

## MODEL PENILAIAN KEMAMPUAN BERPIKIR HISTORIS (HISTORICAL THINKING) DENGAN MODEL RASCH

---

*Ofianto*

Prodi Pendidikan Sejarah, Fakultas Ilmu Sosial, Universitas Negeri Padang  
ofianto.anto@yahoo.com

*Wahidul Basri*

Prodi Pendidikan Sejarah, Fakultas Ilmu Sosial, Universitas Negeri Padang  
wahidulsejarahunp@gmail.com

### ABSTRACT

*This study was conducted to produce a model and instruments of historical thinking skills in the history subject at the senior high school (SHS) and to identify SHS students' historical thinking skills. The study was conducted in two stages, i.e. model development and instrument development along with a small-scale tryout and a large-scale tryout. The test for each tryout consisted of six and five sub-test sets. Each test set contained 20 anchor items. The sample for each tryout comprised 1573 and 2613 testees. The data were analyzed by means of Partial Credit Model (PCM) using the QUEST program. The tryout results on whole indicated that based on the criteria for an INFIT MNSQ mean of 1.0 and a standard deviation of 0.0 the tests fitted PCM. The reliability coefficients of the tests for the tryouts were moderately good; the Cronbach's alpha coefficients were, respectively, 0.65 and 0.54. The lowest score of historical thinking skills was -.352 and the highest was +1.21 in an ideal range of -4.0 to +4.0. On the whole, the testees' scores were not satisfactory. Only 5.89% of the testees were above the expected median.*

**Keywords:** *Instrument development, test, historical thinking skills, polytomous, PCM*

### ABSTRAK

Penelitian dilakukan untuk menghasilkan model beserta instrumen penilaian kemampuan berpikir historis mata pelajaran sejarah SMA dan mengidentifikasi kemampuan berpikir historis siswa SMA. Penelitian ini dilakukan melalui dua tahap, yakni tahap pengembangan model dan tahap pengembangan instrumen yang disertai uji coba terbatas dan uji coba diperluas. Tes untuk masing-masing uji coba terdiri atas enam dan lima perangkat subtes. Masing-masing perangkat tes memuat 20% *anchor* item. Sampel yang digunakan untuk masing-masing uji coba sebanyak 1573 dan 2673 testi. Data dianalisis mengikuti *Partial Credit Model* (PCM) menggunakan program *QUEST*. Hasil keseluruhan uji coba menunjukkan bahwa berdasarkan kriteria mean INFIT MNSQ 1,0 dan simpangan baku 0,0 tes terbukti *fit* dengan PCM. Reliabilitas tes untuk masing-masing uji coba cukup baik, dengan koefisien alpha Cronbach masing-masing 0,65 dan 0,54. Skor keterampilan berpikir historis terendah -,352 dan tertinggi +1,21 dari rentang ideal -4,0 sampai +4,0. Secara keseluruhan skor testi jauh dari memuaskan. Hanya 5,89% testi yang berada di atas nilai tengah yang diharapkan.

**Kata Kunci:** Pengembangan instrumen, tes, kemampuan berpikir historis, politomus, dan PCM

## A. Pendahuluan

Penilaian merupakan komponen penting dalam penyelenggaraan pendidikan. Penilaian dilakukan untuk melihat dan memantau perkembangan mutu pendidikan dari waktu ke waktu<sup>1</sup>. Oleh karena itu, untuk melakukan penilaian terhadap mutu pendidikan dapat menggunakan beragam alat penilaian. Alat penilaian dapat berupa tes dan nontes<sup>2</sup>. Penggunaan beragam alat penilaian dimaksudkan untuk memotret hasil belajar peserta didik secara komprehensif. Dengan demikian, penilaian bermanfaat untuk melihat mutu pendidikan secara menyeluruh dan penilaian memberikan informasi penting untuk perbaikan pembelajaran.

Penilaian hasil belajar sejarah di sekolah menurut Permendiknas No. 20 Tahun 2007<sup>3</sup> tentang Standar Penilaian untuk Satuan Pendidikan Dasar dan Menengah, memuat tiga aspek yakni: aspek akademik, aspek kesadaran sejarah, dan aspek nasionalisme. Dalam melakukan penilaian di sekolah, guru hendaknya

memperhatikan kesesuaian antara standar (kompetensi), isi/konten (materi kurikulum), penilaian, dan strategi pembelajaran<sup>4</sup>.

Penilaian yang dilakukan guru sejarah berdasarkan hasil survei pendahuluan yang dilakukan, menunjukkan penilaian yang diberikan guru, rata-rata menggunakan tes pilihan yang bersifat objektif dan memiliki kecenderungan untuk menuntut peserta didik menghafal fakta-fakta. Kenyataan ini telah diteliti beberapa ahli seperti Bain<sup>5</sup>, Barton & Levstik<sup>6</sup>, dan Lee<sup>7</sup>. Hasilnya menunjukkan, praktik penilaian sejarah saat ini berkisar pada ingatan faktual yang bertumpu pada tes pilihan ganda. Kenyataan lain adalah bahwa, tes tertulis sebagai salah satu alat

---

<sup>1</sup> Allen, M.J. & Yen, W.M. 1979. *Introduction to Measurement Theory*. Belmont, CA: Wadsworth, Inc.; Griffin, P. & Nix, P. 1991. *Educational Assessment and Reporting: A New Approach*. Sydney: Harcourt Brace Jovanovich, Publishers.

<sup>2</sup> Mardapi, D. 2008. *Teknik Penyusunan Instrumen Tes dan Nontes*. Yogyakarta: Mitra Cendikia Press.

<sup>3</sup> Depdiknas. 2007. *Peraturan Menteri Pendidikan Nasional Republik Indonesia Nomor 20, Tahun 2007, tentang Standar Penilaian Pendidikan untuk Satuan Pendidikan Dasar dan Menengah*. Jakarta: Depdiknas.

---

<sup>4</sup> Ashby, R., Lee, P.J. & Shemit, D. 2005. "Putting principles into practice: teaching and planning". Dalam M.S. Donovan and J.D. Bransford (Eds.). *How Students Learn: History, Mathematics, and Science in the Classroom*. Washington, DC: The National Academies Press.

<sup>5</sup> Bain, R. B. 2005. "Applying the principles of how people learning teaching high school history". Dalam M.S. Donovan and J.D. Bransford (Eds.). *How Students Learn: History, Mathematics, and Science in the Classroom*. Washington, DC: The National Academies Press.

<sup>6</sup> Barton, K.C. & Levstik, L.S. 2003. "Why don't more history teachers engage students in interpretation?". *Research and Practice Social Education*, 67 (6), 358-361.

