumlah *judge* yang terlibat 3 orang. Validitas diketahui dengan validitas isi dan reliabilitas instrumen menggunakan korelasi *Inter Rater*. Hasil penghitungan koefisien korelasi, diketahui bahwa skor rater 1 = 0,999, skor rater 2 = 0,996, dan skor rater 3 = 0,991. Hasil penghitungan reliabilitas diestimasi dengan *Koefisien Alpha* sebesar $r_{xx} =$ **0,994**. Hasil pengembangan tes keterampilan dasar olahraga bola tangan bagi mahasiswa menghasilkan 3 (tiga) jenis tes yaitu: (1) Tes keterampilan *passing* dengan sasaran ke tembok (waktu: 30 detik), (2) Tes keterampilan *dribbling* (waktu: 30 detik), (3) Tes keterampilan *flying shoot* melakukan 3 kali tembakan dari tiga posisi: kiri, tengah, dan kanan.

Kata kunci: *pengembangan tes, keterampilan dasar, bola tangan*

DEVELOPMENT OF HANDBALL BASIC SKILLS TEST FOR STUDENTS

Abstract

This article aims to: (1) develop a grain handball basic skills test, and (2) determine the validity and reliability of estimating handball basic skills test. The method used is research and development. Subject of the study 30 students. Data collection such as observation sheets and questionnaires. Number judge involved three people. Validity determined by content validity and reliability of the instrument using the correlation *Inter Rater*. The results of the correlation coefficient calculation, it is known that the rater score 1 = 0.999, the rater score 2 = 0.996, and the score rater 3 = 0.991. The results estimated by the reliability calculation of *Alpha Coefficient* $\mathbf{r_{xx} =}$ **0,994**. The result of the development of the form of handball basic skills test for students produce three (3) types of test are: (1) The test of passing skills the target wall (time: 30 seconds), (2) the tests of dribbling skills (time: 30 seconds) (3) The tests of flying shoot skills from three positions: left, center, and right.

Keywords: *test, development, handball basic skill*

Permalink/DOI: http://dx.doi.org/10.21831/pep.v21i1.15784

Pendahuluan

Olahraga permainan bola tangan (*Sport Handball*) merupakan cabang olahraga yang sebenarnya telah lama dikenal di Indonesia, dan sampai saat ini masih menjadi salah satu matakuliah yang diajarkan di Lembaga Pendidikan Tinggi Kependidikan (LPTK) keolahragaan. Keberadaannya kurang diperhitungkan karena beberapa hal, jenis permainan yang kurang populer, minim sosialisasi, dan tidak memiliki induk organisasi olahraga yang resmi. Sebagai salah satu cabang olahraga permainan, bola tangan memiliki beberapa dampak positif bagi pelakunya antara lain perkembangan fisik, kedisipinan, kerja sama, sosial emosional, dan keterampilan hidup. Hal ini tentu sesuai dengan tujuan pendidikan nasional secara umum. Demikian pula olahraga bola tangan dapat ditelusuri kebenaran sejarahnya dan telah berusia sangat tua. Sebuah fakta yang meyakinkan telah menunjukkan seseorang memainkan bola tangan jauh lebih awal daripada sepak bola. Permainan bolatangan yang dimainkan pada masa Yunani kuno merupakan sebuah isyarat terciptanya olahraga bola tangan modern (IHF, 2012, p.5).

Pada sejarahnya tahun 1928 *International Amateur Handball Federation* (IAHF) telah dideklarasikan bertepatan dengan Olimpiade Amsterdam dengan ketua Avery Brundage dari USA. Setelah tahun 1936 negara anggota IAHF menjadi 23 negara dan dilanjutkan dengan sebuah kompetisi yang disebut dengan "Berlin Olympic Games" di kota Berlin, Jerman. Tahun 1938 untuk pertama kali diselenggarakan Kejuaraan Dunia Bolatangan juga di Jerman. Akhirnya pada tahun 1946 atas usulan dan undangan Denmark dan Swedia, delapan negara memprakarsai Federasi Bola tangan Internasional. Delapan negara tersebut adalah; Denmark, Finlandia, Perancis, Belanda, Norwegia, Polandia, Swedia, dan Swiss (Moustafa, 2010, p.48-50).

