Buletin Psikologi
Volume 32, Nomor 2, 2024: 187– 204
DOI: 10.22146/buletinpsikologi.95539

ISSN 0854 - 7106 (print)
ISSN 2528-5858 (Online)

Aspek-Aspek yang Perlu Diperhatikan *Rater* dalam Verifikasi Isi Instrumen Pengukuran

Aspects to be Considered by Raters in Verifying the Content of Measurement Instruments

Ahmad Saifuddin*[1]
*Fakultas Ushuluddin dan Dakwah, Universitas Islam Negeri Raden Mas Said Surakarta, Indonesia

Naskah Masuk 16 April 2024 **Naskah Diterima 28 Oktober 2024** **Naskah Terbit 11 Desember 2024**

Abstrak. *Rater* memiliki peran penting dalam memvalidasi skor hasil pengukuran, sehingga *rater* diharapkan memiliki keahlian, salah satunya pemahaman tentang psikometrika dan ilmu pengukuran. Akan tetapi, pada kenyataannya terdapat *rater* yang tidak memiliki kompetensi tersebut sehingga terdapat beberapa kesalahan dalam memverifikasi isi instrumen pengukuran. Oleh karena itu, artikel ini bertujuan untuk menyoroti pentingnya kompetensi *rater*, bagaimana *rater* harus membangun kompetensi tersebut, dan bagaimana seharusnya peneliti memilih *rater*. Metode yang digunakan meliputi *literature review* dan pengalaman reflektif yang menghasilkan bahwa terdapat beberapa prinsip dan kaidah dalam memvalidasi instrumen pengukuran yang harus dikuasai dan dipahami oleh *rater*. Secara etika, ketika seseorang kurang memahami kaidah tersebut, maka sebaiknya tidak bersedia untuk memverifikasi isi instrumen pengukuran. Selain itu, para akademisi juga perlu untuk membangun kompetensinya di bidang pengukuran untuk mengisi kelangkaan jumlah seseorang yang pantas menjadi *rater*. Di sisi lain, para peneliti juga hendaknya menelusuri secara mendalam tentang kompetensi dan rekam jejak ilmiah seseorang yang akan diminta menjadi *rater*.

Kata kunci: isi instrumen; penyusunan instrumen pengukuran; *rater*; validasi

Abstract. Raters have an important role in validating test scores, so raters are expected to have expertise, one of which is an understanding of psychometrics and measurement science. However, in reality there are raters who do not have these competencies so that there are several errors in verifying the contents of the item. Therefore, this article aims to highlight the importance of rater competencies, how raters should build these competencies, and how researchers should select raters. The methods used include literature review and reflective experience, which resulted in several principles and rules in verifying the contents of instruments that must be mastered and understood by raters. Ethically, when someone does not understand these rules, they should not be willing to validate measurement instruments. In addition, academics also need to build their competence in the field of measurement to fill the scarcity of people who are suitable to become raters. On the other hand, researchers should also explore in depth the competence and scientific track record of someone who will be asked to become a rater.

Keywords: evidence based on test content; instrument development; raters; validation

*Alamat Korespondensi: ahmad.saifuddin@staff.uinsaid.ac.id

Pengantar

Instrumen pengukuran merupakan komponen penting dalam penelitian kuantitatif karena berperan sebagai alat untuk mengumpulkan data. Sebagai alat pengumpul data, instrumen harus memiliki kualitas yang baik. Salah satu indikator yang menunjukkan kualitas instrumen pengukuran adalah validitas. Validitas memiliki beberapa definisi. Secara singkat, validitas dimaknai sebagai kemampuan instrumen pengukuran untuk mengukur dan mengungkap variabel atau konstruk yang hendak diteliti. Dengan kata lain, validitas mengacu pada ketepatan sasaran instrumen pengukuran (Drost, 2011; Messick, 1989; van Heerden & Mellenbergh, 2003). Artinya, ketika tujuan pengukuran hendak mengungkap konstruk X, maka skor yang terungkap dari instrumen pengukuran juga harus mampu menggambarkan X tersebut. Selain itu, validitas juga dimaknai bukan hanya ketepatan sasaran, tetapi juga ketepatan interpretasi makna skor yang didapatkan oleh pengukuran menggunakan instrumen pengukuran (Borsboom *et al.*, 2003; van Heerden & Mellenbergh, 2003).

Berdasarkan pedoman yang berjudul "*Standards for educational and psychological testing*" (yang kemudian sering disebut *Standards*) dan disusun oleh *American Educational Research Association* (AERA), *the American Psychological Association* (APA), dan *National Council on Measurement in Education* (NCME), validitas mengacu pada sejauh mana bukti dan teori mendukung interpretasi skor pengukuran. Proses validasi melibatkan akumulasi terhadap bukti relevan yang memberikan dasar ilmiah dan kuat terhadap interpretasi skor yang dihasilkan. Pernyataan tentang validitas harus mengacu pada interpretasi tertentu untuk kegunaan tertentu. Sehingga, tidaklah benar menggunakan kalimat "validitas tes" (AERA, APA, and NCME, 2014). Atas dasar ini, maka pencapaian validitas tergantung dari tujuan spesifik pengukuran (Kane, 2015).

Menurut AERA, APA, and NCME (2014) terdapat lima sumber validitas. Pertama, konten pengukuran (*evidence based on test content*) yang merujuk pada keterwakilan aspek-aspek berdasarkan kerangka teoritis pada suatu instrumen pengukuran (Knekta *et al.*, 2019). Guna menguji sumber validitas ini, maka peneliti bisa menerapkan validitas konten. Adapun validitas konten bisa dilakukan dengan meminta tolong ahli (sering disebut dengan *rater*) untuk menilai tingkat kesesuaian item-item, aspek, dan indikator dalam instrumen pengukuran. Penilaian tersebut kemudian diolah dengan formula tertentu (misalkan formula Aiken (1985), Lawshe (1975), atau Hambleton (1980) sehingga diketahui koefisien validitas item dan instrumen pengukuran.

Kedua, *evidence based on internal structure* yang merujuk pada sejauh mana hubungan di antara butir-butir tes dan komponen-komponen tes sesuai dengan konstruk yang menjadi dasar interpretasi skor tes yang diusulkan (AERA, APA, and NCME, 2014). Sumber validitas ini bisa diketahui dengan melakukan analisis faktor terhadap item-item dalam instrumen pengukuran (Knekta *et al.*, 2019). Selain itu, juga bisa menggunakan model *bifactor*, *Confirmatory Factor Analysis* (CFA), dan *Multiple Group Confirmatory Factor Analysis* (MGCFA) (Rios & Wells, 2014).

Ketiga, *evidence based on response processes* yang merujuk pada informasi mengenai bagaimana respons yang dimunculkan oleh responden atau sampel pengukuran (AERA, APA, and NCME, 2014; Knekta *et al.*, 2019). Isu utama dalam sumber validitas ini adalah apakah responden atau sampel pengukuran termotivasi dan jujur dalam menjawab instrumen pengukuran. Selain itu juga apakah pemahaman yang muncul pada responden atau sampel pengukuran sama dengan yang diinginkan oleh pembuat instrumen pengukuran (Knekta *et al.*, 2019). Hal ini berkaitan dengan tingkat kemudahan item untuk dipahami responden atau sampel pengukuran di setiap kondisi. Di sisi lain, *evidence based on response processes* dapat diketahui dari beberapa metode, misalkan mengetahui durasi waktu pengerjaan instrumen pengukuran yang dilakukan seseorang, pergerakan mata ketika seseorang mengerjakan instrumen pengukuran, wawancara, dan *focus group discussion* (Padilla & Ben'itez, 2014).