<sup>7</sup> Lee, P. 2005. "Putting principles into practice: understanding history". Dalam M.S. Donovan & J.D. Bransford (Eds.). *How Students Learn: History, Mathematics, and Science in the Classroom*. Washington, DC: The National Academies Press.

penilaian yang selama ini digunakan untuk mengungkap kemampuan atau hasil belajar siswa, sering dikonstruksi secara tidak sistematis. Akibatnya, banyak tes yang digunakan guru tidak dapat mengungkap kemampuan peserta didik yang sesungguhnya. Hasil penelitian yang dilakukan oleh Mardapi, dkk.<sup>8</sup> mengungkapkan bahwa masih banyak guru-guru dalam membuat soal tes tidak berpedoman pada kisi-kisi tes, tetapi cenderung hanya menggunakan soal-soal yang ada pada buku-buku yang beredar di pasaran.

Terkait dengan hal tersebut, guru hendaknya juga membiasakan menggunakan tes bentuk lainnya, misalnya dengan bentuk uraian, yang lebih sesuai dengan karakteristik mata pelajaran serta tujuan pembelajaran yang diharapkan. Tuntutan dalam rumusan salah satu Kompetensi Dasar (KD) dalam standar isi kurikulum nasional untuk Sekolah Menengah Atas/Madrasah Aliyah (SMA/MA), adalah agar peserta didik dapat mengembangkan kemampuan untuk memahami dan menerapkan prinsip dasar *inquiry*, yang merupakan aplikasi dari keterampilan berpikir historis dalam mata pelajaran sejarah.

Keterampilan berpikir historis dapat didefinisikan sebagai langkah-langkah/proses ilmiah dalam belajar sejarah<sup>9</sup>. Dalam setiap proses berpikir

historis, selalu melibatkan proses berpikir. Dengan demikian, keterampilan berpikir historis juga dapat mendorong berkembangnya kemampuan berpikir kritis dan kreatif pada diri peserta didik.

Berdasarkan uraian tersebut, untuk mengukur kemampuan berpikir historis digunakan tes uraian. Untuk itu perlu disusun instrumen kemampuan berpikir historis yang terdiri atas tes dan pedoman penilaian. Dengan demikian, diperlukan adanya penelitian pengembangan instrumen penilaian kemampuan berpikir historis yang terdiri atas tes dan pedoman penilaian.

## B. Tinjauan Kepustakaan

Teknik penilaian yang dilakukan melalui ujian merupakan kegiatan pengukuran karena melalui ujian dapat diperoleh data numerik untuk mengembangkan karakteristik kemampuan peserta didik<sup>10</sup>. Salah satu mata pelajaran yang diajarkan mulai dari tingkat sekolah dasar hingga tingkat sekolah menengah adalah mata pelajaran sejarah. Pembelajaran sejarah di sekolah bertujuan agar peserta didik memperoleh kemampuan berpikir historis<sup>11</sup>, mendorong

---

<sup>8</sup>Mardapi, D. 1999. "Estimasi kesalahan pengukuran dalam bidang pendidikan dan implikasinya pada ujian nasional". Disampaikan pada *Pidato Pengukuhan Guru Besar, tanggal 4 Mei 1999*. Yogyakarta: Universitas Negeri Yogyakarta.

<sup>9</sup>Seixas, P. & Peck, C. 2004. "Teaching historical thinking". Dalam A. Sears & I. Wright (Eds.), *Challenges and Prospects*

---

*for Canadian Social Studies*. Vancouver: Pacific Educational Press; Seixas, P. 2013. *Linking Historical Thinking Concepts, Content and Competencies*. Vancouver: Pacific Educational Press.

<sup>10</sup> Hargreaves, A., Earl, L. & Schmidt, M. 2002. "Perspectives on alternative assessment reform". *American Educational Research Journal*, 39 (1), 69-95.

<sup>11</sup> Fogu, C. 2009. "Digitalizing historical consciousness". *Journal History and Theory*, 47 (1), 103-121.

peserta didik berpikir kritis-analitis<sup>12</sup>, memanfaatkan pengetahuan tentang masa lampau untuk memahami kehidupan masa kini dan yang akan datang.

Analisis hasil penilaian merupakan informasi penting untuk perbaikan pembelajaran, untuk itu, para ahli psikometri mengembangkan satu model analisis yang dikenal dengan teori tes (*test theory*)<sup>13</sup>. Teori tes yang sudah cukup lama berkembang dan digunakan adalah teori tes klasik (*classical test theory = CTT*)<sup>14</sup>. CTT dalam estimasinya mengandung banyak kesalahan (*error*) dan memberikan hanya sedikit informasi. Dalam perkembangannya, untuk mengatasi kelemahan mendasar CTT dikembangkan teori pengukuran modern (*item response theory = IRT*)<sup>15</sup>. Model IRT memberikan informasi yang lebih banyak dengan

asumsi yang banyak pula. IRT terdiri dari tiga model yakni model Rasch atau model satu parameter logistik (1-PL), model dua parameter logistik (2-PL), dan model tiga parameter logistik (3-PL).

### C. Metode Penelitian

Penelitian ini adalah penelitian pengembangan, yakni untuk mengembangkan tes kemampuan berpikir historis siswa SMA. Prosedur dan langkah-langkah pengembangan yang digunakan dalam penelitian ini mengacu pada prosedur penelitian dan pengembangan yang dikemukakan Borg & Gall<sup>16</sup>. Namun, tahapannya disesuaikan dengan tujuan dan kepentingan penelitian ini. Adapun tahapan penelitian dan pengembangan yang dilakukan pada penelitian ini, meliputi tahap: (1) analisis kebutuhan dan pengumpulan informasi awal, (2) perencanaan dan penyusunan model, (3) uji coba model, (4) evaluasi (5) implementasi, dan 6) diseminasi.

Analisis kebutuhan atau masalah dan pengumpulan informasi awal (*preliminary investigation*) dilakukan dalam bentuk kegiatan pengamatan langsung/survei dan studi literatur. Hasil dari kegiatan ini, dijadikan sebagai dasar untuk menyusun draf awal model penilaian/tes.

Kegiatan yang dilakukan pada tahap perencanaan (*planning*), adalah menyaring dan merumuskan semua informasi yang diperoleh pada tahap pertama.

---

<sup>12</sup> Winerburg, S. 2006. *Berpikir historis: Memetakan Masa Depan, Mengajarkan Masa Lalu*. (Terjemahan Masri Maris). Jakarta: Yayasan Obor Indonesia.

