OPTIMALISASI UJI TINGKAT KOMPETENSI DI SMK UNTUK MENINGKATKAN *SOFT SKILL* LULUSAN*)

Oleh :

Badrun Kartowagiran)**

UNIVERSITAS NEGERI YOGYAKARTA

Mei, 2014

***) Makalah disiapkan dalam ranka Dies UNY ke 50**

****) Dosen UNY**

THE OPTIMILIZATION OF COMPETENCE LEVEL TEST IN SMK TO IMPROVE GRADUATES' SOFT SKILL

By : Badrun Kartowagiran

abstract

The competence level test (Uji tingkat kompetensi = UTK) is a competency test in certain level conducted by educational units at the end of second grade (level 1), grade IV (level 2), grade VIII (level 4), and XI (level 5), by using the lattice compiled by the Government . This test will give more tasks to the educational unit, so educational unit will be overloaded. If there is no special attention, it is possible UTK becomes just a label and it will not be done seriously. UTK conducted with the lowest quality so its product can not be fully utilized . Therefore, UTK optimilization needs to be done.

UTK in Vocational High School (Sekolah Menengah Kejuruan = SMK), is implemented in grade XII (level 6) and carried out by the UN, which means it is also the end of the competency level test. The optimilization of UTK is a great effort to construct instruments and to conduct UTK in a high-quality so that the graduates qualification meets the needs of the nation in the future, especially Indonesian Gold. Graduates have the knowledge and skills appropriate with the competency standards and have soft skills.

An attempt to produce a high -quality UTK question is done by tightening the content validity , predictive validity, and the reliability. Tightening the content validity can be done by using more than 2 raters then calculated with Aiken formula or may use Lawshe formula. Tightening the reliability of the tests can be done by using more than 3 raters with ordinal scales then analyzed using a model of the ICC. Or, if there are two raters and two categories of choice, the interrater reliability is calculated by Cohens Kappa formula. When there are three raters or more and two categories of choices, the reliability coefficient is calculated by using Fleis Kappa formula. Another attempt to optimilize UTK is by implementing it as well as possible, ie: honest, fair, and accountable. Students who pass are those who have the appropriate competence standards and will not pass if the student has not reached the competency criteria that have been determined . In this way students will be more resilient, working harder, learning more earnestly, more appreciative of the time, and more responsibile. This means that the optimilization of UTK can improve the graduates' soft skills.

Keywords : Optimization of UTK and its effect on graduates' competencies

OPTIMALISASI UJI TINGKAT KOMPETENSI DI SMK UNTUK MENINGKATKAN *SOFT SKILL* LULUSAN

Oleh: Badrun Kartowagiran

Abstrak

Uji tingkat kompetensi merupakan uji kompetensi pada tingkat tertentu yang dilakukan oleh satuan pendidikan pada akhir kelas II (tingkat 1), kelas IV (tingkat 2), kelas VIII (tingkat 4), dan kelas XI (tingkat 5), dengan menggunakan kisi-kisi yang disusun oleh Pemerintah. Kegiatan ini menambah tugas bagi satuan pendidikan, sehingga beban sekolah terlalu besar. Apabila tidak ada perhatian khusus, tidak menutup kemungkinan UTK hanya sekedar label, bahkan asal jalan. UTK dilaksanakan sekadarnya sehingga hasilnya tidak dapat dimanfaatkan secara maksimal. Oleh karenanya optimalisasi UTK perlu dilakukan.

Untuk SMK, UTK dilaksanakan pada kelas XII (tingkat 6) dan dilakukan melalui UN yang berarti juga merupakan uji kompetensi akhir jenjang. Optimalisasi UTK adalah usaha keras agar instrumen dan pelaksanaan UTK berkualitas tinggi sehingga kualifikasi lulusannya sesuai dengan kebutuhan bangsa di masa datang, khususnya Indonesia Emas. Lulusannya memiliki kompetensi pengetahuan dan keterampilan sesuai standar yang telah ditentukan dan memiliki *soft skill*.

Upaya untuk menghasilkan soal UTK yang berkualitas tinggi dilakukan dengan cara memperketat validitas isi, validitas prediktif, dan memperketat reliabilitas soal UTK. Pengetatan validitas isi dilakukan dengan cara menggunakan lebih dari 2 rater kemudian dihitung dengan formula Aiken atau menggunakan formula Lawshe. Pengetatan reliabilitas soal tes menggunakan 3 rater dengan skala ordinal kemudian dianalisis dengan menggunakan model ICC. Atau bila raternya ada dua dan pilihannya dua katagori, maka reliabilitas antara raternya dihitung dengan persamaan Cohen Kappa. Bila raternya lebih dari dua dan pilihannya berbentuk katagori maka reliabilitas antar raternya dihitung dengan koefisien Fleis Kappa. Usaha lain untuk mengoptimalkan UTK adalah mengusahakan agar UTK dilaksanakan sebaik mungkin, yakni: jujur, adil, dan dapat dipertanggungjawabkan. Siswa yang dinyatakan lulus adalah mereka yang memiliki kompetensi sesuai standar dan tidak akan lulus bila kompetensi siswa belum mencapai kriteria yang telah ditentukan. Dengan cara demikian siswa akan lebih ulet, bekerja lebih keras, belajar lebih sungguh-sungguh, lebih menghargai waktu, dan lebih tanggung jawab. Ini berarti bahwa optimalisasi UTK dapat meningkatkan *soft skill* lulusan.

Kata Kunci: Optimalisasi UTK dan dan dampaknya pada kompetensi lulusan

PENDAHULUAN

Ada ungkapan menarik yang disampaikan Bahrul Hayat, Ketua Himpunan Evaluasi Pendidikan Indonesia (HEPI), pada saat memberikan sambutan pada acara pelantikan pengurus HEPI JABOTABEK pada tanggal 8 Maret 2014. Dia mengatakan bahwa kurikulum 2013 ini memiliki dua kelemahan, yaitu: kelemahan pada aspek idealistik dan kelemahan pada aspek praksis. Pada aspek idealistik, kurikulum tahun 2013 selalu memuat materi hari ini ke belakang; tidak ada mata pelajaran hari ini ke depan. Oleh karena itu, kita selalu tertinggal dalam menyusun kurikulum. Lebih jauh Bahrul Hayat menjelaskan, kelemahan aspek praksis adalah cara melakukan penilaian atau asesmen. Dengan pendekatan tematik integratif, seorang guru bisa mengajarkan banyak hal dalam satu waktu, tetapi sewaktu menilai, guru harus spesifik substansi yang dinilai.