Keempat, *evidence based on relations to other variables* yang merujuk pada korelasi antara instrumen pengukuran dengan variabel lain (AERA, APA, and NCME, 2014). Korelasi ini bisa ditunjukkan dengan menguji instrumen pengukuran yang disusun dengan instrumen pengukuran lain yang secara teoritis memiliki hubungan dengan instrumen tersebut. Dengan demikian, sumber validitas ini kemudian berkaitan dengan beberapa jenis validitas. Misalkan, validitas diskriminan, validitas konvergen, dan validitas *criterion*.

Kelima, *evidence based on the consequences of testing* yang merujuk pada konsekuensi yang dimunculkan dari skor yang dihasikan oleh pengukuran (AERA, APA, and NCME, 2014). Pertanyaan penting dari jenis *validity evidence* ini adalah apakah penggunaan instrumen akan menyebabkan konsekuensi yang tidak diinginkan bagi responden (Knekta *et al.*, 2019). Lebih lanjut, *evidence based on the consequences of testing* diuji dengan mengevaluasi sistem penilaian yang diciptakan dari suatu instrumen pengukuran (Lane, 2014).

Cizek (2015) kemudian mencoba merekonstruksi paradigma validitas dalam *Standards*. Menurutnya, secara garis besar terdapat dua sumber *validity evidence*, yaitu *sources of evidence for validating score meaning* (sumber bukti untuk memvalidasi makna skor pengukuran) dan *sources of evidence for validating justifying test use* (sumber bukti untuk memvalidasi penggunaan suatu instrumen atau alat tes). *Sources of evidence for validating score meaning* terdiri dari *evidence based on test content* (bukti validitas berdasarkan konten pengukuran); *evidence based on response process* (bukti validitas berdasarkan respons responden); *evidence based on hypothesised relationship among variables* (bukti validitas berdasarkan hubungan antara variabel yang diukur dengan variabel lain); dan *evidence based on test development and administration procedures* (bukti validitas berdasarkan penyusunan tes dan prosedur administrasi).

Adapun *sources of evidence for validating justifying test use* terdiri dari *evidence based on consequences of testing* (bukti validitas berdasarkan konsekuensi yang dimunculkan dari pengukuran); *evidence based on costs of testing* (bukti validitas berdasarkan pembiayaan yang muncul akibat pengukuran yang dilakukan); *evidence based on alternatives to testing* (bukti validitas berdasarkan evaluasi terhadap format dan metode pengukuran serta evaluasi terhadap pilihan metode lain untuk mencapai tujuan pengukuran); dan *evidence based on fundamental fairness* (bukti validitas berdasarkan asas keadilan yang

dirasakan oleh sampel pengukuran dan pemangku kebijakan) (Cizek, 2015).

Paradigma lain tentang validitas menjelaskan bahwa validitas tidak perlu dibagi-bagi (*not need to be decentralized*) dan menganggap validitas sebagai sistem yang universal dan interaktif/terkait antar beberapa validitas. Misalkan, aspek *practical constraints* berdampak pada desain item dan spesifikasi pengukuran. Adapun desain item dan spesifikasi pengukuran ini juga dipengaruhi oleh *latent process studies* dan *logical/theoretical analysis*. Kemudian, spesifikasi pengukuran berdampak pada properti psikometrik suatu instrumen. Di sisi lain, properti psikometrik instrumen pengukuran ini juga dipengaruhi oleh model atau bentuk penilaian/skoring suatu instrumen (Embretson, 2007).

Paradigma semacam itu dipengaruhi oleh pendapat Lissitz dan Samuelsen (2007). Kedua ilmuwan ini menawarkan perspektif baru dalam merumuskan validitas. Perspektifnya diawali dengan pendapat bahwa penyusunan instrumen pengukuran melibatkan faktor internal dan faktor eksternal instrumen pengukuran. Faktor internal ini berupa konten instrumen pengukuran, reliabilitas, dan *theoretical latent process* (berkaitan dengan *review* terhadap item *performance* data, menguji interkorelasi antar item, menguji validitas konvergen dan divergen, analisis faktor, dan analisis respons kognitif responden). Adapun pengujian terhadap faktor eksternal meliputi kegunaan praktis instrumen pengukuran, dampak pengukuran, dan *theoretical nomological network*. Konsep ini dikuatkan oleh Moss (2007). Ilmuwan lain yang setuju dengan konsep *unitary concept of validity* adalah Gorin (2007). Menurutnya validitas instrumen berkaitan dengan definisi operasional terhadap teori yang digunakan dan prosedur penyusunan instrumen pengukuran. Meskipun terdapat dua paradigma yang berbeda, penulis memilih menggunakan paradigma yang dirumuskan oleh AERA, APA, and NCME (2014).

Seperti yang telah dijelaskan bahwa terdapat berbagai metode yang dapat digunakan untuk mencapai validitas instrumen pengukuran, salah satunya adalah validitas konten. Validitas konten merujuk pada kesesuaian dan keselarasan item-item yang ada di dalam instrumen pengukuran dengan konstruk yang hendak diukur dan diungkap (Cohen *et al.*, 2022; Kaplan & Saccuzzo, 2017). Validitas konten ini dilakukan dengan cara meminta ahli (atau kemudian disebut dengan *rater*) untuk menilai kesesuaian tersebut (Gafni, 2016). Mekanisme ini sering disebut dengan *expert judgement* atau *professional judgement* (Berk, 1990). Di sisi lain, jika merujuk *validity evidence* menurut AERA, APA, and NCME (2014), maka tugas *rater* juga mengecek aspek lain dari instrumen pengukuran, misalkan kesesuaian item dengan bentuk respons jawaban pada instrumen, kejelasan instruksi, efektivitas kalimat yang digunakan dalam item, sistem penilaian, serta relasi antara instrumen pengukuran terkait dengan variabel lain.

Adapun batasan dari ahli adalah seseorang yang memiliki karya publikasi dan pengalaman di bidang terkait (Rubio *et al.*, 2003). Selain itu, ahli juga hendaknya memiliki keterlibatan dalam pengembangan suatu disiplin keilmuan (Syaiful & Roebianto, 2020). Oleh karena itu, dalam konteks penyusunan dan pengembangan instrumen pengukuran, dibutuhkan ahli di bidang konstruk yang hendak diukur dan di bidang psikometrika.

Tingkat kompetensi dan ketelitian ahli dalam memvalidasi instrumen pengukuran berdampak pada tingkat validitas konten dari instrumen pengukuran tersebut. Kompetensi ahli terkait ilmu

pengukuran dan psikometrika bermanfaat untuk memahami arah variabel atau konstruk yang diukur sehingga bisa menilai ketepatan sifat item. Selain itu, kompetensi ahli dalam bidang pengukuran dan psikometrika juga berperan di dalam menilai item-item yang ada dalam suatu instrumen sudah sesuai dengan kaidah atau belum. Di sisi lain, ketelitian ahli berperan di dalam menilai keselarasan antara teori, aspek, indikator, dengan item yang telah disusun oleh peneliti.

Meskipun ahli (yang juga disebut dengan *rater*) memiliki peran yang krusial, kenyataannya tidak setiap orang yang berperan sebagai *rater* memiliki kemampuan yang diharapkan. Hal ini penulis temui di beberapa kesempatan. Suatu ketika penulis pernah diminta oleh seorang peneliti untuk menjadi *rater*. Penulis menganggap ada kesalahan fundamental dalam penyusunan item tersebut. Kesalahan tersebut terletak pada ketidaktepatan pembuatan item *favourable* dan *unfavourable*. Item yang seharusnya bersifat *favourable* justru dianggap *unfavourable*, dan sebaliknya. Akan tetapi, penulis mendapati ada *rater* lain yang menilai valid instrumen tersebut.