<sup>13</sup> Rasch, G. 1961. "On general laws and the meaning of measurement in psychology". *The Danish Yearbook of Philosophy*, 4 (1), 321-334; Rasch, G. 1977. "On Specific Objectivity: An attempt at formalizing the request for generality and validity of scientific statements". *The Danish Yearbook of Philosophy*, 14 (3), 58-93.

<sup>14</sup> Van der Linden, W.J. & Hambleton, R.K. 1997. *Handbook of Modern Item Response Theory*. New York: Springer; Hambleton, R.K. & Swaminathan, H. 1985. *Item Response Theory*. Boston, MA: Kluwer Inc.

<sup>15</sup> Masters, G.N. 1999. "Partial credit model". Dalam J.P. Keeves & G.N. Masters (Eds.). *Advances in Measurement in Educational Research and Assessment*. Amsterdam: Pergamon.

---

<sup>16</sup> Borg, W.R. & Gall, M.D. 1989. *Educational Research: An Introduction*. New York: Longman.

Selanjutnya, tahap penyusunan model, pada tahap ini merupakan tahap pengembangan instrumen/tes kemampuan berpikir historis SMA. Tahapan pengembangan tes mengikuti tahapan menurut Oriondo & Dallo-Antonio<sup>17</sup>. Tahap dalam pengembangan tes tersebut mencakup (1) perancangan tes dan 2) uji coba tes. Kegiatan perancangan tes, dilakukan sampai pada perakitan tes yang siap untuk diujicobakan.

Kegiatan perancangan instrumen/tes meliputi: a) menyusun *learning continuum* (LC), b) penyisipan kisi-kisi tes/instrumen kemampuan berpikir historis, berdasarkan *learning continuum*, c) penulisan item, dan d) perbaikan item dan perakitan tes/instrumen. Penskalaan dibuat menjadi penskalaan politomus, sesuai dengan bentuk tes yang digunakan yakni uraian. Untuk penskalaan politomus menggunakan skala 0-2 atau 3 kategori.

Perbaikan item dilakukan setelah dilakukan analisis kualitatif terhadap kualitas item yang disusun. Analisis kualitatif terhadap item tidak dapat terlepas dari LC, dan kisi-kisinya. Oleh karena itu, langkah pertama dilakukan telaah terhadap LC, indikator, beserta kisi-kisi dan itemnya, melalui *focus group discussion* (FGD).

Instrumen kemampuan berpikir historis yang sudah di rancang, dilakukan uji coba terbatas untuk memperoleh data empirik. Hasil uji coba dianalisis, baik secara klasik

maupun dengan pendekatan *item response theory* (IRT). Analisis dilakukan untuk melihat kualitas butir-butir tes, sebelum disusun kembali untuk siap digunakan pada tahap uji coba diperluas atau implementasi.

Selanjutnya, tahap tes, evaluasi, dan revisi. Pada tahap ini, kegiatan yang dilakukan adalah melakukan uji coba model yang dikembangkan melalui uji coba terbatas. Data dari hasil uji coba kemudian dianalisis untuk mengetahui apakah model yang dikembangkan *fit* atau belum.

Uji coba diperluas, dilakukan setelah uji coba terbatas dan telah mengalami perbaikan atau revisi. Hasilnya kemudian dianalisis untuk mengetahui sejauh mana penguasaan siswa terhadap tes kemampuan berpikir historis.

Produk akhir dari model yang dikembangkan selanjutnya disebarluaskan kepada pemakai dan pengambil kebijakan pada tingkat sekolah, di antaranya adalah guru, kepala sekolah, kepala dinas pendidikan kota/kabupaten dan provinsi. Diseminasi dilakukan dalam bentuk pengiriman ringkasan hasil penelitian ke sekolah sampel.

Uji coba produk dilaksanakan sebanyak dua kali, yaitu: (1) uji coba terbatas, dan (2) uji coba diperluas. Kegiatan yang dilakukan pada uji coba terbatas meliputi: pelaksanaan tes, analisis hasil tes. Kegiatan yang dilakukan pada saat uji coba diperluas meliputi: pelaksanaan tes, analisis hasil tes, dan penafsiran hasil tes.

Terkait dengan subjek coba, penelitian ini dilakukan di Sumatera Barat. Subjek uji coba produk adalah siswa SMA. Sekolah SMA yang

---

<sup>17</sup>Oriondo, L.L. & Dallo-Antonio. 1998. *Evaluating Educational Outcomes (test, measurement, and evaluation)*. Quezon City: REX Printing Company, Inc.

dilibatkan pada penelitian ini adalah SMA yang berada di ibu kota provinsi dengan kategori favorit, sampai dengan SMA di ibu kota kecamatan, dengan kategori tidak favorit. Alasan utamanya adalah agar dapat diperoleh variabilitas hasil pengukuran yang maksimal.

Data yang terkumpul dalam penelitian adalah data kuantitatif. Data kuantitatif, berupa hasil tes siswa yang terdiri dari data hasil uji coba terbatas dan data hasil uji coba diperluas. Pengumpulan data dalam penelitian ini dilakukan dengan menggunakan seperangkat tes.

Untuk mengetahui kualitas instrumen tes, dilakukan analisis baik, secara kualitatif dengan validasi pakar (*expert judgement*) dari segi aspek isi (materi), konstruksi, dan bahasa, maupun secara kuantitatif melalui proses uji coba (uji empiris). Data hasil uji coba dianalisis dengan menggunakan program Quest. Tujuan dari analisis ini adalah untuk mengetahui kualitas parameter butir soal dan tingkat kehandalan tes. Kualitas parameter butir soal, hanya tingkat kesukaran butir saja karena menggunakan model 1-PL/model Rasch. Sedangkan tingkat kehandalan tes ditunjukkan oleh nilai koefisien alpha.

Data hasil tes uji coba diperluas, dianalisis dengan menggunakan program Quest. Analisis data

hasil tes uji coba diperluas, dilakukan untuk memperoleh informasi tentang karakteristik parameter butir, parameter kemampuan peserta, dan penguasaan siswa terhadap tes kemampuan berpikir historis di sekolah.

#### **D. Hasil Penelitian dan Pembahasan**

Model penilaian kemampuan berpikir historis yang dihasilkan pada penelitian ini, termasuk ke dalam model prosedural, yaitu model yang memiliki langkah-langkah yang harus di tempuh secara berurutan. Langkah-langkah tersebut mencakup penyiapan tes, uji coba terbatas, dan uji coba diperluas.

##### **Penyiapan Tes**

Kegiatan penyiapan tes diawali dengan perumusan *learning continuum*, penyusunan kisi-kisi, dan penulisan soal untuk keterampilan berpikir historis. Selanjutnya, dilakukan penelaahan instrumen dengan melibatkan beberapa pakar. Perangkat tes yang dibuat seluruhnya ada enam perangkat. Keenam perangkat tersebut memiliki sepuluh item sebagai *anchor/common* item di dalamnya. Kegiatan uji coba terbatas dilakukan terhadap SMA tertunjuk, dengan melibatkan 1573 peserta didik, yang terdiri dari siswa kelas X dan XI.