Olahraga bola tangan dikatakan sebagai olahraga cepat dan dinamis yang dimainkan di dalam ruangan (*indoor*). Dalam catatan sejarah, olahraga ini telah dimainkan di lebih dari 150 negara. Bahkan sampai dengan tahun 2003, IHF memiliki jumlah anggota 150 negara dengan jumlah klub sebanyak 80.000 dan 19 juta atlit putra maupun putri (Rachman & Susanto, 2005, p.12).

Dalam permainannya bola tangan dimainkan di atas lapangan dengan panjang 40 meter x 20 meter. Saat berlangsung permainan, masing-masing tim terdiri atas 6 pemain dan 1 penjaga gawang. Waktu yang digunakan adalah 2 x 30 menit. Setiap tim terdiri dari 12 pemain. Namun, hanya 7 pemain yang ada di lapangan termasuk dengan seorang penjaga gawang. Selebihnya adalah pemain pengganti selama permainan.

Permainan bola tangan terdiri atas beberapa teknik dasar seperti *warming-up*, *dribbling*, *passing*, *shooting*, *possitioning*, *attacking exercise*, *defencing exercise*, dan *fast break exercise* namun dalam permainan hanya tiga teknik dasar yang paling sering digunakan, diantaranya; (1) teknik *dribbling* yaitu upaya pemain untuk membawa bola mendekati daerah pertahanan lawan dengan cara memantulkan bola ke lantai, (2) teknik *passing* yaitu upaya memberikan bola kepada teman dengan menggunakan satu atau dua tangan, (3) teknik *shooting* atau menembak bola ke gawang.

Menggiring (*dribble*) adalah keterampilan untuk menguasai dan membawa bola dengan cara memantulkannya setiap kali ke tanah, dengan satu atau dua tangan. Menggiring bola merupakan keterampilan yang cukup sulit karena memerlukan koordinasi mata dan tangan yang tinggi. Perlu diingat bahwa arah pantulan bola akan tergantung pada arah datang dari bola itu ke tanah. Cepat atau lambatnya pergerakan bola berasal dari kuat lemahnya menggiring bola tersebut.

Melempar atau *passing* adalah pola gerak dasar yang dimaksudkan untuk melepaskan suatu objek menjauhi tubuh pelempar. Gaya melempar memang berbeda-beda sesuai keperluannya tetapi pola dasarnya tetap konsisten atau sama. Bola dilempar kemudian bola tersebut harus ditangkap. Posisi tubuh untuk menangkap harus memungkinkan. Menembak (*shooting*) merupakan salah satu teknik terpenting dalam permainan bola tangan karena dengan teknik *shooting* kemungkinan terciptanya gol sangat besar.

Beberapa teknik *shooting* yang ada ialah *flying shoot*, *drive shoot*, *jump shoot*, dan *straight shoot*. Setiap regu berusaha dengan sekuat tenaga untuk memasukkan bola ke gawang agar kemenangan dapat diraih.

Namun demikian dalam perkuliahan dasar gerak bola tangan di jurusan pendidikan olahraga FIK UNY, belum ada instrumen tes keterampilan bola tangan bagi mahasiswa. Padahal instrumen tes keterampilan disusun guna mendukung pelaksanaan perkuliahan dan untuk mengukur keterampilan mahasiswa. Instrumen keterampilan dasar bolatangan yang akan disusun meliputi tes keterampilan melempar bola (*passing*), tes menggiring bola (*dribbling*), dan tes menembak bola (*shooting*).

Matakuliah dasar gerak bola tangan (*Foundation of Handball Technique*) merupakan matakuliah yang bersifat fakultatif dan diajarkan kepada seluruh program studi yang ada di Fakultas Ilmu Keolahragaan UNY. Matakuliah ini berbobot 1 sks dengan pelaksanaan praktek di lapangan. Matakuliah ini mengenalkan teknik dasar menggiring bola (*dribbling*), lemparan (*passing*), tembakan melayang bolatangan (*flying shoot*), peraturan permainan bola tangan, dan praktik bermain bola tangan (Kurikulum FIK, 2014, p. 42). Bola tangan merupakan cabang olahraga yang sebenarnya telah lama dikenal dan menjadi salah satu matakuliah yang diajarkan di LPTK keolahragaan.