Keraguan Bahrul Hayat ini dapat difahami karena lapangan menunjukkan bahwa sebagian besar sekolah belum siap mengimplementasikan kurikulum tahun 2013 ini. Penelitian Badrun Kartowagiran (2013) terhadap 15 SMP di Daerah Istimewa Yogyakarta (DIY) menunjukkan bahwa sebagian besar (82%) belum siap mengimplementasikan kurikulum tahun 2013.

Ungkapan Bahrul Hayat dan hasil penelitian Badrun Kartowagiran di atas harus ditanggapi secara positif. Pemerintah dan masyarakat harus sadar bahwa untuk dapat mengimplementasikan kurikulum tahun 2013, termasuk melakukan penilaian dengan baik masih diperlukan kerja keras. Masih diperlukan upaya-upaya tambahan agar implementasi kurikulum tahun 2013 berjalan lancar. Sementara itu, satuan pendidikan juga harus menyelenggarakan ujian sekolah dan uji tingkat kompetensi (UTK) dengan menggunakan kisi-kisi yang disusun oleh Pemerintah. UTK dilakukan oleh satuan pendidikan pada akhir kelas II (tingkat 1), kelas IV (tingkat 2), kelas VIII (tingkat 4), dan kelas XI (tingkat 5), Ujian tingkat kompetensi pada akhir kelas VI (tingkat 3), kelas IX (tingkat 4A), dan kelas XII (tingkat 6). UTK pada akhir kelas VI (tingkat 3), kelas IX (tingkat 4A) dan kelas XII (tingkat 6) dilakukan melalui UN (Permendikbud R.I. nomor 66 Tahun 2013 tentang Standar Penilaian).

Uraian di atas memberi gambaran bahwa tugas satuan pendidikan saat ini sangat padat karena harus menyelenggarakan ujian sekolah dan uji tingkat kompetensi (UTK). Apabila tidak

ada perhatian khusus, tidak menutup kemungkinan UTK hanya sekedar label, bahkan asal jalan. UTK dilaksanakan sekadarnya dan akhirnya hasil tidak dapat dimanfaatkan secara maksimal. Oleh karenanya optimalisasi UTK perlu dilakukan.

OPTIMALISASI UJI UTK DI SMK

Optimalisasi UTK adalah usaha memaksimumkan hasil UTK dan pemanfaatannya. Jangan sampai hasil UTK tidak tepat sehingga tidak dapat dimanfaatkan. Atau, hasil UTK tepat namun karena tidak kontekstual maka hasilnya tidak dapat dimanfaatkan secara optimal. Agar hasil UTK tepat maka ada dua hal yang harus diusahakan, yakni: kualitas soal yang digunakan dan kualitas pelaksanaan UTK harus tinggi. Ini berarti bahwa optimalisasi UTK dapat tercapai manakala instrumen dan pelaksanaan UTK berkualitas tinggi sehingga kualifikasi lulusannya sesuai dengan kebutuhan bangsa di masa datang, khususnya Indonesia Emas. Lulusannya memiliki kompetensi pengetahuan dan keterampilan sesuai standar dan memiliki *soft skill*.

1. Kualitas Instrumen (soal UTK)

Syarat instrumen yang baik adalah instrumen yang memiliki validitas dan reliabilitas tinggi atau memenuhi persyaratan psikometrik. Menurut pendekatan teori tes klasik, validitas suatu alat ukur adalah sejauhmana alat ukur itu mampu mengukur apa yang seharusnya diukur (Nunnally, 1978). Sejauhmana besaran skor tampak (X) mendekati besaran skor murni (T), semakin jauh perbedaan antara skor tampak dan skor murni berarti semakin kecil validitas alat ukur tersebut.

Instrumen yang baik juga harus memiliki reliabilitas tinggi, yakni memiliki keajegan atau kestabilan hasil pengukuran. Alat ukur yang reliabel adalah alat ukur yang mampu membuahkan hasil pengukuran yang stabil (Lawrence, 1994). Dalam ilmu sosial hal ini sulit sekali terjadi karena banyak faktor yang mempengaruhinya. Jika dilakukan pengukuran pada kelompok yang sama dua kali secara berurutan, beberapa variasi skor dapat terjadi karena adanya fluktuasi pada memori sesaat, perhatian, kelelahan, ketegangan emosional, tebak-tebak dan sejenisnya. Sebaliknya jika dilaksanakan dalam waktu yang lama antara tes pertama dan tes kedua variasi skor kemungkinan disebabkan oleh pengaruh pengalaman belajar, perubahan kesehatan, lupa dan lain-lain. Variasi skor juga mungkin terjadi jika hasil tes uraian dikoreksi oleh orang yang berbeda atau pengukuran kinerja siswa dilakukan oleh orang yang berbeda. Variasi skor juga

akan terjadi jika digunakan sampel tugas yang berbeda dari domain yang sama. Adanya variasi hasil pengukuran ini menunjukkan adanya kesalahan pengukuran.

Hal senada disampaikan Nunnally (1978) yang mengatakan bahwa banyak faktor yang mempengaruhi ketepatan pengukuran. Jenis dan jumlah penyebab kesalahan ini tergantung pada karakteristik tes dan bagaimana tes itu digunakan. Hal penting yang perlu diperhatikan adalah harus dibedakan antara kesalahan pengukuran yang menyebabkan variasi penampilan dari butir-kebutir dalam suatu tes dan kesalahan yang dimanifestasikan dalam variasi penampilan dalam bentuk tes berbeda diberikan pada waktu sama atau berbeda waktunya. Kesalahan tipe pertama dikarenakan sampling butir. Semakin banyak butir yang diambil semakin berkurang kesalahannya, asalkan korelasi antara butir yang satu dengan lainnya tinggi atau korelasi antara skor butir dengan skor keseluruhan itu tinggi. Kesalahan tipe kedua dikarenakan tingkat paralelisme dua tes yang digunakan, semakin tinggi kualitas indikator semakin kecil kemungkinannya kesalahan tipe kedua muncul.