**Tabel 1. Karakteristik SMA untuk Uji Coba Terbatas Tes Keterampilan Berpikir Historis**

| No. | Nama SMA            | Lokasi           | Kefavoritan Berdasarkan Lulusan yang Diterima PTN |
|-----|---------------------|------------------|---|
| 1   | SMA 1 Solok         | Kota Solok       | Favorit di Kota Solok                             |
| 2   | SMA 1 Payakumbuh    | Kota Payakumbuh  | Favorit di Kota Payakumbuh                        |
| 3   | SMA 1 Gunung Talang | Kab. Solok       | Favorit di Kab. Solok                             |
| 4   | SMA 1 Batu Sangkar  | Kab. Tanah Datar | Favorit di Kab. Tanah Datar                       |
| 5   | SMA 2 Solok         | Kota Solok       | Tidak favorit di Kota Solok                       |

### Hasil Kegiatan Uji Coba Terbatas

Penskoran dilakukan dengan poliotomus tiga kategori dengan skala 0-2. Data dianalisis menggunakan program QUEST. Hasilnya, ada dua item yang dinyatakan tidak *fit* dengan model, yakni nomor 23 dan nomor 24. Pada kedua item tersebut, tidak sampai sebagian testi mampu mencapai kategori-2 dan sedikit sekali testi yang berhasil mencapai kategori-3.

Reliabilitas menurut CTT, dalam bentuk indeks *alpha* Cronbach sebesar 0,65 tetap sama setelah dua item dikeluarkan dari analisis.

Sementara reliabilitas menurut IRT, yang diestimasi berdasarkan analisis testi (*case/person*), yang dinyatakan dalam bentuk indeks sparasi *person* sebesar 0,82.

Tabel 3 menunjukkan nilai rata-rata tingkat kesukaran item yang meningkat dari yang paling mudah ke yang paling sukar. Gradasi untuk aspek keterampilan dasar, adalah keterampilan berpikir kronologis, keterampilan mengidentifikasi kesinambungan dan perubahan, dan keterampilan menganalisis sebab akibat.

**Tabel 2. Hasil Estimasi Item dan Estimasi Testi Hasil Uji Coba Terbatas**

| No. | Uraian  | Sebelum Dua Item Data Dikeluarkan (I=111) |                      | Setelah Dua Item Data Dikeluarkan (I=109) |                      |
|-----|---|---|----------------------|---|----------------------|
|     |   | Estimasi Untuk Item                       | Estimasi Untuk Testi | Estimasi Untuk Item                       | Estimasi Untuk Testi |
| 1   | Nilai rata-rata dan simpangan baku              | 0,00 ± 1,08                               | -0,61 ± 0,86         | 0,00 ± 1,06                               | -0,58 ± 0,85         |
| 2   | Nilai rata-rata dan baku yang sudah disesuaikan | 0,00 ± 1,02                               | -0,61 ± 0,78         | 0,00 ± 1,00                               | -0,58 ± 0,77         |
| 3   | Indeks sparasi                                  | 0,89                                      | 0,82                 | 0,89                                      | 0,82                 |
| 4   | Nilai <i>alpha</i> Cronbach                     |   | 0,54                 |   | 0,54                 |
| 5   | Nilai rata-rata dan simpangan baku INFIT MNSQ   | 0,98 ± 0,10                               | 0,99 ± 0,47          | 0,98 ± 0,10                               | 0,99 ± 0,48          |
| 6   | Nilai rata-rata dan simpangan baku OUTFIT MNSQ  | 0,99 ± 0,15                               | 1,00 ± 0,51          | 0,98 ± 0,13                               | 1,00 ± 0,51          |
| 7   | Nilai rata-rata dan simpangan baku INFIT t      | -0,22 ± 1,06                              | -0,24 ± 1,09         | -0,19 ± 1,06                              | -0,24 ± 1,09         |
| 8   | Nilai rata-rata dan simpangan baku OUTFIT t     | -0,17 ± 1,07                              | -0,15 ± 1,05         | -0,14 ± 1,06                              | -0,14 ± 1,05         |
| 9   | Item atau testi skor 0                          | 0   | 0                    | 0   | 0                    |
| 10  | Item atau testi skor <i>perfect</i>             | 0   | 0                    | 0   | 0                    |

Untuk aspek keterampilan penelitian sejarah berturut-turut adalah keterampilan membangun arti penting sejarah, keterampilan merekam data atau informasi/sumber sejarah, keterampilan merancang penelitian sejarah, keterampilan melaporkan hasil penelitian sejarah dan keterampilan menggunakan dan menganalisis sumber-sumber sejarah.

Nilai rata-rata tingkat kesukaran item sub aspek keterampilan menggunakan dan menganalisis sumber-sumber sejarah ternyata tertinggi di antara sub aspek keterampilan ber pikir historis lainnya, sedangkan yang termudah pikir historis lainnya, sedangkan yang termudah adalah sub aspek keterampilan membangun arti penting

adalah sub aspek keterampilan membangun arti penting sejarah. Sebaran item berdasarkan tingkat kesukaran dalam bentuk nilai *difficulty* hasil analisis menggunakan program QUEST, menunjukkan 5,40% item keterampilan dasar yang tergolong agak sukar (dari 1,0 sampai <1,5) dan tidak ada item keterampilan dasar yang tergolong sukar (dari 1,5 sampai 2,0). Item keterampilan penelitian sejarah ternyata ada 5,40% yang tergolong agak sukar (dari 1,0 sampai <1,5), ada 1,3% yang tergolong sukar (dari 1,5 sampai 2,0), dan ada 1,35% tergolong sangat sukar ( $\geq 2,0$ ).