Bola tangan (*handball*) diartikan sebagai permainan beregu yang menggunakan bola sebagai alatnya dan dimainkan dengan menggunakan satu atau kedua tangan. Bola tersebut dapat dilempar, dipantulkan, atau ditembakkan. Induk organisasi dari bola tangan ini adalah *International Handball Federation* (IHF) dan di Indonesia sendiri adalah Asosiasi Bola Tangan Indonesia (ABTI). Tujuan dari permainan ini adalah memasukkan bola sebanyak-banyaknya ke gawang lawan, dan mencegah agar tim lawan tidak dapat memasukkan bola ke gawang kita sendiri. Kunci keberhasilan agar dapat bermain dengan baik, seseorang harus mengerti dan benar-benar dapat menguasai teknik-teknik dasar yang ada seperti *passing, dribble, dan shooting*. Melempar atau *passing* adalah pola gerak dasar yang dimaksudkan untuk melepaskan suatu objek menjauhi tubuh pelempar. Gaya melempar memang berbeda-beda sesuai keperluannya tetapi pola dasarnya tetap konsisten atau sama. Bola dilempar kemudian bola tersebut harus ditangkap.

Penilaian suatu keterampilan dapat dilakukan dengan berbagai cara, salah satunya melalui pengamatan. Pengamatan dilakukan untuk mengetahui perkembangan dan sikap anak dalam kehidupan sehari-hari secara terus menerus (Depdiknas, 2003, p. 12). Berbagai alat penilaian yang dapat digunakan untuk memperoleh gambaran perkembangan perilaku anak, antara lain: (1) Portofolio yaitu penilaian berdasarkan kumpulan hasil kerja. (2) Unjuk kerja (*performance*) merupakan penilaian yang menuntut tugas dalam perbuatan yang dapat diamati, misalnya **praktek olahraga**, (3) Penugasan (*Project*) merupakan tugas yang memerlukan waktu relatif lama dalam pengerjaannya. (4) Hasil karya (*Product*) merupakan hasil kerja anak setelah melakukan suatu kegiatan.

Instrumen adalah suatu alat yang memenuhi persyaratan akademis sebagai alat untuk mengukur suatu objek ukur atau mengumpulkan data mengenai suatu variabel. Instrumen tersebut dapat digunakan untuk mengumpulkan data penelitian. Instrumen dibagi menjadi dua macam, yakni tes dan non-tes. Instrumen kelompok tes, misalnya tes prestasi belajar, tes inteligensi, tes bakat, tes keterampilan; sedangkan non-tes misalnya pedoman wawancara, angket atau kuesioner, pedoman observasi, daftar cocok *(check list*), dan skala penilaian (Sukmadinata, 2004, p. 47). Keterampilan gerak dasar bola tangan diukur menggunakan tes. Tes sebagai instrumen pengumpulan data adalah serangkaian latihan yang digunakan untuk mengukur keterampilan, pengetahuan, intelegensi, kemampuan atau motorik (Sugiyono, 2003, p. 138).

Terdapat empat konsep mendasar dalam menyusun tes yaitu validitas, reliabilitas, objektivitas dan norma. Valid berarti instrumen dapat digunakan untuk mengukur apa saja yang seharusnya diukur, reliabel berarti instrumen yang bila digunakan beberapa kali untuk mengukur objek yang sama, akan menghasilkan data yang sama. Reliabilitas instrumen keterampilan gerak bola tangan diestimasi dengan cara melakukan uji coba instrumen beberapa kali kepada responden, apabila koefisien korelasi positif dan signifikan maka instrumen dinyatakan reliabel. Instrumen keterampilan dasar bola tangan, disusun sendiri oleh peneliti danterdiri atas (1) ketepatan isi materi instrumen, (2) kelengkapan isi materi instrumen, (3) keterlaksanaan instrumen.

Metode Penelitian

Desain atau rancangan penelitian ini berbentuk penelitian pengembangan instrumen tes keterampilan *research & development* (Borg & Gall, 1983, p. 774) yang bertujuan sebagai alat ukur keberhasilan keterampilan bola tangan pada mahasiswa. Instrumen penelitian disusun sendiri oleh peneliti terdiri atas (1) ketepatan isi materi instrumen, (2) kelengkapan isi materi instrumen, (3) keterlaksanaan instrumen.