Di kesempatan lain, penulis diminta oleh peneliti lain untuk memvalidasi instrumen. Konstruk yang hendak diukur menyasar beberapa domain, yaitu domain sikap, perilaku, dan pengetahuan. Oleh karena itu, instrumen pengukuran tersebut juga ada tiga bentuk instrumen meskipun mengukur satu konstruk. Namun, peneliti membuat instrumen tersebut hanya satu bentuk, yaitu berbentuk skala likert yang *notabene* hanya cocok untuk menyasar domain sikap. Dengan demikian, instrumen pengukuran tersebut hendaknya bersifat tidak valid. Akan tetapi, penulis mendapati seorang *rater* menilai valid instrumen tersebut.

Meskipun sudah dianggap usang oleh sebagian kalangan, validitas konten atau verifikasi isi instrumen dengan *professional judgement* ini masih banyak diterapkan oleh sebagian akademisi. Kesalahan validasi, terutama karena *rater* yang kurang berkompetensi, bisa mengakibatkan instrumen pengukuran menjadi tidak valid. Berdasarkan kesenjangan dan urgensitas tersebut, maka penting untuk membahas tentang keahlian *rater*. Oleh karena itu, tujuan dari artikel ini adalah untuk menjelaskan tentang keahlian apa saja yang diperlukan oleh *rater*; bagaimana *rater* harus membangun kompetensi tersebut; dan bagaimana seharusnya peneliti memilih *rater*.

Pembahasan

Keahlian Rater Dalam Memahami Bentuk Pengukuran dan Sifat Item

Seperti yang telah dijelaskan bahwa *rater* seharusnya memiliki beberapa kompetensi penting agar dapat memvalidasi instrumen pengukuran sesuai dengan prosedur dan kaidah keilmuan yang berlaku. *Rater* hendaknya memiliki pengetahuan tentang psikometrika dasar. Di dalam psikometrika dasar, dipelajari beberapa hal penting ketika menyusun instrumen pengukuran, misalkan bentuk instrumen pengukuran. *Rater* hendaknya memahami bentuk pengukuran dan sifat item yang disesuaikan dengan domain pengukuran. Kemampuan ini penting guna memenuhi sumber validitas berupa *evidence based on test content*.

Terdapat berbagai macam bentuk instrumen pengukuran, baik dari jenis soal atau itemnya maupun dari bentuk respons jawaban. Bentuk instrumen pengukuran ini ditentukan oleh domain

yang menjadi target pengukuran. Apabila domain sikap yang menjadi target pengukuran, maka bentuk instrumen pengukuran yang relevan adalah Likert (Holt, 2014; Likert, 1932; Lozano *et al.*, 2008).

Ciri khas dari pengukuran sikap ini adalah tidak ada jawaban benar dan salah. Selain itu, rentang jawaban dari instrumen Likert ini bermula dari suatu kutub menuju kutub lain (Fishman *et al.*, 2021). Oleh karena itu, jumlah rentang jawaban pada instrumen Likert paling masyhur adalah lima pilihan jawaban, mulai dari sangat tidak sesuai, tidak sesuai, netral, sesuai, sampai dengan sangat sesuai. Meskipun demikian, respons jawaban dari instrumen Likert juga bisa dibuat sampai dengan delapan rentang (Lozano *et al.*, 2008). Rentang jawaban ini bersifat setara.

Selain Likert, domain sikap juga bisa diukur dengan bentuk *semantic differential* (Aros *et al.*, 2009; Rosenberg & Navarro, 2018). Pada dasarnya, *semantic differential* ini mirip dengan instrumen Likert, yaitu terdapat beberapa rentang jawaban. Akan tetapi, *semantic differential* biasanya memiliki rentang yang lebih panjang.

Domain lain yang bisa menjadi target pengukuran adalah domain perilaku. Perilaku dapat dimaknai sebagai pergerakan individu yang terlihat secara kasat mata (Henriques & Michalski, 2020). Oleh karena itu, bentuk pengukuran perilaku yang relevan adalah dengan menghitung jumlah perilaku yang muncul. Adapun bentuk respons dari instrumen yang mengukur domain perilaku mirip dengan Likert. Akan tetapi, respons jawaban dari pengukuran perilaku ini menunjukkan intensitas. Hal ini disebabkan perilaku merupakan bentuk pergerakan seseorang secara kasat mata dan tingkat keseringannya bisa dihitung.

Contoh bentuk respons jawaban dari pengukuran perilaku ini adalah rentang jawaban yang terdiri dari selalu, sering, kadang-kadang, jarang, dan tidak pernah. Akan tetapi, respons "selalu" dan "tidak pernah" terkesan mutlak dan hamper mustahil dialami oleh manusia. Sehingga, respons tersebut diganti dengan "hampir selalu" dan "hampir tidak pernah" (Azwar, 2021).

Rater juga hendaknya memahami perbedaan sikap dan perilaku karena kedua hal ini sering dipahami secara tumpeng tindih. Sikap dapat diartikan sebagai kecenderungan psikologis yang diekspresikan berdasarkan hasil mengevaluasi entitas tertentu (Haddock & Maio, 2008). Oleh karena berdasarkan evaluasi, maka sikap memiliki dua kecenderungan, yaitu kecenderungan penilaian positif dan negatif (Fishman *et al.*, 2021; Haddock & Maio, 2008). Atas dasar ini, item pada instrumen sikap biasanya terdiri dari dua sifat, yaitu *favourable* dan *unfavourable*. Item yang bersifat *favourable* adalah item yang mendukung terjadinya atau terdapatnya gejala atau aspek dari konstruk yang diukur pada diri seseorang yang diukur. Sedangkan, *unfavourable* adalah kebalikannya (Azwar, 2016).

Adanya item yang bersifat *unfavourable* ini bertujuan untuk menghindari munculnya respons yang monoton meskipun berpotensi mengganggu validitas faktorial dan tidak mengganggu validitas kriteria. Dengan demikian, meskipun menurut sebagian ilmuwan item *unfavourable* kurang direkomendasikan, keberadaannya tetap dibutuhkan dengan mencermati beberapa hal penting, misalkan item yang bersifat *unfavourable* harus benar-benar mencerminkan rendahnya tingkat atribut ukur dan tidak mengukur atribut ukur lain (Widhiarso, 2016).

Poin penting dari kedua sifat item ini adalah mendukung atau tidak mendukung terjadinya variabel atau konstruk yang diukur pada diri seseorang, bukan sifat positif atau negatif (Azwar, 2021;

Saifuddin, 2020; Widhiarso, 2016). Positif atau negatif tergantung dari arah konstruk yang hendak diukur. Dalam konteks psikologi, apabila konstruk yang diukur mengarah pada kesehatan mental, maka item yang sifatnya *favourable* akan berupa kalimat yang positif. Sebaliknya, item *unfavourable* akan berupa kalimat yang negatif. Contoh yang lebih konkret adalah instrumen yang mengukur motivasi belajar. Item-item yang sifatnya *favourable* harus mencerminkan adanya gejala motivasi belajar pada diri seseorang, misalkan, "saya senang mendengarkan penjelasan dosen". Item *unfavorable* dari instrumen pengukuran motivasi belajar tidak mencerminkan adanya gejala motivasi belajar, misalkan, "saya merasa berat untuk mempelajari materi kuliah".