**Tabel 3. Nilai Tingkat Kesukaran Item Aspek dan Subaspek Keterampilan Berpikir Historis Menurut PCM Hasil Uji Coba Terbatas**

| No | Aspek dan Sub Aspek Keterampilan HT  | Nilai Tingkat Kesukaran Item |               |              |
|----|--|------------------------------|---------------|--------------|
|    |  | <i>Difficulty</i>            | <i>Delta</i>  |              |
| 1. | <b>Keterampilan dasar (<i>Basic Skill</i>)</b>   | <b>-0.989</b>                | <b>-2.677</b> | <b>0.697</b> |
|    | a. Keterampilan berpikir kronologis  | -1.776                       | -3.336        | -0.221       |
|    | b. Keterampilan mengidentifikasikesinambungan dan perubahan                                | -1.027                       | -2.673        | 0.618        |
|    | c. Keterampilan menganalisis sebab dan akibat  | -0.348                       | -2.190        | 1.492        |
| 2. | <b>Keterampilan/kemampuan penelitian sejarah (<i>historical research capabilities</i>)</b> | <b>0.508</b>                 | <b>-0.685</b> | <b>1.703</b> |
|    | a. Keterampilan membangun arti penting sejarah   | -0.450                       | -1.993        | 1.093        |
|    | b. Keterampilan merekam data/informasi/sumber sejarah                                      | 0.462                        | -0.862        | 1.788        |
|    | c. Keterampilan menggunakan dan menganalisis sumber-sumber sejarah                         | 0.917                        | -0.405        | 2.238        |
|    | d. Keterampilan merancang penelitian sejarah   | 0.689                        | -0.305        | 1.690        |
|    | e. Keterampilan melaporkan hasil penelitian sejarah  | 0.726                        | 0.112         | 1.340        |



**Tabel 4. Distribusi Item Aspek Keterampilan Berpikir Historis Berdasarkan Nilai Tingkat Kesukaran Hasil Uji Coba Terbatas**

| Kisaran Tingkat Kesukaran | Keterampilan Dasar<br>(Basic Skill) |                   | Keterampilan/Kemampuan Penelitian Sejarah (Historical Research Capabilities) |                   |
|---------------------------|-------------------------------------|-------------------|--|-------------------|
|                           | Frekuensi absolut                   | Frekuensi relatif | Frekuensi absolut  | Frekuensi relatif |
| < -2,0                    | 4                                   | 10,81%            | 0  | 0,00              |
| -2,0 sampai <-1,5         | 5                                   | 13,51%            | 0  | 0,00              |
| -1,5 sampai <-1,0         | 6                                   | 16,21 %           | 4  | 5,40%             |
| -1,0 sampai <-0,5         | 11                                  | 29,72%            | 3  | 4,05 %            |
| -0,5 sampai <0,0          | 5                                   | 13,51 %           | 9  | 12,16 %           |
| 0,0 sampai <0,5           | 3                                   | 8,10 %            | 16   | 21,62 %           |
| 0,5 sampai <1,0           | 2                                   | 5,40 %            | 23   | 31,08 %           |
| 1,0 sampai <1,5           | 1                                   | 2,70 %            | 16   | 21,62%            |
| 1,5 sampai <2,0           | 0                                   | 0,00%             | 2  | 2,70%             |
| ≥ 2,0                     | 0                                   | 0,00%             | 1  | 1,35 %            |
| <b>Jumlah</b>             | <b>37</b>                           | <b>100 %</b>      | <b>74</b>  | <b>100 %</b>      |

#### Hasil Kegiatan Uji Coba Diperluas

Ringkasan hasil estimasi item dan estimasi testi (*case/person*) data testi sebanyak (N) 2673 dan item (I) sebanyak 93, dengan level peluang sebesar 0,50, menggunakan Program QUEST disajikan pada Tabel 5. Tabel 5 menunjukkan bahwa, secara keseluruhan item dalam bentuk tes dinyatakan *fit* dengan model karena

memenuhi persyaratan *fit statistics* yang dipersyaratkan dalam program QUEST, yakni *fit* dengan model apabila item-item yang dianalisis memiliki nilai rata-rata INFIT MNSQ dan OUTFIT MNSQ mendekati 1,0 dengan simpangan baku 0,0 atau memiliki nilai rata-rata INFIT t dan OUTFIT t mendekati 0,0 dengan simpangan baku 1,0.

**Tabel 5. Hasil Estimasi Untuk Item (I) Keterampilan Berpikir Historis dan Estimasi Untuk Testi (N) Menurut Partial Credit Model pada Tahap Uji Coba Diperluas**

| No. | Uraian                                    | Estimasi untuk Item | Estimasi untuk Testi ( <i>case/person</i> ) |
|-----|---|---------------------|---|
| 1   | Rata-rata dan simpangan baku              | 0,00 ± 0,96         | -0,58 ± 0,71                                |
| 2   | Rata-rata dan baku yang sudah disesuaikan | 0,00 ± 0,93         | -0,58 ± 0,60                                |
| 3   | Indeks sparasi                            | 0,93                | 0,72  |
| 4   | Indeks <i>alpha</i> Cronbach              |                     | 0,41  |
| 5   | Rata-rata dan simpangan baku INFIT MNSQ   | 0,99 ± 0,05         | 0,99 ± 0,51                                 |
| 6   | Rata-rata dan simpangan baku OUTFIT MNSQ  | 0,99 ± 0,10         | 0,99 ± 0,56                                 |
| 7   | Rata-rata dan simpangan baku INFIT t      | -0,16 ± 1,05        | -0,25 ± 1,08                                |
| 8   | Rata-rata dan simpangan baku OUTFIT t     | -0,14 ± 1,04        | -0,16 ± 1,05                                |
| 9   | Item/ <i>case</i> skor 0                  | 0                   | 0   |
| 10  | Item/ <i>case</i> skor <i>perfect</i>     | 0                   | 0   |

Reliabilitas menurut CTT, dalam bentuk indeks *alpha* Cronbach sebesar 0,54. Sementara reliabilitas menurut IRT yang diestimasi berdasarkan

analisis testi (*case/person*), yang dinyatakan dalam bentuk indeks sparasi *person* sebesar 0,72.

**Tabel 6. Nilai Tingkat Kesukaran Item Aspek dan Subaspek Keterampilan Berpikir Historis pada Tahap Uji Coba Diperluas**

| No. | Aspek dan Sub Aspek Keterampilan HT  | Nilai Tingkat Kesukaran Item |               |              |
|-----|--|------------------------------|---------------|--------------|
|     |  | <i>difficulty</i>            | <i>delta</i>  |              |
|     |  |                              | 1             | 2            |
| 1.  | <b>Keterampilan Dasar (<i>Basic Skill</i>)</b>   | <b>-0.705</b>                | <b>-2.307</b> | <b>0.897</b> |
|     | a. Keterampilan berpikir kronologis  | -1.072                       | -2.641        | 0.488        |
|     | b. Keterampilan mengidentifikasikan sambungan dan perubahan                                | -0.698                       | -2.150        | 0.758        |
|     | c. Keterampilan menganalisis sebab dan akibat  | -0.420                       | -2.261        | 1.419        |
| 2.  | <b>Keterampilan/kemampuan penelitian sejarah (<i>historical research capabilities</i>)</b> | <b>0.369</b>                 | <b>-0.650</b> | <b>1.390</b> |
|     | a. Keterampilan membangun arti penting sejarah   | -0.13                        | -1.363        | 1.102        |
|     | b. Keterampilan merekam data/informasi/sumber sejarah                                      | 0.24                         | -1,00         | 1.481        |
|     | c. Keterampilan menggunakan dan menganalisis sumber-sumber sejarah                         | 0.461                        | -0.552        | 1.475        |
|     | d. Keterampilan merancang penelitian sejarah   | 0.643                        | 0.135         | 1.153        |
|     | e. Keterampilan melaporkan hasil penelitian sejarah  | 0.933                        | 0.178         | 1.691        |