Menurut para ahli (Nunnally, 1978, Allen & Yen, 1979, Fernandes, 1984, Woolfolk & McCane, 1984, dan Lawrence, 1994), validitas dapat dikelompokkan menjadi tiga tipe, yaitu: (1) validitas kriteria, (2) validitas isi, dan (3) validitas konstruk. Validitas kriteria dibedakan menjadi dua, yaitu validitas prediktif dan validitas konkuren. Fernandes (1984) mengatakan validitas berdasarkan kriteria dimaksudkan untuk menjawab pertanyaan: *"How well test performance predicts future performance (predictive validity) or estimate current performance on some valued measure other than the test itself (concurrent validity)?"*. Senada hal ini, Nunnally (1978) berpendapat validitas prediktif diestimasi manakala instrumen dimaksudkan sebagai prediktor bagi performansi di waktu yang akan datang. Sementara itu instrumen dikatakan memiliki validitas konkuren tinggi bila skor hasil pengukuran dengan instrumen yang dikembangkn berkorelasi tinggi dengan skor hasil pengukuran menggunakan instrumen yang sudah valid. Dalam analisis validitas prediktif dan konkuren, performansi yang hendak diprediksikan disebut dengan kriteria. Besar kecilnya harga estimasi validitas prediktif atau konkuren suatu instrumen digambarkan dengan keofisien korelasi antara prediktor dengan kriteria tersebut.

Validitas isi suatu instrumen adalah sejauhmana butir-butir dalam instrumen itu mewakili komponen-komponen dalam keseluruhan kawasan isi objek yang hendak diukur (aspek representasi) dan sejauh mana butir-butir itu mencerminkan ciri perilaku yang hendak diukur

(aspek relevansi) (Fernandes, 1984; Nunnally, 1978). Validitas konstruk adalah validitas yang menunjukkan sejauhmana instrumen mengungkap suatu trait atau konstruk teoritik yang hendak diukurnya (Saifuddin Azwar, 2013; Allen & Yen, 1979; Nunnally, 1978). Pengujian validitas konstruk merupakan proses yang terus berlanjut sejalan dengan perkembangan konsep ***trait*** yang akan diukur. Perubahan dan perkembangan konsep seperti ini merupakan hal biasa dalam bidang psikologi karena variabel itu pada dasarnya merupakan konsep hipotetik yang tidak selalu mudah untuk dioperasionalkan.

Konsep validitas konstruk sangat bermanfaat pada tes yang mengukur trait yang tidak memiliki kriteria eksternal. Untuk itu prosedur validasi konstruk diawali dari suatu identifikasi dan batasan mengenai variabel yang hendak diukur dan dinyatakan dalam bentuk konstruk logis berdasarkan teori mengenai variabel tersebut. Dari teori ini ditarik suatu konskuensi praktis mengenai hasil pengukuran pada kondisi tertentu, dan konskuensi inilah yang akan diuji. Apabila hasilnya sesuai dengan harapan maka instrumen itu dianggap memiliki validitas konstruk yang baik.

Uji Tingkat Kompetensi (UTK) merupakan uji kompetensi pada tingkat tertentu, oleh karenanya kualifikasi lulusan UTK harus dikaitkan dengan Kerangka Kualifikasi Nasional Indonesia (KKNI). Menurut Pasal 5 Perpres R.I. Nomor 8 Tahun 2012 Tentang Kerangka Kualifikasi Nasional Indonesia, lulusan pendidikan menengah paling rendah setara dengan jenjang 2. Selanjutnya dalam Perpres itu dijelaskan bahwa lulusan yang memiliki kemampuan setingkat jenjang 2 itu harus memiliki kemampuan dan tanggung jawab sebagai berikut.

1. Mampu melaksanakan satu tugas spesifik, dengan menggunakan alat, dan informasi, dan prosedur kerja yang lazim dilakukan, serta menunjukkan kinerja dengan mutu yang terukur, di bawah pengawasan langsung atasannya.
2. Memiliki pengetahuan operasional dasar dan pengetahuan faktual bidang kerja yang spesifik, sehingga mampu memilih penyelesaian yang tersedia terhadap masalah yang lazim timbul.
3. Bertanggung jawab pada pekerjaan sendiri dan dapat diberi tanggung jawab membimbing orang lain.

KKNI jenjang 2 seperti yang dijelaskan di atas merupakan Standar Kompetensi Lulusan (SKL) bagi lulusan SMK beberapa tahun mendatang. Selanjutnya SKL ini digunakan sebagai acuan dalam mengembangkan standar isi, standar proses, dan standar penilaian. Ini berarti

bahwa cakupan materi yg diujikan dalam UTK adalah materi yang harus diberikan agar tujuan UTK yang setara dengan KKNI jenjang 2 ini dicapai. Proses pembelajaran yang harus dilakukan di SMK adalah kegiatan-kegiatan untuk mempelajari pengetahuan, dan atau berlatih keterampilan, dan atau berlatih mengamalkan sikap spiritual dan sikap sosial sedemikian rupa agar SKL tercapai. Penilaian kompetensi lulusan atau UTK harus menggunakan berbagai teknik penilaian sehingga mampu mengungkap kompetensi pengetahuan, kompetensi keterampilan, dan kompetensi sikap dari peserta UTK.

Uji tingkat kompetensi (UTK) termasuk tes prestasi belajar, maka teknik validasi yang paling tepat adalah validitas isi. UTK memiliki validitas isi manakala materi yang diujikan mewakili komponen-komponen yang ada dalam KKNI jenjang 2. UTK harus mencakup uji pengetahuan atau Teori Kejuruan dan Praktik Kejuruan. Butir-butir UTK juga harus mampu mendorong munculnya perilaku yang ada dalam KKNI jenjang 2, misal cermat, tanggung jawab, dan jujur.