Adapun instrumen pengukuran yang mengukur konstruk yang mengarah pada gangguan psikologis, maka item *favourable* akan mengarah pada gejala gangguan tersebut. Sebagai contoh adalah instrumen pengukuran stres. Maka, item yang *favourable* hendaknya menggambarkan gejala stres, misalkan, "saya merasa tertekan ketika terpikirkan masalah hidup". Sedangkan, item yang *unfavourable* justru tidak mencerminkan gejala stres, misalkan, "saya tetap tenang meskipun memperoleh banyak masalah". Berdasarkan penjelasan tersebut, maka item *favourable* dan *unfavourable* tidak disifati positif atau negatif, tetapi lebih pada apakah item tersebut mencerminkan gejala konstruk/atribut ukur atau tidak. Apabila terdapat kesalahan dalam memahami item *favourable* selalu positif dan item *unfavourable* selalu negatif, maka akan berpotensi kesalahan penilaian terhadap item.

Terkait sifat *favourable* dan *unfavourable*, *rater* juga perlu memahami bahwa meskipun sifat keduanya bersifat kebalikan, bukan berarti peneliti boleh membuat item yang berkebalikan. Misalkan, pada item *favourable* peneliti membuat item "saya suka mendengarkan penjelasan dosen ketika kuliah", sedangkan pada item *unfavourable* peneliti membuat item "saya malas mendengarkan penjelasan dosen ketika kuliah". Tujuan utama diadakannya item *unfavourable* adalah agar sampel penelitian tidak monoton di dalam memberikan respons sehingga rentan menyebabkan proses belajar pada sampel penelitian (Azwar, 2021; Widhiarso, 2016). Maka dari itu, pembuatan item *favourable* dan *unfavourable* tidak seharusnya dibalik. Atas dasar ini pula, maka peneliti hendaknya membuat item yang variatif. Apabila suatu perilaku atau sikap atau konteks sudah digunakan dalam suatu item, maka hendaknya tidak digunakan dalam item lain. Kondisi item semacam ini perlu dipahami oleh *rater* sehingga menjadi pertimbangan *rater* dalam memvalidasi item.

Pemahaman *rater* terhadap sifat item yang *favourable* dan *unfavourable* ini juga dapat menjadi upaya untuk memenuhi sumber validitas berupa *evidence based on the consequences of testing*. Ketika *rater* salah menilai sifat item, maka akan mengakibatkan peneliti memercayai penilaian tersebut meskipun salah. Akibatnya, skor yang dihasilkan bukan rentan kurang tepat dan representatif. Dengan demikian, akan memunculkan risiko kesalahan interpretasi terhadap skor. Akibat lainnya yang bisa muncul adalah kerentanan kesalahan pengambilan keputusan yang diterapkan kepada responden yang bersangkutan sebab kesalahan interpretasi skor tersebut. Masih berkaitan dengan sistem penilaian yang dipengaruhi oleh sifat item dan bentuk instrumen, maka *rater* juga perlu menilai apakah sistem penilaian yang dirancang oleh peneliti sudah sesuai dengan bentuk pengukuran dan domain pengukuran atau belum. Penilaian *rater* terhadap ketepatan sistem penilaian ini juga meminimalisasi kesalahan penilaian yang bisa berakibat pada munculnya risiko pada responden atau

sampel pengukuran.

Keahlian Rater dalam Memahami Domain Pengukuran

Keterampilan lain yang hendaknya dimiliki oleh seorang *rater* adalah memahami domain ukur yang menjadi target pengukuran dari suatu instrumen yang divalidasi. *Rater* hendaknya memahami bentuk pengukuran dan sifat item yang disesuaikan dengan domain pengukuran. Kemampuan ini penting guna memenuhi *evidence based on test content*. Terdapat banyak jenis dari domain ukur atau juga sering disebut dengan objek ukur, misalkan domain kognitif. Domain kognitif ini biasanya menggambarkan kapasitas pengetahuan sehingga mengenal jawaban benar dan salah. Jika merujuk pada respons benar dan salah, maka bentuk soal dalam instrumen kognitif bisa berwujud pilihan ganda atau *true-false question*. Bentuk respons yang lain dari instrumen kognitif ini adalah jawaban yang berjenjang, mulai dari jawaban yang paling sempurna sampai jawaban yang tidak sempurna. Berbeda dengan pengukuran sikap yang tidak mengenal batas waktu pengerjaan, pengukuran kognitif bisa melibatkan waktu pengerjaan (Kyllonen & Zu, 2016). Maka dari itu, instrumen kognitif akan menjadi kurang tepat apabila dibentuk menjadi skala Likert.

Apabila satu atribut atau konstruk ukur melibatkan banyak domain, maka akan ada beberapa instrumen dalam satu atribut/konstruk ukur. Contohnya adalah religiositas. Religiositas dibagi menjadi beberapa dimensi. Pertama, dimensi *ideological*, yaitu dimensi religiositas yang mengukur tingkat keyakinan dan kepercayaan kepada Tuhan beserta setiap sesuatu yang wajib diyakini dalam beragama, contohnya kepercayaan dan keyakinan kepada nabi dan rasul, malaikat, hari kiamat, dan sebagainya. Kedua, dimensi *intellectual*, yaitu dimensi religiositas yang menggambarkan pengetahuan keagamaan. Ketiga, dimensi *ritualistic*, yaitu dimensi religiositas yang mengukur implementasi peribadatan dalam beragama. Keempat, dimensi *experiential*, yaitu dimensi yang menggambarkan kedalaman penghayatan seseorang dalam beragama, khususnya terhadap Tuhan. Kelima, dimensi *consequantial*, yaitu dimensi religiositas yang menggambarkan perubahan perilaku ke arah yang lebih baik sebagai akibat dari beragama (Stark & Glock, 1968).

Ketika mencermati dimensi-dimensi tersebut, maka instrumen religiositas mengandung tiga bentuk stimulus dan respons. Pertama, bentuk Likert (misalkan, dengan jawaban sangat sesuai, sesuai, netral, tidak sesuai, dan sangat tidak sesuai) yang mencakup dimensi *ideological* dan *experiential*. Kedua, bentuk pilihan ganda yang mencakup dimensi *intellectual* karena dimensi ini mencerminkan tingkat dan kapasitas pengetahuan sehingga soalnya atau itemnya mengandung jawaban benar dan salah. Ketiga, bentuk instrumen perilaku (misalkan, dengan jawaban hampir selalu, sering, kadang, jarang, dan hampir tidak pernah) yang mencakup dimensi *ritualistic* dan *consequential*. Oleh karena itu, *blueprint* instrumen pengukuran religiositas pun ada tiga buah.

Hal-hal semacam ini yang terkadang kurang diperhatikan oleh sebagian *rater*. Di dalam dunia psikometrika, terdapat banyak bentuk stimulus dan respons yang dipengaruhi oleh domain pengukuran. Selain bentuk stimulus dan respons yang telah dijelaskan, terdapat bentuk stimulus dan respons yang lain. Misalkan, stimulus dan respons dengan model pemecahan masalah atau sering disebut dengan model hipotetik, baik stimulus tersebut benar-benar dialami oleh sampel atau

responden penelitian maupun pengandaian. Secara lebih konkret, instrumen pengukurannya bisa berbentuk pilihan ganda, namun jawabannya adalah solusi-solusi dari pertanyaan. Sehingga, sampel atau responden penelitian diminta memilih solusi yang paling tepat (Azwar, 2021).