Penelitian dilakukan pada matakuliah olahraga pilihan permainan bola tangan prodi PJKR yang dilaksanakan pada semester enam (genap) di GOR Fakultas Ilmu Keolahragaan Universitas Negeri Yogyakarta. Subjek penelitian adalah mahasiswa prodi PJKR yang mengambil matakuliah sejumlah 30 mahasiswa. Instrumen pengumpulan data menggunakan pedoman observasi. Observasi digunakan untuk mendapatkan atau menjaring informasi dari para ahli sebagai *expert judgement* untuk memberikan masukan dan saran tentang instrumen tes keterampilan yang akan dihasilkan. Jumlah *judge* yang terlibat sejumlah 2 orang ahli pembelajaran bola tangan. Sebelum digunakan untuk pengambilan data yang asli kepada mahasiswa, dilakukan uji coba instrumen untuk mengetahui validitas serta reliabilitasnya. Validitas dan reliabilitas instrumen dibuktikan dengan uji korelasi **Inter Rater**. Analisa data menggunakan **Anova-General Multifacet Model** (Thorndike, 1982, p. 161).

Validitas dilakukan melalui analisis faktor terhadap instrumen dengan cara mengkorelasikan jumlah skor item pengamatan dengan skor total. Reliabilitas menggunakan nilai korelasi ICC atau dapat juga dihitung dengan rumus manual berdasar tabel ANOVA.

Secara garis besar langkah-langkah penyusunan dan pengembangan instrumen (Hadi, 2004, pp. 22-24), adalah sebagai berikut.

1. Berdasarkan sintesis dari teori-teori yang dikaji tentang suatu konsep dari variabel yang hendak diukur, maka dirumuskan konstruk dari variabel tersebut. Konstruk dalam penelitian ini adalah tes keterampilan dasar olahraga bola tangan.
2. Menetapkan besaran atau parameter. Pada penelitian ini besaran atau parameter keberhasilan keterampilan dasar olahraga bola tangan adalah tes *passing*, tes *dribbling*, dan tes *shooting*.
3. Jenis tes keterampilan yang akan dibuat harus melalui proses validasi, baik validasi teoretik maupun validasi empirik.
4. Revisi berdasarkan saran dari pakar atau berdasarkan hasil *judgements*.
5. Setelah konsep tes dianggap valid secara teoretik atau secara konseptual, dilakukanlah penggandaan tes secara terbatas untuk keperluan ujicoba.
6. Uji coba tes keterampilan dasar olahraga bola tangan di lapangan merupakan bagian dari proses validasi empirik. Melalui ujicoba tersebut, tes diberikan kepada sejumlah responden yaitu mahasiswa sebagai sampel uji coba, untuk kemudian diketahui valid atau tidaknya sebuah perangkat tes.
7. Selanjutnya dihitung koefisien reliabilitas. Koefisien reliabilitas dengan rentangan nilai (0-1) adalah besaran yang menunjukkan kualitas atau konsistensi hasil ukur tes. Makin tinggi koefisien reliabilitas makin tinggi pula kualitas tes.
8. Apabila ketujuh tahapan tersebut selesai kemudian tes jadi.

Penghitungan validitas dilakukan melalui analisis faktor terhadap instrumen dengan cara mengkorelasikan jumlah skor item pengamatan dengan skor total. Uji korelasi dilakukan untuk mencari besarnya hubungan dan arah hubungan. Nilai korelasi berkisar dalam rentang 0 sampai 1 atau 0 sampai -1 (Trihendardi, 2004, p.146)

Nilai korelasi ICC dapat juga dihitung dengan rumus manual berdasar tabel ANOVA seperti di bawah ini:

$$r = \frac{MS_{people} - MS_{residual}}{MS_{people} - (df_{people} \times MS_{residual})}$$

Hubungan antara ICC dengan alpha dapat diketahui melalui rumus berikut :

$$\boldsymbol{\alpha = \frac{k \times r}{1 + (k - 1) \times r}}$$

Besarnya nilai koefisien korelasi (r) dikategorikan sebagai berikut:

1. 0.7 – 1.00 baik positif maupun negatif, menunjukkan derajat hubungan yang tinggi,
2. 0.4 – 0.7 baik positif maupun negatif, menunjukkan derajat hubungan substansial,
3. 0.2 – 0.4 baik positif maupun negatif, menunjukkan derajat hubungan yang rendah,
4. < 0.2 baik positif maupun negatif, hubungan dapat diabaikan (Trihendardi, 2006, p. 145)

Penelitian ini bersifat uji coba pengembangan tes keterampilan gerak atau motorik fisik, teknik analisis data yang digunakan adalah dengan menilai tingkat kelayakan, kualitas, dan ketepatan tes yang dihasilkan. Tes dikatakan layak/tepat apabila langkah-langkah penelitian dapat dilaksanakan di setiap uji coba dan semua unsur yang terlibat. Selanjutnya tes dikatakan layak/ tepat, apabila dapat dipakai untuk mengukur keterampilan dasar olahraga bola tangan pada mahasiswa.