Untuk memastikan apakah UTK memiliki validitas isi atau tidak dapat dilakukan dua langkah, yakni mencermati validitas tampang dan mencermati validitas logis. Sesuai dengan namanya, validitas tampang adalah validitas instrumen yang didasarkan pada penilaian pakar terhadap tes itu. Menurut pakar apakah format tes itu sudah layak dan butir-butir dalam instrumen itu sudah mengukur apa yang seharusnya diukur. Bila ya, maka dikatakan bahwa tes itu memiliki validitas tampang yang baik.

Setelah tes itu memenuhi validitas tampang, selanjutnya tes itu dicek validitas logiknya. Pada dasarnya, validitas logik adalah sejauhmana butir-butir tes itu representatif mewakili semua materi, pengetahuan, keterampilan, dan perilaku yang akan diukur. Agar mudah memilih butir yang mewakili atribut yang akan diukur, tes harus dirancang secermat mungkin. Rancangan ini dapat berupa tabel spesifikasi yang berisi tujuan tes dan kisi-kisi tes. Kisi-kisi merupakan panduan penulisan bahan ajar, dan panduan penyusunan butir-butir soal. Kisi-kisi memuat kompetensi inti, kompetensi dasar, dan indikator pencapaian. Butir-butir soal ditulis mengacu pada indikator pencapaian.

Untuk memantapkan kecermatan validitas isi, butir-butir soal tadi dinilai ketepatannya oleh lebih dari satu pakar penilai (panel). Para penilai ini memberikan penilaian terhadap setiap butir tes, yakni sejauhmana butir-butir tes itu representatif mewakili materi pengetahuan,

keterampilan, dan perilaku yang akan diukur. Penilaian dilakukan dengan cara memberikan skor 1 (sangat tidak mewakili atau sangat tidak relevan) sampai dengan 5 (sangat mewakili atau sangat relevan). Selanjutnya digunakan persamaan V dari Aikens (Saifuddin Azwar, 2013):

$V = \sum s /[n(c-1)]$

$S = r – lo$

$\sum s = s1 + s2 + dst$

Lo = angka penilaian validitas yang terendah

c = angka penilaian validitas yang tertinggi

r = angka yang diberikan oleh seorang penilai

Selain menggunakan persamaan Aiken, validitas isi juga dapat diestimasi menggunakan rumus Lawshe, yakni *content validity ratio* (CVR) diteruskan dengan ke *Content Validity Index* (CVI). CVR adalah validitas isi dari suatu butir menurut penilaian para ahli yang disebut dengan *Subject Matter Experts* (SME). Penilaian SME terhadap suatu butir bergradasi, yakni: *esensial, berguna tetapi tidak esensial*, dan *tidak diperlukan*. Suatu butir dianggap memiliki validitas isi tinggi manakala butir itu esensial bagi operasionalisasi konstruk teoritik tes yang disusun. Rumus CVR yang dimaksudkan adalah sebagai berikut (Saifuddin Azwar, 2013).

$CVR = [(2ne/ n) – 1]$

ne = banyaknya SME yang menilai suatu butir tes itu ***esensial***

n = banyaknya SME yang melakukan penilaian

Sebagai contoh, suatu butir dinilai tingkat esensialnya oleh 10 penilai (SME); enam penilai menyatakan bahwa butir itu *esensial*, tiga penilai menyatakan butir itu *berguna tetapi tidak esensial*, dan 1 penilai menyatakan bahwa butir itu *tidak diperlukan*. Dengan demikian $CVR = [(2.6)/10 -1] = 0,20$

Angka CVR bergerak dari -1,00 sampai dengan +1,00 bila harga CVR positif atau > 0 maka 50% SME menilai butir itu esensial. Semakin tinggi harga CVR, semakin baik validitas isi butir itu. Dalam hal ini butir dikatakan memiliki validitas baik bila $CVR \geq 0,3$. Sementara itu validitas isi suatu tes atau *Content Validity Index* (CVI) adalah rata-rata dari CVR semua butir, sehingga $CVI = (\sum CVR)/k$; k = jumlah butir dalam tes. Dalam hal ini, tidak semua butir dapat

dimasukkan dalam rumus CVI, namun hanya butir-butir tepilih atau butir yang memiliki harga CVR ≥ 0,3. Hal ini dapat difahami karena sebaiknya tes itu terdiri dari butir-butir yang baik.

Selain memiliki validitas, butir-butir soal UTK juga harus memiliki karakteristik, misal tingkat kesulitan dan daya beda yang baik. Karakteristik butir-butir soal UTK dapat dihitung menurut pendekatan teori tes klasik dan/atau teori respon butir. Tulisan ini hanya membatasi pada pendekatan teori tes klasik, karena pendekatan ini yang lebih murah dan lebih mudah dilaksanakan.

Menurut pendekatan teori tes klasik karakteristik butir meliputi tingkat kesukaran (p), daya pembeda (d), dan efektivitas distraktor. Selain itu, dengan analisis kuantitatif pendekatan teori klasik juga dapat diketahui reliabilitas soal tes, dan kesalahan baku pengukuran. Untuk melihat tingkat kesukaran, daya pembeda, dan efektivitas distraktor dilakukan analisis setiap butir tes, sedangkan reliabilitas dan kesalahan pengukuran baku dapat dilihat dengan cara menganalisis soal tes secara keseluruhan. Tingkat kesukaran (p) dapat diperoleh dengan beberapa cara, antara lain: (1) skala kesukaran linier; (2) skala bivariat; (3) indeks Davis; dan (4) proporsi menjawab benar. Cara yang paling mudah dan paling banyak digunakan adalah skala rata-rata atau proporsi menjawab benar atau *proportion correct* (p), yaitu jumlah peserta tes yang menjawab benar pada butir yang dianalisis dibandingkan dengan peserta tes seluruhnya.