Keahlian Rater dalam Memahami Instrumen Pengukuran Secara Holistik

Keterampilan lain yang diperlukan oleh *rater* dalam memvalidasi instrumen pengukuran adalah pemahamannya terhadap konstruk atau variabel yang akan diukur. *Rater* hendaknya memahami bentuk pengukuran dan sifat item yang disesuaikan dengan domain pengukuran. Kemampuan ini penting guna memenuhi sumber validitas berupa *evidence based on test content*. Biasanya, peneliti akan menuliskan definisi operasional serta penjelasan tentang setiap dimensi atau aspek dari konstruk yang diukur. Meskipun demikian, *rater* membutuhkan ketelitian dan kecermatan untuk mengecek ulang keselarasan antara aspek atau dimensi, indikator keperilakuan, dan item-item dalam instrumen pengukuran. Keselarasan tersebut menjadi jaminan bahwa item bukan hanya menggambarkan indikator keperilakuan, tetapi juga menggambarkan aspek/dimensi bahkan konstruk yang akan diukur. Hal ini dapat dicapai jika *rater* memahami teori yang menjadi dasar penyusunan instrumen pengukuran secara mendalam. Selain itu, *rater* juga perlu memiliki cara berpikir yang berkesinambungan dan berkaitan dalam memahami tingkat keselarasan antara arah variabel, aspek, indikator keperilakuan, dengan item pernyataan.

Terkadang, di dalam membuat item, peneliti terlalu berfokus pada aspek dan indikator keperilakuan sehingga menyebabkan peneliti membuat item yang kurang holistik. Akibatnya, item pada instrumen hanya mencerminkan aspek dan indikator keperilakuan namun kurang mencerminkan konstruk atau variabel yang diukur. Padahal, seharusnya item yang ada di dalam instrumen pengukuran bukan sebatas mencerminkan aspek dan indikator keperilakuan, namun juga mampu menggambarkan konstruk yang diukur (Saifuddin, 2021). Oleh karena itu, *rater* perlu memiliki kecermatan, ketelitian, dan cara berpikir yang holistik di dalam memandang dan menilai item-item yang ada di instrumen pengukuran yang divalidasi.

Pengetahuan dan pemahaman terhadap konstruk yang diukur menjadi penting juga karena sebagian penelitian menyusun definisi operasional dan penjelasan tentang aspek atau dimensi dari konstruk yang diukur bukan dari penyusun teorinya secara langsung. Sehingga, penjelasan tersebut berpotensi melenceng dari yang dimaksudkan dari penyusun teori. Dengan demikian, selain kecermatan dan ketelitian terhadap keselarasan antara aspek, indikator keperilakuan, dengan item, *rater* hendaknya mencari berbagai referensi, termasuk dari penyusun teori, untuk mengklarifikasi ketepatan definisi operasional serta penjelasan tentang dimensi/aspek dan indikator keperilakuan.

Manfaat lain dari kecermatan dan ketelitian *rater* adalah memunculkan sikap kritis agar peneliti mengecek ulang keselarasan antara item dengan kondisi sampel penelitian. Terkadang, peneliti membuat item yang cukup kompleks sehingga item tersebut mungkin bisa terjadi pada konteks seseorang tertentu namun tidak terjadi dalam konteks orang lain. Misalkan, terdapat instrumen untuk mengukur penerimaan diri. Salah satu itemnya adalah "Keadaan ekonomi orang tua tidak menjadi penghambat dalam mengembangkan potensi diri". Dalam konteks seseorang yang memang

memiliki tingkat ekonomi yang rendah, item tersebut bisa menjadi relevan. Namun, apabila sampel penelitiannya beragam sehingga ada kemungkinan sebagian sampel memiliki tingkat ekonomi yang rendah maupun tinggi, maka item tersebut menjadi kurang relevan, khususnya bagi sampel yang tingkat ekonominya tinggi. Hal ini disebabkan mereka tidak mengalami konteks item tersebut.

Keterampilan *rater* dalam berpikir holistik ketika memvalidasi instrumen pengukuran ini juga sebagai upaya untuk memenuhi sumber validitas berupa *evidence based on relations to other variables*. Penilaian validitas item dan instrumen pengukuran juga berkaitan dengan bagaimana korelasi antara instrumen pengukuran yang disusun dengan variabel lainnya. Oleh karena itu, *rater* hendaknya memahami bahwa instrumen yang sedang divalidasinya sangat mungkin berkaitan dengan variabel lain sehingga *rater* perlu menggunakan keterampilan dan pengetahuannya untuk mencegah agar instrumen tersebut tidak memiliki item yang tumpang tindih dengan variabel lain. Terlebih lagi, terdapat beberapa variabel yang mirip namun memiliki rumusan dan definisi operasional yang berbeda, misalkan motivasi belajar dan minat belajar.

Di sisi lain, terdapat beberapa variabel yang memiliki aspek yang merupakan variabel lain. Misalkan, di dalam resiliensi terdapat aspek regulasi emosi, optimisme, dan efikasi diri (Reivich & Shatt'e, 2002), di mana dalam konteks lain keempat aspek tersebut merupakan variabel-variabel tersendiri. Misalkan, terdapat variabel regulasi emosi yang disusun oleh Gross (2002), optimisme yang disusun oleh Seligman (2006), dan efikasi diri disusun oleh Bandura (1977, 2006), di mana masing-masing variabel ini memiliki aspek/dimensinya sendiri-sendiri. Ini artinya, *rater* hendaknya mencermati indikator dari ketiga aspek resiliensi tersebut berada dalam koridor definisi operasional yang dirumuskan oleh ketiga ilmuwan tersebut tanpa keluar dari konteks variabel resiliensi.

Keterampilan *rater* dalam berpikir holistik juga dapat memenuhi prinsip *evidence based on internal structure*. *Rater* perlu memiliki kemampuan yang tinggi dan cermat untuk menilai apakah suatu item benar-benar mencerminkan indikator keperilakuan tertentu dan aspek tertentu sehingga selaras. Hal ini disebabkan terkadang peneliti membuat suatu item pada suatu indikator keperilakuan dan aspek namun item tersebut justru lebih cocok untuk indikator keperilakuan dan aspek yang lain.

Keahlian Rater dalam Memahami Asal Instrumen Pengukuran

Kecermatan dan ketelitian juga bermanfaat bagi *rater* untuk memahami asal instrumen pengukuran yang akan divalidasi, apakah instrumen pengukuran tersebut dibuat sendiri oleh peneliti, atau adaptasi, atau modifikasi. Terkadang, peneliti membuat sendiri item-item untuk menyusun instrumen pengukuran yang dibutuhkan berdasarkan teori tertentu. Akan tetapi, terkadang peneliti juga melakukan adaptasi terhadap instrumen yang sudah ada. Adaptasi ini biasanya dilakukan ketika ada instrumen pengukuran yang sudah diuji oleh peneliti lain namun memiliki budaya dan bahasa yang berbeda. Sehingga, seorang peneliti harus melakukan sejumlah rangkaian perubahan/kontekstualisasi bahasa dan penyesuaian konten budaya yang ada di dalam item instrumen tersebut (Ambuehl & Inauen, 2022; Pillet *et al.*, 2023). Tak jarang pula adaptasi instrumen ini dilakukan dalam bentuk merevisi item agar lebih sesuai dengan budaya sampel yang menjadi target penelitian, mengurangi, dan menambah item. Di sisi lain, peneliti juga bisa melakukan

modifikasi instrumen. Modifikasi instrumen dilakukan ketika peneliti mengubah instrumen yang sudah ada tanpa adanya penyesuaian bahasa. Dengan kata lain, modifikasi dilakukan pada instrumen pengukuran yang berbahasa sama dengan bahasa sampel penelitian (Saifuddin, 2020). Di sisi lain, ada juga yang berpendapat bahwa pada dasarnya modifikasi dan adaptasi memiliki kemiripan karena adaptasi adalah salah satu bentuk dari modifikasi dan keduanya mengandung proses perubahan item (Finn & Kayande, 2004) dan respons pengukuran (Steinberg & Rogers, 2022).