Hasil Penelitian dan Pembahasan

Data Analisis Kebutuhan

Analisis kebutuhan dalam penyusunan tes diperlukan untuk menyusun dan menggali permasalahan tes keterampilan dasar bola tangan bagi mahasiswa. Kegiatan ini dilakukan dengan cara menganalisis proses perkuliahan di lapangan, melakukan observasi, dan melakukan studi pustaka/ kajian literatur. Produk yang dihasilkan antara lain: (1) *draft* tes keterampilan dasar bola tangan bagi mahasiswa, (2) validitas dan reliabilitas tes keterampilan dasar bola tangan bagi mahasiswa, (3) butir-butir tes keterampilan dasar bola tangan bagi mahasiswa yang valid untuk diseminasikan menjadi instrumen final.

Validasi Ahli Draft Tes Keterampilan

Produk awal tes keterampilan dasar bola tangan sebelum diujicobakan dalam uji kelompok kecil dilakukan validasi oleh para ahli yang sesuai dengan bidang penelitian. Untuk memvalidasi produk yang akan dihasilkan, melibatkan dua (2) orang ahli keterampilan bola tangan sekaligus menguasai bidang pendidikan jasmani yang berasal dari dosen. Validasi dilakukan dengan cara memberikan *draft* produk awal, dengan disertai lembar evaluasi ahli.

Hasil evaluasi berupa nilai untuk aspek kualitas menggunakan skala likert 1 sampai 4. Data yang diperoleh dari pengisian kuesioner oleh para ahli, merupakan pedoman untuk menyatakan apakah produk tes keterampilan dasar bola tangan dapat digunakan untuk uji coba skala kecil dan skala luas. Berikut ini adalah *draft* tes keterampilan dasar bola tangan:

Tes Keterampilan Passing (waktu: 30 detik)

Tujuan : Untuk mengukur kemampuan melempar dan menangkap bola secara terus menerus.

Alat yang digunakan:

a. Bola tangan, 2 buah
b. Dinding/tembok
c. Stop watch
d. Pita pengukur
e. Kapur/ lakban

Tes Keterampilan Dribbling (jarak: 40 m)

Tujuan : untuk mengukur kemampuan menggiring bola di lapangan.

Alat yang digunakan:

a. Bola
b. Stop watch
c. Pita pengukur
d. Kapur/ lakban

Tes Keterampilan Flying Shoot (melakukan 6 kali tembakan)

Tujuan : untuk mengukur keterampilan menembak secara berturut-turut dari tiga posisi.

Alat yang digunakan:

a. Gawang
b. Tali
c. Pita pengukur
d. Bola, 6 buah

Berdasarkan hasil pengisisan kuesioner yang dilakukan oleh masing-masing ahli didapat rata-rata lebih dari 3 (tiga) atau masuk dalam kategori penilaian "baik/tepat/jelas". Oleh karena itu dapat disimpulkan bahwa instrumen tes keterampilan dasar bola tangan dapat digunakan untuk uji coba skala kecil. Masukan yang berupa saran dan komentar pada produk, sangat diperlukan untuk perbaikan pada tahap berikutnya.

Tabel 1. Hasil Pengisian Kuesioner Ahli

No	Aspek yang Dinilai	Nilai Ahli A1	Nilai Ahli A2
1.	Kesesuaian dengan kompetensi perkuliahan	4	4
2.	Kesesuaian dengan materi dasar bola tangan	3	3
3.	Ketepatan tes keterampilan dasar bola tangan.	2	3
4.	Kesesuaian alat dan fasilitas.	3	4
5.	Kemudahan tes keterampilan untuk dilakukan.	3	3
6.	Kesesuaian tes keterampilan dengan usia mahasiswa.	4	4
7.	Mendorong perkembangan aspek fisik .	3	4
8.	Mendorong perkembangan aspek kognitif.	3	4
9.	Mendorong perkembangan aspek psikomotor.	4	4
10.	Mendorong perkembangan aspek afektif.	2	3
11.	Dapat dilakukan siswa putra maupun putri.	4	4
12.	Mampu mengukur keterampilan dasar.	3	4
13.	Meningkatkan minat dan motivasi mahasiswa.	3	4
14.	Aman untuk diterapkan dalam perkuliahan.	3	4
Jumlah Skor		44	52
Rata-rata		3,14	3,71

A1 = Ahli 1
A2 = Ahli 2

Uji Coba Skala Kecil

Setelah produk tes keterampilan dasar bola tangan divalidasi oleh para ahli serta dilakukan revisi, kemudian produk diujicobakan kepada mahasiswa. Uji coba ini dilakukan terhadap 30 mahasiswa.