Tingkat kesukaran *(p)* mengandung banyak kelemahan, antara lain tingkat kesukaran sebenarnya merupakan ukuran kemudahan butir karena semakin tinggi indeks *p*, semakin mudah butir tersebut. Sebaliknya semakin rendah *p* semakin sulit. Oleh karenanya ada beberapa ahli pengukuran yang menyebut tingkat kesukaran ini dengan tingkat kemudahan. Tingkat kesukaran merupakan salah satu parameter butir soal, yang disimbolkan (P_i), yakni rasio antara jawaban benar dan banyaknya penjawab butir soal.

Besarnya tingkat kesukaran berkisar antara nol dan satu. Suatu butir kadang-kadang dikategorikan ke dalam ekstrim sukar yaitu apabila nilai p mendekati nol dan ekstrim mudah apabila nilai p mendekati satu. Menurut Fernandes (1984), butir soal yang menghasilkan rerata skor sekitar 50 % dari skor maksimum dapat dikatakan bahwa butir soal itu mempunyai tingkat kesukaran yang tepat. Sementara itu, Thomas dan Dawson (1972) menjelaskan bahwa butir soal yang memiliki tingkat kesukaran 0,25 - 0,75 sudah dikatakan baik.

Daya pembeda atau daya beda suatu butir tes berfungsi untuk menentukan dapat tidaknya suatu butir tes membedakan kelompok dalam aspek yang diukur sesuai dengan perbedaan yang ada pada kelompok itu. Tujuan dari penelaahan daya pembeda adalah untuk melihat kemampuan butir tes tertentu dalam membedakan antara pengambil tes yang berkemampuan tinggi dan pengambil tes yang berkemampuan rendah.

Ada beberapa cara yang digunakan untuk menghitung daya pembeda, yaitu: (1) indeks diskriminasi, (2) indeks korelasi, dan (3) indeks keselarasan. Pada tulisan ini hanya dibahas dua cara untuk menghitung daya pembeda dengan metode korelasi, yaitu korelasi *point biserial* dan korelasi *biserial*. Korelasi *point biserial* maupun korelasi *biserial* adalah korelasi *product moment* yang diterapkan pada data, variabel-variabel yang dikorelasikan sifatnya masing-masing berbeda satu sama lain. Korelasi *point biserial* adalah korelasi dua variabel, satu variabel berskala nominal atau dikotomi yaitu bernilai 1 untuk jawaban benar dan 0 untuk jawaban salah, sedangkan variabel lainnya berskala interval atau rasio. Korelasi biserial adalah korelasi dua variabel, satu variabel berskala ordinal, sedangkan variabel lainnya berskala interval atau rasio.

Indeks daya beda butir soal dapat digunakan sebagai bahan pertimbangan sebuah butir baik atau tidak baik. Butir soal yang baik adalah butir soal yang mempunyai indeks daya beda lebih dari 0,2 (Fernandes, 1984). Sementara Ebel (1972) menjelaskan suatu butir soal dikatakan berkualitas apabila indeks diskriminasi atau daya pembedanya paling sedikit 0,41.

Hal penting yang juga harus diperhatikan dalam menganalisis empirik butir soal adalah kemampuan distraktor atau alternatif jawaban yang disediakan menarik peserta tes untuk memilihnya. Jangan sampai tidak seorang peserta tes-pun memilih alternatif jawaban yang disediakan. Fernandes (1984) yang mengutip pendapat Brawn menjelaskan distraktor dikatakan baik apabila paling tidak dipilih oleh 2 % dari seluruh peserta. Sementara itu, Nitko (1996) mengatakan distraktor dikatakan berfungsi manakala paling tidak dipilih oleh seorang peserta tes dari kelompok rendah. Pemilih dari kelompok rendah harus lebih banyak daripada kelompok atas. Distraktor juga dapat dikatakan berfungsi manakala peserta tes (siswa) dari kelompok atas dapat membedakan antara distraktor dan kunci jawaban sehingga yang memilih kunci jawaban lebih banyak daripada yang memilih distraktor.

Dalam menganalisis distribusi jawaban juga perlu memperhatikan kemungkinan salah kunci, yaitu manakala siswa dari kelompok atas yang memilih pengecoh lebih banyak daripada

yang memilih kunci jawaban. Selain itu, juga perlu dideteksi ada tidaknya unsur tebakan dalam memilih alternatif jawaban. Hal ini dapat dilihat apabila jawaban peserta tes (siswa) merata, baik jawaban dari siswa kelompok atas maupun kelompok bawah.

Hal penting lainnya dalam menuliskan butir-butir soal UTK adalah sebagian besar butir-butir soal itu sebaiknya memenuhi *the higher level of thinking (HOT)*. Menurut Moore, B dan Stanley T (2010), dari peringkat kognitif Bloom itu, urutan nomor 1 – 3, yakni pengetahuan, pemahaman, dan aplikasi dikategorikan *the lower level of thinking*. Sementara itu tingkat 4 -6, yakni analisis, evaluasi, dan kreasi termasuk *the higher level of thinking (HOT)*. Ini berarti bahwa sebagian besar butir-butir soal UTK sebaiknya berada pada tingkat analisis, evaluasi, dan kreasi.

Dengan demikian jelaslah bahwa untuk menilai kualitas butir tes dengan pendekatan teori tes klasik tidak cukup hanya memperhatikan tingkat kesukaran dan daya pembeda butir tes yang bersangkutan. Penilaian kualitas butir tes juga harus memperhatikan tingkat kognitif Bloom butir itu dan juga keberfungsian pilihan jawaban, terutama distraktor-distraktornya. Pilihan jawaban itu harus tampak sebagai jawaban yang benar bagi subjek dari kelompok yang berkemampuan rendah. Sebaliknya harus tampak sebagai jawaban yang salah bagi subjek dari kelompok yang berkemampuan tinggi.

Telah dijelaskan bahwa selain valid, soal UTK juga harus reliabel, andal, stabil, atau konsisten. Bila instrumen menggunakan metode rating atau pemberian skor berdasarkan *judgment* subyektif terhadap atribut tertentu yang dilakukan melalui pengamatan sistematis secara langsung maupun tidak langsung, maka reliabilitasnya dapat dihitung menggunakan persamaan Ebel (Saifuddin Azwar, 2013). Estimasi reliabilitas dapat dilakukan dengan cara memberi angka ulang pada atribut yang sama pada waktu berbeda kemudian mengkorelasikan kedua hasil rating itu. Biasanya teknik korelasi yang digunakan adalah koefisien korelasi jenjang Spearman (*rank-order correlation*). Teknik ini banyak kelemahan karena besarnya varians error dikarenakan adanya pengaruh faktor ingatan (memory) dari fihak rater.