Peneliti sering kali melakukan adaptasi terhadap instrumen pengukuran dari peneliti lain yang berasal dari negara lain, sehingga instrumen pengukuran tersebut pada dasarnya berbahasa asing, misalkan berbahasa Inggris. Oleh karena itu, peneliti perlu melakukan prosedur adaptasi dengan tepat. Dalam proses validasi instrumen tersebut, peneliti perlu memberikan keterangan bahwa instrumen tersebut adalah hasil adaptasi. Di sisi lain, *rater* hendaknya juga bersikap teliti sehingga mampu menangkap keterangan tersebut.

Dalam konteks validasi instrumen hasil adaptasi, *rater* juga perlu memahami prosedur adaptasi instrumen pengukuran, misalkan merujuk pedoman yang telah disusun oleh ITC (International Test Commision) (International Test Commision, 2017). Prosedur yang telah ditetapkan tersebut menunjukkan bahwa adaptasi instrumen pengukuran bukan hanya menerjemahkan atau mengalihbahasakan item-item di dalam instrumen pengukuran. Akan tetapi, adaptasi instrumen pengukuran juga persoalan memahami konteks item-item dari bahasa asal sehingga konteks tersebut tidak hilang ketika dialihbahasakan ke bahasa target. Proses alih bahasa ini pun berlangsung dua arah, yaitu dari bahasa asal menuju bahasa target dan dari bahasa target dialihbahasakan ke bahasa asal. Bahkan, dalam proses adaptasi diperbolehkan menambah, mengurangi, atau mengganti item yang ada untuk disesuaikan dengan kondisi budaya sampel penelitian yang ditarget. Selain itu, setelah dialihbahasakan, akan ada uji keterbacaan yang disusul dengan uji validasi.

Prosedur semacam ini harus diinformasikan oleh peneliti, sehingga apabila di dalam form validasi *rater* tidak memperoleh informasi tersebut maka *rater* perlu menanyakan kepada peneliti. Apabila peneliti belum melaksanakan prosedur adaptasi secara tepat, maka *rater* hendaknya tidak bersedia memvalidasi. Sikap semacam ini penting bagi *rater* karena sebagian peneliti didapati kurang memahami prosedur adaptasi. Bahkan, di antara mereka memahami bahwa adaptasi instrumen pengukuran itu sebatas menerjemahkan atau mengalihbahasakan instrumen pengukuran saja. Selain itu, *rater* juga diharapkan memiliki kemampuan memahami konteks yang ada di dalam item sehingga *rater* bisa mempertahankan konteks yang ada di item bahasa asal. Dengan demikian, *rater* bukan hanya memahami item terjemahan secara kebahasaan saja. Atas dasar ini, maka *rater* sebaiknya bukan hanya fokus pada item yang telah dialihbahasakan, tetapi juga mengetahui item dengan bahasa aslinya. Ini artinya, peneliti perlu mencantumkan item dengan bahasa asli maupun item hasil alih bahasa.

Penulis ingin memberikan contoh penting bahwa adaptasi bukan sebatas mengalihbahasakan saja, namun juga mempertahankan konteks dari item yang asli. Davis *et al.* (2015) menyusun instrumen yang mengukur tentang spiritualitas. Item pertama berbunyi, "*I felt near to God*", sedangkan item kedua berbunyi, "*I felt close to God*". Apabila diartikan secara bahasa saja, maka keduanya memiliki arti yang sama, yaitu "saya merasa dekat dengan Tuhan". Akan tetapi, konteks kedekatan pada kedua

item tersebut berbeda. Konteks "dekat" pada item pertama, yaitu makna dari "*near*", adalah perasaan dekat secara fisik, bahwa seseorang merasa berada dalam lingkup pengawasan dan pengetahuan Tuhan. Akan tetapi, konteks "dekat" pada item kedua, yaitu makna dari "*close*", adalah dekat secara emosional. Oleh karena itu, terjemahan dari item nomor satu adalah "saya merasa dekat dengan Tuhan", sedangkan terjemahan dari item nomor dua adalah "saya merasa menyatu dengan Tuhan".

Contoh lain adalah instrumen kecanduan gawai yang disusun oleh Kwon *et al.* (2013) dengan nama *Smartphone Addiction Scale-Short Version*. Terdapat item nomor empat yang berbunyi "*Wont be able to stand not having a smartphone*". Apabila diterjemahkan secara bahasa, maka artinya "Tidak akan tahan apabila tidak memiliki gawai". Akan tetapi, dalam konteks bahasa Indonesia, struktur kalimat tersebut kurang wajar karena tidak memiliki subjek. Di sisi lain, kurang tepat kata "*having*" diterjemahkan menjadi "memiliki". Meskipun secara bahasa terjemah tersebut bisa dianggap tepat, konteks yang diinginkan oleh item tersebut bukan sebatas memiliki, tetapi apakah ponsel atau gawai tersebut berada dalam jangkauannya atau tidak. Konteks ini penting dipertahankan karena instrumen ini hendak mengukur kecanduan gawai. Adapun kecanduan gawai, bukan sebatas mengukur memiliki atau tidak memiliki ponsel/gawai, namun juga mengukur apakah ponsel/gawai tersebut berada dalam jangkauannya atau tidak. Oleh karena itu, terjemahan yang tepat untuk item tersebut adalah "saya tidak tahan untuk tidak memegang ponsel".

Keahlian Rater dalam Memahami Kaidah Item yang Baik

Keterampilan lain yang perlu dikuasai oleh *rater* adalah ketelitiannya di dalam menilai item sesuai dengan kaidah atau tidak. Terdapat beberapa kaidah di dalam membuat item, misalkan tidak mengandung *social desirability* yang tinggi (Azwar, 2021) karena bisa berdampak pada respons dan waktu menjawab (Eichenbrenner & Helmes, 2016). *Social desirability* merupakan kecenderungan seseorang memberikan atau memilih jawaban yang sesuai dengan norma yang berlaku di lingkungan sosial, bukan jawaban yang sebenarnya ada di dalam dirinya (Eichenbrenner & Helmes, 2016). Dengan kata lain, *social desirability* mendorong seseorang memilih jawaban yang seharusnya, bukan yang sebenarnya. Padahal, dalam konteks pengukuran kepribadian, sikap, dan perilaku, seseorang hendaknya memilih jawaban yang sebenarnya, bukan seharusnya. Oleh karena itu, *social desirability* yang tinggi bisa merusak kejujuran sampel penelitian dalam menjawab yang pada akhirnya bisa memunculkan bias pengukuran. Dengan demikian, *social desirability* perlu dikendalikan (Larson, 2018). Potensi *social desirability* pada item perlu ditangkap oleh *rater* sehingga menjadi pertimbangan di dalam memvalidasi item. Kemampuan *rater* dalam menilai item yang *social desirability* ini berkaitan dengan sumber validitas berupa *evidence based on response processes*.