Uji coba bertujuan untuk mengetahui dan mengidentifikasi berbagai permasalahan seperti kelemahan, kekurangan, ataupun keefektifan tes keterampilan. Pengamatan yang dilakukan oleh *rater*, merupakan salah satu indikator untuk mengetahui keefektifan tes. Pengamatan oleh *rater* dilakukan selama tes berlangsung. Berdasarkan hasil pengamatan, didapatkan tiga bentuk tes keterampilan bola tangan dan nilai validitas instrumen.

Validitas Tes

Jumlah subjek atau mahasiswa yang digunakan dalam uji coba skala kecil adalah sejumlah 30 mahasiswa. Uji korelasi dilakukan dengan uji inter rater (antar penilai), analisa data menggunakan **Anova-General Multifacet Model** dari Thorndike yaitu untuk menguji dua variabel bertipe ordinal dan skala dengan distribusi normal/parametrik. Data hasil uji validitas pengamatan menunjukkan derajat hubungan yang tinggi sebesar rata-rata **0.995**.

Tabel 2. Tingkat Validitas Tes Bola Tangan

No	Perbandingan skor	Koef. kor	Status
1	Rater 1 – skor total rater	r = 0.999	Valid
2	Rater 2 – skor total rater	r = 0.996	Valid
3	Rater 3 – skor total rater	r = 0.991	Valid

Dengan demikian berdasarkan penghitungan statistik validitas uji coba tes keterampilan, diketahui terdapat tingkat hubungan positif yang tinggi, sehingga instrumen dinyatakan valid dan dapat digunakan untuk pengambilan data pada skala luas. Berdasarkan hasil analisis faktor tersebut dapat disimpulkan bahwa tes tersebut memiliki *construct validity* yang baik, artinya tes tersebut dapat digunakan untuk mengukur gejala sesuai dengan yang didefinisikan (Sugiyono, 2003, p. 170).

Reliabilitas Tes

Uji reliabilitas antar rater terdiri dari dua jenis, uji koefisien korelasi Kesepakatan Antar Rater dari Kappa dan uji koefisien korelasi Antar-Kelas (*Intraclass Correlation Coefficients*, ICC). Uji reliabilitas antar rater dari Kappa digunakan apabila rater berjumlah dua orang sedangkan uji reliabilitas antar rater ICC digunakan apabila rater lebih dari 2 orang (Widhiarso, 2006, p.15). Penelitian ini menggunakan 3 rater sehingga menggunakan koefisien korelasi Antar Kelas. ICC menunjukkan perbandingan antara variasi yang diakibatkan atribut yang diukur dengan variasi pengukuran secara keseluruhan. Berdasarkan penghitungan statistik reliabilitas uji coba skala kecil, diketahui terdapat nilai reliabilitas Antar-Rater yang tinggi yaitu **0.994**, sehingga instrumen dinyatakan reliabel dan dapat digunakan untuk pengambilan data.

Pembahasan

Berdasarkan hasil distribusi frekuensi pada 30 mahasiswa, diketahui: (1) menurut *rater 1* yang termasuk dalam kategori baik berjumlah 22 mahasiswa (74%), kategori sedang berjumlah 6 mahasiswa (20%), dan kategori kurang berjumlah 2 mahasiswa (6%), (2) menurut *rater 2* yang termasuk dalam kategori baik berjumlah 20 mahasiswa (68%), kategori sedang berjumlah 8 mahasiswa (26%), dan kategori kurang berjumlah 2 mahasiswa (6%), (3) menurut *rater* 3 yang termasuk dalam kategori baik berjumlah 20 mahasiswa (68%), kategori sedang berjumlah 8 mahasiswa (26%), dan kategori kurang berjumlah 2 mahasiswa (6%). Rata-rata distribusi frekuensi psikomotorik pada 30 mahasiswa, diketahui bahwa: (1) yang termasuk kategori baik berjumlah 21 mahasiswa (70%), (2) yang termasuk kategori sedang berjumlah 7 mahasiswa (23%), (3) yang termasuk kategori kurang berjumlah 2 mahasiswa (7%).