Cara yang lebih praktis adalah memperbanyak rater, rater lebih dari satu tetapi setara kepakarannya dan independen satu sama lain. Bila rating dilakukan oleh beberapa raters maka makna reliabilitas hasil rating merupakan konsistensi diantara para raters (*interrater reliability*) atau ada juga yang menyebut dengan *Intraclass Correlation Coefficients* (ICC) dengan rumus:

$r_{xx} = (S_s^2 - S_e^2)/[S_s^2 + (k-1)S_e^2]$ ….. (1)

r_{xx} = reliabilitas antar rater

S_s^2 = varian antar subyek (dalam hal ini butir) yang dikenai rating

S_e^2 = varian error, yaitu varians interaksi antara butir atau subyek (s) dan rater (r)

K = banyaknya rater yang memberikan penilaian

Sebagai contoh, tiga orang pakar (rater) diminta untuk menilai soal UTK, apakah soal ini memiliki validitas isi yang baik atau tidak. Ketiga rater tadi diberi cheklist (lembar penilaian) untuk menilai soal UTK) yang terdiri dari 10 butir, dan setiap butir dinilai dengan skor 1 – 4; dengan ketentuan: (1) sangat tidak tepat, (2) tidak tepat, (3) tepat, dan (4) sangat tepat. Contoh hasil penilaian yang dilakukan oleh tiga rater dapat dilihat pada Tabel 1.

Tabel 1. Contoh data hasil penilaian validitas isi butir-butir soal UTK

NOMOR BUTIR	RATER 1	RATER 2	RATER 3
1	3	4	4
2	3	4	4
3	2	2	2
4	4	4	3
5	2	2	4

NOMOR BUTIR	RATER 1	RATER 2	RATER 3
6	1	3	2
7	3	4	3
8	4	4	4
9	4	3	4
10	3	4	3

Persamam nomor 1 dapat diselesaikan secara manual dan dapat juga dihitung dengan menggunakan SPSS. Hasil hitungan dengan bantuan SPSS dapat dilihat pada Tabel 2 berikut.

Tabel 2. Hasil analisis Anova dari ke tiga rater

		Sum of Squares	df	Mean Square	F	Sig
Between People		14,133	9	1,570		
Within People	Between Items	1,400	2	,700	1,734	,205
	Residual	7,267	18	,404		
	Total	8,667	20	,433		
Total		22,800	29	,786		

Tabel 2 menunjukkan bahwa tidak ada perbedaan yang signifikan (p = 0,205) antara rata-rata skor rater 1, 2, dan 3. Selanjutnya harga reliabilitas antar rater (*Intraclass Correlation*) dapat dilihat pada Tabel 3.

Tabel 3. ***Intraclass Correlation Coefficient***

	Intraclass Correlation (a)	95% Confidence Interval		F Test with True Value 0			
	Lower Bound	Upper Bound	Value	df1	df2	Sig	Lower Bound
Single Measures	,491(b)	,099	,817	3,890	9,0	18	,007
Average Measures	,743(c)	,247	,931	3,890	9,0	18	,007

Two-way mixed effects model where people effects are random and measures effects are fixed.
a Type C intraclass correlation coefficients using a consistency definition-the between-measure variance is excluded from the denominator variance.
b The estimator is the same, whether the interaction effect is present or not.
c This estimate is computed assuming the interaction effect is absent, because it is not estimable otherwise.

Tabel 3 menunjukkan bahwa reliabilitas antar rater atau ***Intraclass Correlation Coefficient*** sebesar 0,491, suatu harga yang cukup moderate atau cukup reliabel. Apabila raternya hanya dua maka dapat digunakan cara Wilson (2008) sebagai berikut.

	Ratings of Rater 1				
Ratings of Rater 2	*1*	*2*	*3*	*4*	*5*
5	0	0	1	2	4
4	0	0	2	3	2
3	0	2	3	1	0
2	1	1	1	0	0
1	1	1	0	0	0

Once the data are recorded you can calculate inter-rater agreement with the following formula

$$\text{Inter-Rater Agreement} = \frac{\text{Number of Cases Assigned the Same Scores}}{\text{Total Number of Cases}} \times 100$$

In our example the calculation would be:

Inter-Rater Agreement = 12/25 × 100
Inter-Rater Agreement = 48%

Apabila penilaian rater terhadap butir soal UTK bukan berbentuk rating tetapi berbentuk katagori maka teknik estimasi reliabilitasnya sedikit berbeda. Pilihan bentuk rating dapat diubah menjadi bentuk katagori, misal pilihan ***sangat tidak tepat*** dan ***tidak tepat*** diganti menjadi ***tidak tepat*** dan pilihan ***tepat*** dan ***sangat tepat*** diganti menjadi ***tepat***. Teknik untuk mengestimasi reliabilitasnya digunakan rumus Cohen's Kappa (Gwet KL, 2012):

$$Kc = \frac{pa-pe}{1-pe} \quad \ldots\ldots\ldots (2)$$

Dalam hal ini:

$Pa = \frac{n11+n22}{n}$ dan Pe = [(n1+/n)(n+1/n)][(n2+/n)(n+2/n)]……, perhatikan Tabel 3

Sebagai contoh, dua orang pakar (rater) diminta untuk menilai soal UTK, apakah soal ini memiliki validitas isi yang baik atau tidak. Dua orang rater tadi diberi cheklist (lembar penilaian) untuk menilai soal UTK) yang terdiri dari 10 butir, dan setiap butir dinilai dengan pilihan **tepat** dan **tidak tepat**. Rater A mengatakan bahwa dari 10 butir soal UTK; 7 butir tepat dan 3 butir tidak tepat, sedangkan Rater B mengatakan 5 butir tepat dan 5 butir lainnya tidak tepat (lihat Tabel 3). Ada 5 butir yang dinilai tepat oleh Rater A dan Rater B, dan 3 butir yang dinilai tidak tepat oleh Rater A dan Rater B.