Sumber validitas berupa *evidence based on response processes* ini juga bisa diupayakan dengan keterampilan *rater* dalam menilai efektivitas kalimat dalam item-item. Semakin efektif suatu kalimat, maka akan semakin mudah dipahami. Keefektifan kalimat ini juga bisa dinilai dengan *cognitive labs*. *Cognitive labs* merupakan strategi untuk mengecek tingkat pemahaman responden penelitian terhadap item pernyataan, instruksi pengerjaan, dan kesesuaian respons dengan item pernyataan (Willis, 1999; Willis *et al.*, 1991; Wilson & Peterson, 1999). Dalam konteks *rater*, *rater* dapat memposisikan dirinya

sebagai responden yang awam ketika menilai instruksi pengerjaan instrumen, item, dan bentuk respons dalam instrumen pengukuran. Kondisi ini membantu *rater* untuk menilai apakah instruksi sudah cukup jelas, tidak ambigu, dan mudah dipahami. Begitu pula itemnya apakah juga sudah mudah dipahami dan tidak multitafsir serta menilai kesesuaian antara item denbgan bentuk jawaban dalam instrumen pengukuran.

Kaidah lain adalah item dibuat dalam konteks saat ini. Item juga bisa dibuat dalam konteks perilaku di dalam item tersebut sudah dilakukan atau seolah-olah dilakukan. Sehingga, item yang baik tidak mengandung kata "akan", misalkan "saya akan mendengarkan penjelasan dosen ketika kuliah". Kata akan menyebabkan kerentanan terhadap *social desirability* yang tinggi karena kalimat yang mengandung kata "akan" cenderung mencerminkan idealitas, bukan realitas. Selain itu, kata akan juga mengindikasikan bahwa perilaku tersebut belum terjadi. Padahal, pengukuran hendaknya mengukur perilaku atau sikap yang sedang, sudah terjadi, atau seandainya terjadi, bukan akan terjadi (Saifuddin, 2020, 2021). *Rater* perlu menangkap apakah item-item yang ada di dalam instrumen pengukuran yang divalidasi terdapat penggunaan kata akan atau tidak.

Karakteristik item yang baik yang perlu dipahami oleh *rater* dan menjadi pertimbangan *rater* dalam memvalidasi item adalah apakah item tersebut berisi sikap atau perilaku yang relevan dengan sampel pengukuran. Ini artinya, *rater* harus memastikan apakah sikap atau perilaku yang terkandung di dalam item kemungkinan besar ada di dalam sampel pengukuran atau tidak. Hal ini didasarkan pada pemahaman bahwa suatu konstruk pengukuran hendaknya dicerminkan dari perilaku atau fenomena yang hendak diukur (Schmittmann *et al.*, 2013).

Etika Rater dalam Memvalidasi Instrumen Pengukuran

Selain keterampilan dan kemampuan yang telah dijelaskan, *rater* juga memiliki etika lain. Misalkan, *rater* hendaknya selalu memperbarui pengetahuannya sehingga mampu memberikan saran yang kontekstual dan terkini. Selain itu, apabila seseorang merasa tidak memiliki kemampuan untuk memvalidasi instrumen pengukuran, maka sebaiknya seseorang tersebut tidak bersedia menjadi *rater*.

Ketika merujuk bahwa *rater* hendaknya memiliki sejumlah keterampilan, maka bisa memunculkan permasalahan lain, yaitu ketersediaan seseorang yang memiliki keterampilan tersebut. Dengan kata lain, apabila *rater* benar-benar harus memiliki berbagai macam keterampilan, maka seseorang yang memiliki keterampilan tersebut tidak banyak. Namun, sangat mungkin dalam satu waktu banyak peneliti yang membutuhkan *rater* untuk memvalidasi instrumen pengukuran mereka. Di sisi lain, setiap peneliti selalu membutuhkan *rater* dalam jumlah banyak. Dalam hal ini, setiap akademisi hendaknya belajar untuk menguasai keterampilan menjadi *rater* sehingga ketersediaan *rater* yang berkompetensi menjadi banyak. Adapun peneliti hendaknya mengecek ulang kompetensi *rater* melalui berbagai rekam karya dari seseorang yang akan dimintai menjadi *rater*, misalkan melalui akun Google Scholar, *website* institusi tempat seseorang yang akan dimintai menjadi *rater* bekerja, atau *platform* lainnya.

Penutup

Validitas konten atau verifikasi isi dalam instrumen pengukuran menjadi salah satu prosedur penting, meskipun bisa ditunjang dengan jenis validitas lain. Oleh karena itu, validitas konten hendaknya dilakukan dengan prosedur yang tepat. Salah satu variabel yang menentukan berkualitasnya hasil validitas konten adalah keterampilan dan pengetahuan yang dimiliki oleh *rater*. Maka dari itu, penting bagi *rater* untuk memiliki sejumlah keterampilan, misalkan keterampilan memahami domain pengukuran, sifat item, keselarasan antara bentuk instrumen pengukuran dengan domain yang diukur, kecermatan dan ketelitian, pemahaman tentang adaptasi instrumen pengukuran, serta pemahaman terhadap kaidah item yang baik.

Saran

Keterampilan dan pengetahuan *rater* menjadi salah satu faktor penting untuk menghasilkan validasi instrumen pengukuran yang berkualitas. Oleh karena itu, setiap *rater* atau akademisi didorong untuk selalu meningkatkan dan memperbarui keterampilan dan pengetahuannya sehingga dapat memvalidasi instrumen pengukuran dengan baik. Selain itu, setiap akademisi juga didorong untuk mempelajari berbagai keterampilan yang diperlukan untuk menjadi *rater* agar jumlah ketersediaan *rater* menjadi banyak. Ketika seseorang tidak memiliki keterampilan sebagai seorang *rater* atau tidak memahami konstruk yang diukur, maka hendaknya seseorang tersebut tidak bersedia menjadi *rater*. Di sisi lain, para peneliti didorong untuk menelusuri dengan cermat tentang kapasitas seseorang yang akan dimintai menjadi *rater* melalui rekam karyanya.

Pernyataan

Ucapan Terima Kasih

Terima kasih banyak diucapkan kepada para *reviewer* yang telah memberikan komentar dan catatan yang konstruktif serta saran referensi. Terima kasih juga disampaikan kepada Muhammad Dwirifqi Kharisma Putra, S.Psi., M.Si. (Fakultas Psikologi, Universitas Gadjah Mada Yogyakarta) yang telah membantu penulis dalam bentuk memberikan pemahaman dan arahan serta sejumlah referensi dalam merevisi artikel ini.

Pendanaan

Artikel ini tidak melibatkan pendanaan dari pihak manapun.

Kontribusi Penulis

Penulis adalah tunggal sehingga artikel ini merupakan karya dan hasil pemikiran dari penulis.

Pernyataan Konflik Kepentingan

Penulis menyatakan tidak ada konflik kepentingan dengan pihak mana pun.

Orcid ID

Ahmad Saifuddin https://orcid.org/0000-0002-3863-8953

Daftar Pustaka

AERA, APA, and NCME. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

Aiken, L. R. (1985). Three coefficients for analyzing the reliability and validity of ratings. *Educational And Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Ambuehl, B., & Inauen, J. (2022). Contextualized measurement scale adaptation: A 4-Step tutorial for health psychology research. *International Journal of Environmental Research and Public Health*, *19*(12775), 1–24. https://doi.org/10.3390/ijerph191912775

Aros, M., Narvaez, G., & Aros, N. H. (2009). The semantic differential for the discipline of design: a tool for the product evaluation. *Product Engineering*, 422–433.

Azwar, S. (2016). *Reliabilitas dan validitas* (4th Ed). Pustaka Pelajar.

Azwar, S. (2021). *Penyusunan skala psikologi* (3rd Ed). Pustaka Pelajar.