Berdasarkan langkah-langkah penelitian pengembangan untuk menghasilkan produk yang telah dilakukan, maka didapatkan produk akhir berupa tes keterampilan dasar olahraga bola tangan. Indikator keberhasilan produk ini ialah adanya kesamaan persepsi antar *rater* berupa lembar penilaian hasil pengamatan terhadap seluruh subjek yang diujicobakan dalam penelitian. Berdasarkan uji coba yang dilakukan pada 30 mahasiswa yang memiliki karakteristik sama, didapatkan hasil yang hampir sama, artinya produk yang diujicobakan bisa diterapkan pada kelompok mahasiswa dengan karakteristik yang sama.

Secara permainan, olahraga bola tangan akan berjalan dengan benar apabila pemain mampu menguasai keterampilan dasar bola tangan. Keterampilan dasar yang dimaksud antara lain : (1) *Ball handling*, (2) *Dribbling*, (3) *Passing*, (4) *Shooting*, dan (5) *Positioning* (Thum, 2005, pp. 45-47). *Dribbling* adalah keterampilan menggiring bola yang bertujuan melakukan penyerangan dengan menggiring bola kemudian mendekatkannya pada area tembakan. *Driblling* dalam permainan bola tangan dikenal dengan teknik "tiga langkah sekali pantul." Menggiring bola merupakan suatu pergerakan memantulkan bola ke lantai secara kontinyu dengan menggunakan sebelah tangan atau bertukar tangan tanpa memegang bola. Keterampilan menggiring digunakan dalam 3 situasi; (1) Bergerak bebas bila tidak ada penjagaan lawan; (2) Satu lawan satu; (3) Pemain lawan tidak dapat membuat halangan setelah menerima bola.

Passing atau melempar bola adalah keterampilan memberikan bola kepada kawan dalam permainan. Tujuan *passing* adalah untuk memberikan bola kepada kawan yang kemudian melakukan *shooting*, atau untuk tujuan strategi penyerangan maupun pertahanan. Terdapat berbagai macam keterampilan *passing* dalam permainan bola tangan.

Shooting atau menembak adalah keterampilan melakukan tembakan sebagai bagian usaha untuk mencetak goal dalam permainan bola tangan. Tujuan *shooting* adalah untuk mencetak goal sebanyak-sebanyaknya ke gawang lawan. Terdapat berbagai macam keterampilan *shooting* dalam permainan bola tangan.

Dalam menyusun tes perlu diperhatikan empat konsep mendasar yang ada yaitu *Validitas, Reliabilitas, Objektivitas dan Norma*. Valid berarti tes dapat digunakan untuk mengukur apa saja yang seharusnya diukur, reliabel berarti tes yang bila digunakan beberapa kali untuk mengukur objek yang sama, akan menghasilkan data yang sama (Sugiyono, 2003, p.140). Pada pengukuran keterampilan dasar bola tangan mahasiswa maka menggunakan validitas konstruk (*construct validity*) dan validitas isi (*content validity*). Instrumen yang mempunyai validitas isi adalah instrumen yang berbentuk tes dan sering digunakan untuk mengukur prestasi belajar. Untuk menyusun tes keterampilan dasar bola tangan, maka disusun berdasarkan tiga komponen dasar: melempar bola (*passing*), menggiring bola (*dribbling*), dan menembak dengan bola (*shooting*). Adapun untuk menguji validitas konstruk menggunakan pendapat ahli (*expert judgement*). Untuk menguji reliabilitas instrumen keterampilan dasar bola tangan, dilakukan dengan cara melakukan uji coba tes beberapa kali kepada responden/model.

Tes sebagai instrumen pengumpulan data adalah serangkaian latihan yang digunakan untuk mengukur keterampilan pengetahuan, intelegensi, motorik yang dimiliki oleh individu atau kelompok. Ada beberapa alasan mengapa tes perlu dilakukan yaitu: (1) mengklasifikasikan peserta didik, (2) mendiagnosa kebutuhan dan kelemahan peserta didik, (3) evaluasi pembelajaran, (4) evaluasi program, (5) *marking/griding*, (6) motivasi, (7) alat pembelajaran, (8) prediktor penelitian (Suntoda, 2007, p. 67)