Berdasarkan angka yang tersaji, ternyata hasil analisis item pada tahap uji coba diperluas sebagian besar sama dengan hasil ketika uji coba terbatas. Nilai rata-rata tingkat kesukaran item mulai dari keterampilan dasar dan keterampilan penelitian sejarah menunjukkan ada gradasi yang meningkat, yakni mulai dari yang paling mudah ke yang paling sukar.

Hal ini sama dengan hasil uji coba terbatas.

#### **Hasil Pengukuran Uji Coba Diperluas**

Hasil pengukuran menunjukkan kisaran skor mentah terendah 2 dan tertinggi 39, dari batas skor maksimum 50 (kategori-1 = skor 0, kategori-2 = skor 1, dan kategori-3 = skor 2).

**Tabel 7. Frekuensi Absolut dan Relatif Skor Terkonversi Keterampilan Berpikir Historis pada Kisaran dari -2,00 sampai 2,00 dengan Interval Kelas 0,5**

| No.           | Kelas Interval Skor Terkonversi | Frekuensi Absolut | Frekuensi Relatif | Frekuensi Kumulatif |
|---------------|---------------------------------|-------------------|-------------------|---------------------|
| 1             | Skor 0 (tidak dikalibrasi)      | 0                 | 0,00              | 0,00                |
| 2             | <-2,00                          | 46                | 1,72              | 1,72                |
| 3             | -2,00 s/d -1,50                 | 244               | 9,12              | 10,84               |
| 4             | >-1,50 s/d -1,00                | 321               | 12,00             | 22,84               |
| 5             | >-1,00 s/d -0,50                | 738               | 27,60             | 50,44               |
| 6             | >-0,50 s/d 0,00                 | 1166              | 43,62             | 94,06               |
| 7             | >0,00 s/d 0,50                  | 122               | 4,56              | 98,62               |
| 8             | >0,50 s/d 1,00                  | 28                | 1,04              | 99,66               |
| 9             | >1,00                           | 8                 | 0,29              | 100,00              |
| <b>Jumlah</b> |                                 | <b>2673</b>       | <b>100,00</b>     |                     |

Setelah dikalibrasi, kisaran skor terkonversi terendah -3,52 dan yang tertinggi 0,09 dari kisaran skor antara -4,00 sampai +4,00. Skor terkalibrasi kemudian dikelompokkan dengan interval skor terkonversi 0,5. Hasilnya menunjukkan bahwa banyak testi yang memperoleh skor terkonversi lebih besar dari 0,00 hanya 5,89%. Dengan demikian, bila batas 0,00 didudukkan sebagai skor tengah, ada 94,11% testi berada di bawah skor tengah. Dengan kata lain, sebagian besar testi tidak berhasil mengerjakan sampai batas 50% jawaban benar.

### Pembahasan

#### *Karakteristik Item pada Kegiatan Uji Coba Terbatas*

Hasil analisis data uji coba terbatas menurut *Partial Credit Model* menunjukkan adanya item-item yang memiliki nilai *delta-1* lebih tinggi dari nilai *delta-2*, namun item secara keseluruhan *fit* dengan model. Hal ini tidak bertentangan dengan teori yang mendukungnya, sebagaimana yang

dikemukakan oleh Wright & Master<sup>18</sup> bahwa sesuai dengan karakteristik analisis menurut PCM, dimungkinkan adanya item yang memiliki nilai *delta-1* lebih tinggi daripada *delta-2*, yang artinya kemampuan untuk naik dari tahapan ketegori-2 ke kategori-3 dimungkinkan dapat lebih rendah daripada kemampuan untuk naik dari tahapan kategori-1 ke kategori-2. Hasil analisis menunjukkan, bahwa di antara 111 item yang diujikan ada dua item yang tidak *fit* dengan *Partial Credit Model* (PCM), yakni item nomor 23 dan 24.

#### *Tingkat Kesukaran Item Tes*

Sub aspek keterampilan dasar, pertanyaan mengenai keterampilan menganalisis sebab dan akibat merupakan keterampilan yang paling sukar. Selanjutnya, secara berturut-turut diikuti keterampilan: (a) keterampilan mengidentifikasi kesinambungan dan perubahan, dan (b) keterampilan berpikir kronologis.

<sup>18</sup>Wright, B.D. & Masters, G.N. 1982. *Rating Scale Analysis*. Chicago: Mesa Press.

Keterampilan menganalisis sebab dan akibat merupakan keterampilan yang membutuhkan pemahaman mendalam dari peserta didik. Menganalisis sebab dan akibat harus diikuti oleh kemampuan peserta didik dalam hal menyajikan data sejarah secara sistematis, sehingga mudah dipahami sebab dan akibat sebuah peristiwa sejarah tertentu. Siswa dalam hal ini, bukan hanya menghafal fakta-fakta penyebab suatu peristiwa seperti yang terdapat di dalam buku-buku pelajaran dan penyampaian guru. Lebih jauh, siswa dituntut untuk menyajikan data-data dari berbagai sumber sebab-sebab dan akibat-akibat sebuah peristiwa sejarah tertentu. Di saat bersamaan, siswa juga dituntut untuk memilah-milah hasil sajian yang terdapat di dalam berbagai sumber sejarah yang mereka hadapi. Artinya, peserta didik tidak hanya menyarikan saja hasil-hasil pengamatannya akan tetapi, siswa juga dituntut menyajikan pengamatan/observasi mereka ke dalam berbagai bentuk sajian data misalnya tabel, bagan, peta sejarah, dan lain sebagainya.

Dalam hal keterampilan penelitian sejarah (*historical research capabilities*) sub-aspek yang memiliki tingkat kesukaran tertinggi adalah keterampilan melaporkan hasil penelitian sejarah. Hal ini bisa dipahami, melihat kurangnya penerapan hal tersebut di sekolah. Tingkat kesulitan pada aspek ini diikuti secara berturut-turut keterampilan merancang penelitian sejarah, keterampilan menggunakan dan menganalisis sumber-sumber sejarah, keterampilan merekam data/informasi/sumber sejarah, dan keterampilan membangun arti penting sejarah. Peserta didik

mengalami kesulitan untuk memikirkan tindakan alternatif jika kegiatan tersebut jarang dilakukan.