Tabel 3. Contoh tabel persiapan analisis dengan Kai-Kuadrad

		Rater B		
		(1)	**(2)**	Total
Rater A	**(1)**	**n_{11} (5)**	**n_{12} (2)**	n_1+(7)
	(2)	**n_{21} (0)**	**n_{22} (3)**	n_2+ (3)
	Total	n+ 1 (5)	n+2 (5)	**n (10)**

$Pa = \frac{n11+n22}{n} \rightarrow Pa = \frac{5+3}{10} = 0,8$

Pe = [(n1+/n)(n+1/n)] + [(n2+/n)(n+2/n)]

Pe = [(7/10)(5/10)] + [(3/10)(5/10)] → Pe = 0,5

$Kc = \frac{0,8-0,5}{1-0,5}$ → Kc = 0,6

Selain secara manual, estimasi kesepakatan dua rater model Cohen Kappa juga dapat dilakukan dengan menggunakan SPSS. Contoh hitungannya adalah sebagai berikut.

NOMOR BUTIR	RATER A	RATER B
1	1	1
2	2	2
3	1	1
4	1	1
5	1	2
6	1	2
7	2	2
8	1	1
9	1	1
10	2	2

Keterangan: 1 = tepat dan 2 = tidak tepat

Hasil analisis dengan SPSS dengan langkah-langkah:

- Analyze > Descriptive statistics
- Msukkan Rater A pada row dan Rater B pada coloum
- Masuk ke Menu Statistics > Kappa tekan continue
- Masuk ke menu Cells lalu tekan Total di bawah percentage, kemudian tekan Continue
- Klik OK

Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
A * B	10	100,0%	0	,0%	10	100,0%

A * B Crosstabulation

			B		Total
			1,00	2,00	1,00
A	1,00	Count	5	2	7
		% of Total	50,0%	20,0%	70,0%
	2,00	Count	0	3	3
		% of Total	,0%	30,0%	30,0%
Total		Count	5	5	10
		% of Total	50,0%	50,0%	100,0%

Tabel A*B crosstabulation di atas menunjukkan bahwa ada 5 butir yang disepakati ***tepat*** dan 3 butir soal disepakati ***tidak tepat*** oleh Rater A dan Rater B. Selanjutnya hasil hitungan agreement Cohen's Kappa dapat dilihat pada Tabel berikut.

Symmetric Measures

		Value	Asymp. Std. Error(a)	Approx. T(b)	Approx. Sig.
Measure of Agreement	Kappa	,600	,232	2,070	,038
N of Valid Cases		10			

a Not assuming the null hypothesis.
b Using the asymptotic standard error assuming the null hypothesis.

Tabel ***Symmetric Measures*** menunjukkan harga *Measure of agreement* Kappa sebesar 0,6, yakni sama besarnya dengan hasil hitungan secara manual. Hasil ini dapat disimpulkan bahwa tingkat kesepakatan antara Rater A dan Rater B tentang soal UTK cukup baik, yakni 0,6 (Keren D. Multon, 2012). Hal ini selaras dengan pendapat para ahli lain yang mengatakan bahwa reliabilitas instrumen dapat dikatakan baik manakala besarnya minimum 0,7 (Feldt, L.S. and Brennan, R.L. 1989).

Cara yang lebih mudah dalam mengestimasi reliabilitas interrater dijelaskan oleh Salkind (2013) yang mengatakan bahwa besarnya reliabilitas interrater sama dengan perbandingan antara jumlah kesepakatan dan jumlah kemungkinan kesepakatan. Contoh, perhatikan Tabel 4 berikut.

Tabel 4. Tabel persiapan estimasi reliabilitas interrater

Butir	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Rater 1	+	-	+	+	+	+	+	-	-	-
Rater 2	-	+	+	+	+	-	+	-	-	-

$$\textit{Interrater reliability} = \frac{Number\ of\ agreements}{Number\ of\ possible\ agreements}$$

Interrater reliability = $\frac{8}{10}$ = 0,8. Jadi reliabilitas interater soal UTK = 0,8, berarti soal UTK itu reliabel. Harga ini lebih besar daripada yang disarankan oleh Feldt, L.S. and Brennan, R.L. (1989) yang mengatakan bahwa reliabilitas instrumen dianggap baik bila harganya paling kecil 0,7. Sementara itu, bila raternya lebih dari dua dan pilihannya berbentuk katagori maka reliabilitas interraternya dapat dihitung dengan koefisien Fleis Kappa.

2. Kualitas Pelaksanaan UTK dan *Soft Skill* lulusan

Dengan langkah-langkah yang telah dijelaskan di atas akan diperoleh soal UTK yang berkualitas tinggi. Untuk mendapatkan hasil yang akurat, selain soal tes yang baik pelaksanaan tes juga harus baik. Oleh karenanya, harus diupayakan sekuat tenaga agar pelaksanaan UTK di SMK yang berarti juga pelaksanaan UN berkualitas tinggi. Pelaksanaan UTK yang baik, harus jujur dan tidak akan meluluskan siswa yang belum memenuhi persyaratan. Pelaksanaan UN yang baik juga harus adil, dan dapat dipertanggungjawabkan. Dalam pelaksanan UTK yang baik, siswa yang dapat lulus adalah siswa yang betul-betul memenuhi persyaratan akademik dan non-akademik (berperilaku baik). Hal ini akan mendorong siswa untuk belajar lebih sungguh-sungguh, lebih ulet, bekerja lebih keras, lebih menghargai waktu, lebih tanggung jawab, dan lebih religius karena lebih banyak berdoa. Hal ini selaras dengan pendapat Khairil Anwar Notodipuro (2012) yang mengatakan bahwa UN yang diselenggarakan dengan baik akan mendorong siswa lebih tanggung jawab, ulet, lebih menghargai waktu, dan lebih religius.