Tes keterampilan dasar olahraga bola tangan merupakan bagian yang integral dalam proses penilaian hasil belajar mengajar dan latihan, melalui tes dan pengukuran kita akan memperoleh data yang objektif (Susanto, 2015, p. 75). Berdasarkan hasil pengembangan tes keterampilan dasar bola tangan bagi mahasiswa, diperoleh hasil tes sebagai berikut :

1. Tes keterampilan *passing* dengan sasaran ke tembok (waktu: 30 detik)
2. Tes keterampilan *dribbling* (waktu: 30 detik)
3. Tes keterampilan *flying shot* melakukan 3 kali tembakan dari tiga posisi: kiri, tengah, dan kanan

Berikut ini adalah produk jadi berupa tes keterampilan dasar olahraga bola tangan bagi mahasiswa antara lain:

1. Tes keterampilan *passing* (waktu: 30 detik)
   Tujuan : mengukur kemampuan melempar dan menangkap bola secara terus menerus.
2. Tes keterampilan *dribbling* (jarak: 40 m)
   Tujuan : Untuk mengukur kemampuan menggiring bola di lapangan.
3. Tes keterampilan *flying shoot (*melakukan 6 kali tembakan dari 3 posisi)
   Tujuan : mengukur keterampilan menembak secara berturut-turut dari tiga posisi.

Untuk melakukan penilaian gerak, penting memperhitungkan prinsip-prinsip tes keterampilan. Prinsip tes tersebut salah satunya adalah fungsi pengembangan keterampilan motorik. Penilaian dalam bidang olahraga ada yang bersifat objektif dan ada yang subjektif. Dalam penilaian objektif tentunya berdasarkan hasil pengukuran yang objektif. Pada penilaian yang bersifat subjektif umumnya dilakukan terhadap *performance*.

Simpulan

Berdasarkan hasil penelitian pengembangan dan pembahasan di atas maka hasil penelitian menunjukkan telah tersusunnya 3 (tiga) jenis tes keterampilan dasar olahraga bola tangan bagi mahasiswa. Tes tersebut antara lain: (1) Tes keterampilan *passing* dengan sasaran ke tembok (waktu: 30 detik), (2) Tes keterampilan *dribbling* (waktu: 30 detik), (3) Tes keterampilan *flying shoot* melakukan 3 kali tembakan dari tiga posisi: kiri, tengah, dan kanan. Adapun nilai validitas tes **0.995** dan nilai reliabilitas tes **0.994**.

Daftar Pustaka

Borg, W. R. & Gall, M. D. (1983). *Educational research: an introduction* (4th ed.). New York: Longman Inc.

Departemen Pendidikan Nasional. (2003). Undang-undang Nomor 20 Tahun 2003 tentang Sistem Pendidikan Nasional. Jakarta.

Hadi, S. (2004). *Metodologi riset.* (2nd ed.). Yogyakarta: Penerbit Fakultas Psikologi UGM

IHF. (2012). *International handball federation. rules of the game.* Basel Switszerland.

FIK UNY (2014). *Kurikulum program studi pendidikan jasmani, kesehatan, dan rekreasi 2014, KKNI & KBK.* Yogyakarta: FakultasIlmu Keolahragaan UNY.

Moustafa, H. (2010). Teaching handball at School. introduction to handball for student aged 5 to 11. *Handbook. International Handball Federation.* (IHF).

Rachman, H. A. & Susanto, E. (2005). *Bolatangan, sebuah pengantar dalam pembelajaran.* Universitas Negeri Yogyakarta.

Sugiyono. (2003). *Metode penelitian bisnis.* Bandung: Alfabeta.

Suntoda, A. (2007). Pedoman dan instrumen praktikum tes dan pengukuran olahraga. *Panduan Praktikum.* Bandung.

Sukmadinata, N. S. (2004). *Kurikulum dan pembelajaran kompetensi.* Bandung: Kesuma Karya.

Susanto, E. (2015). *Olahraga permainan bola tangan.* Yogyakarta: UNY Press.

Thum, Hans-Peter. (2005). *Handball elementary course for physical education teachers and students.* State University of Yogyakarta.

Thorndike, Robert L. (1982). *Applied Psycometrics.* Houghton Mifflin Company Boston Massachusetts.

Trihendardi, Cornelius. (2004). *Langkah mudah memecahkan kasus statistik: deskriptif, parametrik dan non-parametrik dengan SPSS 12*. Yogyakarta: Andi Offset.

Widhiarso, Wahyu. (2006). *Mengestimasi reliabilitas, SPSS untuk psikologi*. Yogyakarta: Fakultas Psikologi.