#### *Karakteristik Item Tes Uji Coba Diperluas*

Seluruh item yang digunakan pada tahap uji coba diperluas *fit* dengan model. Nilai rata-rata tingkat kesukaran item pada kegiatan uji coba terbatas, untuk aspek keterampilan dasar dan keterampilan tinggi berturut-turut se-besar -0,989 dan 0,508. Pada tahap uji coba diperluas urutan nilai rata-rata tingkat kesukaran berturut-turut sebesar -0,705 dan 0,369. Data tersebut menunjukkan adanya pola respons yang sama antara hasil uji coba terbatas dan hasil tahap uji coba diperluas, dan berdasarkan tingkat kesukarannya pun dapat dikatakan sama.

Nilai rata-rata tingkat kesukaran sub-aspek dari aspek keterampilan dasar pada kegiatan uji coba terbatas, mulai dari yang paling sukar berturut-turut yaitu dalam menganalisis sebab dan akibat -0,348 (subaspek c), mengidentifikasi kesinambungan dan perubahan -1,027 (subaspek b), dan dalam hal ber-pikir kronologis -1,776 (subaspek a). Nilai rata-rata tingkat kesukaran subaspek dari aspek keterampilan dasar pada tahap pengukuran mulai dari yang paling sukar yaitu dalam menganalisis sebab dan akibat -0,420 (subaspek c), mengidentifikasi kesinambungan dan perubahan -0,698 (subaspek b), dan dalam hal berpikir kronologis -1,072 (subaspek a). Dengan demikian, tidak ada perbedaan pola respons testi. Begitupun respons yang paling mudah masih sama, yakni dalam berpikir kronologis.

### *Hasil Ujian pada Tahap Uji Coba Diperluas*

Hasil ujian pada tahap uji coba diperluas menunjukkan bahwa skor keterampilan berpikir historis yang diperoleh dari 2673 testi tidak memuaskan, hanya 5,89% testi yang berada di atas batas nilai tengah. Kemungkinan yang melatarbelakangi hal ini diduga ada tiga faktor. Faktor pertama, keterampilan berpikir historis memang tidak diajarkan sepenuhnya secara terintegrasi dalam kegiatan pembelajaran pada setiap pokok bahasan yang memungkinkan. Dengan demikian peluang berlatih menjadi sangat kecil. Faktor kedua, keterampilan berpikir historis dalam pokok bahasan pembelajaran sejarah tidak dilaksanakan dalam strategi penerapan keterampilan berpikir historis untuk menemukan konsep, bukan untuk mengklarifikasi sebagai hasil hapalan fakta-fakta. Pembelajaran sejarah yang mengandalkan hapalan fakta-fakta dan konsep menjadikan siswa tidak bisa berpikir historis secara semestinya. Faktor ketiga, keterampilan berpikir historis mungkin sudah diajarkan sesuai dengan tuntutan KI dan KD dalam kurikulum 2013, namun peserta didik tidak pernah dibiasakan mengerjakan tes nonobjektif yang memungkinkan mereka memberikan banyak alternatif jawaban benar.

## **E. Penutup**

### **Simpulan**

Berdasarkan hasil penelitian dan pembahasan, dapat ditarik beberapa simpulan sebagai berikut: (1) model penilaian yang dikembangkan termasuk ke dalam model prosedural, (2) informasi yang diperoleh dari model

penilaian kemampuan berpikir historis yang dikembangkan adalah diperolehnya susunan *learning continuum* keterampilan berpikir historis, karakteristik item berupa tingkat kesukaran butir (*item difficulty*) dan kemampuan testi ( $\theta$ ), dan item-item tes yang telah memperoleh bukti empiris *fit* dengan *Partial Credit Model* (PCM) berdasarkan data politomus tiga kategori, (3) validitas instrumen tes kemampuan berpikir historis yang disusun sudah dipenuhi melalui *expert judgement*, dan secara empiris terbukti *fit* dengan *Partial Credit Model* (PCM) berdasarkan data politomus tiga kategori, (4) reliabilitas instrumen tes kemampuan berpikir historis dalam bentuk indeks *alpha* Cronbach cukup baik yakni sebesar 0,64, (5) hasil penilaian secara keseluruhan menunjukkan bahwa, testi belum menguasai keterampilan berpikir historis yang diujikan, hal ini terlihat dari hanya 5,89% testi yang berada di atas nilai tengah harapan berdasarkan data politomus tiga kategori menurut *Partial Credit Model* (PCM). Ini disebabkan oleh karena peserta didik kemungkinan besar kurang dilatih melaksanakan keterampilan berpikir historis untuk menemukan konsep dan tidak dibiasakannya mengerjakan tes bentuk uraian nonobjektif.

### **Saran**

Berdasarkan simpulan penelitian, beberapa saran dapat dirumuskan sebagai berikut: (1) penelitian ini hanya menggunakan SMA negeri sebagai sampel, oleh karena itu disarankan penelitian lebih lanjut dapat menggunakan sampel yang lebih luas sehingga dapat diketahui

penguasaan keterampilan sejarah secara lebih luas pada jenjang tersebut. Penelitian lanjutan dapat pula dikembangkan pada jenjang SD/MI, SMP/MTs, bahkan pada jenjang perguruan tinggi, (2) pengukuran pada 13 SMA tertunjuk hanya untuk mengetahui keberfungsian tes yang sudah dikembangkan, yang telah memiliki bukti empirik. Oleh karena itu perlu dilakukan penelitian lebih lanjut untuk mengetahui penguasaan keterampilan berpikir historis untuk pembandingan antar tempat ataupun antar tahun dengan sampel yang representatif. Penelitian lebih lanjut juga dapat untuk mengetahui keterkaitan antara penguasaan keterampilan berpikir historis dengan strategi yang dikenakan dalam pembelajaran mata pelajaran sejarah, (3) melihat rendahnya penguasaan keterampilan berpikir historis mata pelajaran sejarah SMA pada 13 SMA tertunjuk, disarankan agar para guru di lapangan melatih peserta didik melalui pembelajaran yang cocok untuk mengembangkan keterampilan berpikir historis untuk mendampingi pembelajaran yang selama ini lebih berfokus pada pola-pola menghafal. Dengan demikian, akan terjadi pengembangan pola berpikir historis bukan menghafal sejarah, (4) Rendahnya penguasaan keterampilan berpikir historis dalam mata pelajaran sejarah SMA dapat disebabkan oleh rendahnya penguasaan kemampuan berpikir historis itu sendiri di SMA. Oleh karena itu para guru di lapangan hendaknya melatih keterampilan berpikir historis secara terpadu dalam setiap kegiatan pembelajaran sesuai dengan karakteristik pokok bahasan. Dengan demikian, peserta didik akan membiasakan diri menemukan fakta,