Hasil penelitian Djemari Mardapi dan Badrun Kartowagiran (2010) menunjukkan bahwa dengan adanya UN maka ada 81% siswa dari sekolah kategori tinggi dan 65% siswa dari sekolah kategori rendah menambah jam belajar sekitar 10 jam/minggu dengan cara mengikuti les di sekolah. Sementara itu, Khairil Anwar Notodipuro (2012) yang mengutip hasil penelitian Djemari Mardapi, penelitian Supriyoko, dan penelitian Furqon mengatakan bahwa UN dapat mendorong siswa untuk lebih semangat belajar, rajin mencari sumber bacaan, dan rajin masuk sekolah.

Uraian di atas menjelaskan bahwa dengan mengoptimalkan UTK di SMK yang berarti mengoptimalkan UN maka akan mendorong siswa untuk bekerja lebih keras, belajar lebih sungguh-sungguh, lebih menghargai waktu, lebih ulet, dan lebih tanggung jawab. Butir-butir perilaku positif ini merupakan bagian dari butir-butir soft skill. Hal ini sejalan dengan pendapat Perreault (Mitchel, 2008) yang menjelaskan bahwa *soft skills* merupakan kualitas personal, atribut atau tingkat komitmen seseorang, yang membedakan orang tersebut dengan orang lain

yang memiliki kecerdasan dan pengalaman sama. Sementara itu, Mitchel (2008) yang mengutip pendapat James dan James mengatakan bahwa *soft skills* merupakan cara baru untuk mendeskripsikan seperangkat kemampuan atau talenta seseorang yang tampak saat dia bekerja. Lebih jauh James dan James menjelaskan bahwa *soft skills* seperti kemampuan untuk bekerja dalam tim, keterampilan berkomunikasi, keterampilan kepemimpinan, layanan langganan, dan keterampilan pemecahan masalah sangat bermanfaat untuk perkembangan karir.

PENUTUP

Kualitas soal UTK yang tinggi akan membuahkan hasil UTK yang akurat, sehingga siswa yang dinyatakan lulus betul-betul memiliki kompetensi tinggi atau sesuai dengan standar yang telah ditentukan. Lain halnya bila soal UTK berkualitas rendah maka siswa yang dinyatakan lulus belum tentu mereka memiliki kompetensi tinggi atau sesuai standar yang telah ditentukan.

Pelaksanaan UTK yang baik, yakni jujur, disiplin, dan akuntabel hanya meluluskan siswa yang betul-betul sudah memenuhi standar. Hal ini mendorong siswa untuk lebih ulet, bekerja lebih keras, belajar lebih sungguh-sungguh, lebih menghargai waktu, dan lebih tanggung jawab.

Uraian di atas dapat disimpulkan bahwa dengan optimalisasi UTK, yakni mengusahakan agar soal dan pelaksanaan UTK berkualitas tinggi, dapat mendorong lulusan berkualitas tinggi dan meningkatkan *soft skill* lulusannya.

DAFTAR PUSTAKA

Allen, M.J. & Yen, W.M. 1979. *Introduction to measurement theory*. Monterey, CA: Brooks/Cole Publishing Company.

Badrun Kartowagiran, Amat Jaedun, dan Heri Retnowati. Evaluasi kesiapan SMP di D.I. Yogyakarta dalam mengimplementasikan kurikulum tahun 2013. *Laporan Penelitian.* Yogyakarta: tidak diterbitkan.

Djemari Mardapi dan Badrun Kartowagiran. 2010. Dampak Ujian Nasional. *Laporan Penelitian*. Yogyakarta: tidak diterbitkan

Dawson, J.B. & Thomas, G.H. 1972. *Item analysis and examination statics*. Birmingham: The Union of Educational Institutions.

Ebel, R.L. 1972. *Essentials of educational measurement*. (3rd. ed.) Englewood Cliffts,NJ: Prentice Hall Inc.

Fernandes, H.J. X. 1984. *Evaluation of educational program*. Jakarta: National Education Planning , Evaluating and Curriculum Development.

Feldt, L.S. and Brennan, R.L. 1989. "Reliability", *Educational measurement,* edited by Robert L Linn. New York: Macmillan Publishing Company.

Gwet, K.L.2012. *Handbook of inter-rater reliability*. MD: Advanced Analytics

Lawrence M.R. 1994. Question to ask when evaluaating test. *Eric digest*. Artikel: ED385607. Sumber: http://www.ericfacility.net/ericdigest/ ed.385607.html tanggal 10 Februari 2003.

Moore, B., Stanly, T. 2010. *Critical thinking and formative assessments.* Larchmount, NY: Eye On Education, Inc

Multon, Keren D. 2012. "Interrater reliability". Encyclopedia of research design.Ed. Neil J. Salkind. Thousand Oaks, CA: SAGE, 2010. 627 – 629. SAGE Reference online. Web. 18 July 2012

Nitko, A.J. 1996. *Penilaian berkelanjutan berdasarkan kurikulum (PB2K): Kerangka, konsep, prosedur, dan kebijakan* (terj. AM. Ahmad) Jakarta: Pusat Pengembangan Agribisnis.

Perpres R.I. Nomor 8 Tahun 2012 Tentang Kerangka Kualifikasi Nasional Indonesia

Permendikbud R.I. Nomor 66 Tahun 2013 Tentang Standar Penilaian

Reynolds, C.R., Livingston, R.B., dan Wilson, V. 2008. *Measurement and Assessment in Education.* Englewood Cliffs, NJ: Prentice-Hall,Inc.

Saifuddin Azwar. 2013. *Validitas dan reliabilitas*. Ed.4. Yogyakarta: Pustaka Pelajar

Salkind, N.J. 2013. Test & measurement for people who hate test and measurement. Los Angeles: SAGE Publications, Inc

Thomas, A. dan Thorne, G. (2007). *Higher Order Thinking*. Center for Development and learning. Diambil dari CDL pda tanggal 6 Agustus 2011.

Woolfolk, A.E. & McCune, L.N. 1984. *Educational Psychology for Teachers*. Englewood Cliffs, NJ: Prentice Hall, Inc.