konsep, dan teori dengan menggunakan keterampilan berpikir historis sebagaimana yang dilakukan para pakar dalam bidang sejarah pada khususnya maupun bidang ilmu pengetahuan sosial pada umumnya, (5) Pengukuran keterampilan berpikir historis dalam mata pelajaran sejarah SMA yang dilaksanakan pada 13 SMA tertunjuk yang berada di Provinsi Sumatera Barat pada tahap pengukuran/uji coba diperluas keterampilan berpikir historis diharapkan dapat diperluas oleh pihak terkait pada SMA lain di provinsi tersebut, juga pada SMA di provinsi lainnya, (6) Mengingat tahun demi tahun peserta didik selalu berganti, maka secara periodik penguasaan keterampilan berpikir historis di SMA perlu diukur agar dapat diketahui tingkat penguasaan peserta didik pada tahun yang bersangkutan, (7) Guru hendaknya dapat memanfaatkan mekanisme *assessment for learning* dengan memanfaatkan hasil pengukuran keterampilan berpikir historis yang dilaksanakan di SMA bersangkutan sehingga dapat dipakai sebagai pijakan untuk memperbaiki mutu rancangan pelaksanaan pembelajaran maupun untuk memberikan remediasi kepada peserta didik, (8) Perlu adanya penghargaan dan iklim yang kondusif oleh pihak terkait guna mendorong guru melakukan ujian dengan menggunakan soal bentuk uraian terbuka guna merangsang perkembangan keterampilan berpikir historis peserta didik, dan (9) Guru di lapangan perlu menyadarkan peserta didik akan pentingnya mengenal berbagai bentuk tes agar memiliki wawasan yang lebih luas dan memahami permasalahan yang terkandung dalam status jenis item tes.

Simpulan dapat bersifat generalisasi temuan sesuai permasalahan penelitian, dapat pula berupa rekomendasi untuk langkah selanjutnya. Saran

dapat berupa masukan bagi peneliti berikutnya, dapat pula rekomendasi implikatif dari temuan penelitian

## DAFTAR KEPUSTAKAAN

- Allen, M.J. & Yen, W.M. 1979. *Introduction to Measurement Theory*. Belmont, CA: Wadsworth, Inc.
- Ashby, R., Lee, P.J. & Shemit, D. 2005. "Putting principles into practice: teaching and planning". Dalam M.S. Donovan and J.D. Bransford (Eds.). *How Students Learn: History, Mathematics, and Science in the Classroom*. Washington, DC: The National Academies Press.
- Bain, R. B. 2005. "Applying the principles of how people learning teaching high school history". Dalam M.S. Donovan and J.D. Bransford (Eds.). *How Students Learn: History, Mathematics, and Science in the Classroom*. Washington, DC: The National Academies Press.
- Barton, K.C. & Levstik, L.S. 2003. "Why don't more history teachers engage students in interpretation?". *Research and Practice Social Education*, 67 (6), 358-361.
- Borg, W.R. & Gall, M.D. 1989. *Educational Research: An Introduction*. New York: Longman.
- Depdiknas. 2007. *Peraturan Menteri Pendidikan Nasional Republik Indonesia Nomor 20, Tahun 2007, tentang Standar Penilaian Pendidikan untuk Satuan Pendidikan Dasar dan Menengah*. Jakarta: Depdiknas.
- Fogu, C. 2009. "Digitalizing historical consciousness". *Journal History and Theory*, 47 (1), 103-121.
- Griffin, P. & Nix, P. 1991. *Educational Assessment and Reporting: A New Approach*. Sydney: Harcourt Brace Jovanovich, Publishers.
- Hambleton, R.K. & Swaminathan, H. 1985. *Item Respons Theory*. Boston, MA: Kluwer Inc.
- Hargreaves, A., Earl, L. & Schmidt, M. 2002. "Perspectives on alternative assesment reform". *American Educational Research Journal*, 39 (1), 69-95.
- Keeves, J.P. & Master, G.N. 1999. "Introduction". Dalam G.N. Masters & J.P. Keeves (Eds.). *Advances in Measurement in Education Research and Assessment*. Amsterdam: Pergamon, An imprint of Elsevier Science.
- Lee, P. 2005. "Putting principles into practice: understanding history". Dalam M.S. Donovan & J.D. Bransford (Eds.). *How Students Learn: History, Mathematics, and Science in the Classroom*. Washington, DC: The National Academies Press.

- Mardapi, D. 1999. "Estimasi kesalahan pengukuran dalam bidang pendidikan dan implikasinya pada ujian nasional". Disampaikan pada *Pidato Pengukuhan Guru Besar, tanggal 4 Mei 1999*. Yogyakarta: Universitas Negeri Yogyakarta.
- Mardapi, D. 2008. *Teknik Penyusunan Instrumen Tes dan Nontes*. Yogyakarta: Mitra Cendikia Press.
- Masters, G.N. 1999. "Partial credit model". Dalam J.P. Keeves & G.N. Masters (Eds.). *Advances in Measurement in Educational Research and Assessment*. Amsterdam: Pergamon.
- Oriundo, L.L. & Dallo-Antonio. 1998. *Evaluating Educational Outcomes (test, measurement, and evaluation)*. Quezon City: REX Printing Company, Inc.
- Rasch, G. 1961. "On general laws and the meaning of measurement in psychology". *The Danish Yearbook of Philosophy*, 4 (1), 321-334.
- Rasch, G. 1977. "On Specific Objectivity: An attempt at formalizing the request for generality and validity of scientific statements". *The Danish Yearbook of Philosophy*, 14 (3), 58-93.
- Seixas, P. & Peck, C. 2004. "Teaching historical thinking". Dalam A. Sears & I. Wright (Eds.), *Challenges and Prospects for Canadian Social Studies*. Vancouver: Pacific Educational Press.
- Seixas, P. 2013. *Linking Historical Thinking Concepts, Content and Competencies*. Vancouver: Pacific Educational Press.
- Van der Linden, W.J. & Hambleton, R.K. 1997. *Handbook of Modern Item Response Theory*. New York: Springer.
- Winerburg, S. 2006. *Berpikir historis: Memetakan Masa Depan, Mengajarkan Masa Lalu*. (Terjemahan Masri Maris). Jakarta: Yayasan Obor Indonesia.
- Wright, B.D. & Masters, G.N. 1982. *Rating Scale Analysis*. Chicago: Mesa Press.