Bandura, A. (1977). Self-efficacy: Toward a unifying theory of behavioral change. *Psychological Review*, *84*(2), 191–215. https://doi.org/10.1037/0033-295X.84.2.191

Bandura, A. (2006). Guide for constructing self-efficacy scales. In F. Pajares & T. Urdan (Eds.), *Self-efficacy beliefs of adolescents* (pp. 307–337). https://doi.org/10.1017/CBO9781107415324.004

Berk, R. A. (1990). Importance of expert judgment in content-related validity evidence. *Western Journal of Nursing Research*, *12*(5), 659–671. https://doi.org/10.1177/019394599001200507

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2003). The theoretical status of latent variables. *Psychological Review*, *110*(2), 203–219. https://doi.org/10.1037/0033-295X.110.2.203

Cizek, G. J. (2015). Validating test score meaning and defending test score use: different aims, different methods. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 212–225. https://doi.org/10.1080/0969594X.2015.1063479

Cohen, R. J., Schneider, W. J., Tobin, R., Swerdlik, M., & Sturman, E. (2022). *Psychological testing and assessment: An introduction to tests and measurement*. McGraw Hill.

Davis, D. E., Rice, K., Hook, J. N., Van Tongeren, D. R., DeBlaere, C., Choe, E., & Worthington, E. L. (2015). Development of the sources of spirituality scale. *Journal of Counseling Psychology*, *62*(3), 503–513. https://doi.org/10.1037/cou0000082

Drost, E. A. (2011). Validity and reliability in social science research. *Educational Research And Perspectives*, *38*(1), 105–123.

Eichenbrenner, L.-E., & Helmes, E. (2016). Social desirability and affect: Linking domains of content. *Advances in Social Sciences Research Journal*, *3*(11), 119–125. https://doi.org/10.14738/assrj.311.2277

Embretson, S. E. (2007). Construct validity: A universal validity system or just another test evaluation procedure? *Educational Researcher*, *36*(8), 449–455. https://doi.org/10.3102/0013189x07311600

Finn, A., & Kayande, U. (2004). Scale modification: Alternative approaches and their consequences. *Journal of Retailing*, *80*(1), 37–52. https://doi.org/10.1016/j.jretai.2004.01.003

Fishman, J., Yang, C., & Mandell, D. (2021). Attitude theory and measurement in implementation science: a secondary review of empirical studies and opportunities for advancement. *Implementation Science*, *16*(87), 1–10. https://doi.org/10.1186/s13012-022-01204-9

Gafni, N. (2016). Comments on implementing validity theory. *Assessment in Education: Principles, Policy & Practice*, *23*(6), 284–286. https://doi.org/10.1080/0969594X.2015.1111195

Gorin, J. S. (2007). Reconsidering issues in validity theory. *Educational Researcher*, *36*(8), 456–462. https://doi.org/10.3102/0013189X07311607

Gross, J. J. (2002). Emotion regulation: Affective, cognitive, and social consequences. *Psychophysiology*, *39*, 281–291. https://doi.org/10.1017.S0048577201393198

Haddock, G., & Maio, G. R. (2008). Attitudes: content, dtructure and functions. In M. Hewstone, W. Stroebe, & K. Jonas (Eds.), *Introduction to social psychology: A european perspective* (4th Ed, pp. 112–133). Blackwell Publishing.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art*. John Hopkins University Press.

Henriques, G., & Michalski, J. (2020). Defining behavior and its relationship to the science of psychology. *Integrative Psychological and Behavioral Science*, *54*(2). https://doi.org/10.1007/s12124-019-09504-4

Holt, G. D. (2014). Asking questions, analysing answers: relative importance revisited. *Construction Innovation*, *14*(1), 2–16. https://doi.org/10.1108/CI-06-2012-0035

International Test Commision. (2017). *The ITC guidelines for translating and adapting tests* (2nd Ed). https://doi.org/005

Kane, M. T. (2015). Explicating validity. *Assessment in Education: Principles, Policy & Practice*, *23*(2), 198–211. https://doi.org/10.1080/0969594X.2015.1060192

Kaplan, R. M., & Saccuzzo, D. P. (2017). *Psychological testing: Principles, applications, and issues* (9th Ed). Cengage Learning.

Knekta, E., Runyon, C., & Eddy, S. (2019). One size doesn't fit all: Using factor analysis to gather validity evidence when using surveys in your research. *CBE Life Sciences Education*, *18*(1). https://doi.org/10.1187/cbe.18-04-0064

Kwon, M., Kim, D. J., Cho, H., & Yang, S. (2013). The smartphone addiction scale: Development and validation of a short version for adolescents. *PLoS ONE*, *8*(12), e83558. https://doi.org/10.1371/journal.pone.0083558

Kyllonen, P. C., & Zu, J. (2016). Use of response time for measuring cognitive ability. *Journal of Intelligence*, *4*(4), 1–29. https://doi.org/10.3390/jintelligence4040014

Lane, S. (2014). Validity evidence based on testing consequences. *Psycothema*, *26*(1), 127–135. https://doi.org/10.7334/psicothema2013.258

Larson, R. B. (2018). Controlling social desirability bias. *International Journal of Market Research*, *61*(5), 534–547. https://doi.org/10.1177/1470785318805305

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

Likert, R. (1932). A tehcnique for the measurement of attitudes. *Archives of Psychology*, *22*(140), 5–55.

Lissitz, R. W., & Samuelsen, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, *36*(8), 437–448. https://doi.org/10.3102/0013189X07311286

Lozano, L. M., Garc'ia-Cueto, E., & Muniz, J. (2008). Effect of the number of response categories on the reliability and validity of rating scales. *Methodology*, *4*(2), 73–79. https://doi.org/10.1027/1614-2241.4.2.73

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement*. Collier Macmillan Publishers.

Moss, P. A. (2007). Reconstructing validity. *Educational Researcher*, *36*(8), 470–476. https://doi.org/10.3102/0013189X07311608

Padilla, J. L., & Ben'itez, I. (2014). Validity evidence based on response processes. *Psicothema*, *26*(1), 136–144. https://doi.org/10.7334/psicothema2013.259

Pillet, J.-C., Carillo, K. D., Vitari, C., & Pigni, F. (2023). Improving scale adaptation practices in information systems research: Development and validation of a cognitive validity assessment method. *Focus On Research Methods*, *33*(4), 842–889. https://doi.org/10.1111/isj.12428

Reivich, K., & Shatt'e, A. (2002). *The resilience factor: 7 essential skills for overcoming life's inevitable obstacles*. Broadway Books.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psycothema*, *26*(1), 108–116. https://doi.org/10.7334/psicothema2013.260

Rosenberg, B. D., & Navarro, M. A. (2018). Semantic differential scaling. In B. B. Frey (Ed.), *The SAGE Encyclopedia of Educational Research, Measurement, and Evaluation*. SAGE Publications, Inc. https://doi.org/10.4135/9781506326139.n624

Rubio, D. M., Berg-Weger, M., Tebb, S. S., Lee, E. S., & Rauch, S. (2003). Objectifying content validity: Conducting a content validity study in social work research. *Special Work Research*, *27*(2), 94–104. https://doi.org/10.1093/swr/27.2.94

Saifuddin, A. (2020). *Penyusunan skala psikologi (Pertama)*. KENCANA (Divisi dari Prenadamedia Group).

Saifuddin, A. (2021). *Validitas dan reliabilitas alat ukur psikologi*. RajaGrafindo Persada.

Schmittmann, V. D., Cramer, A. O., Waldorp, L. J., Epskamp, S., Kievit, R. A., & Borsboom, D. (2013). Deconstructing the construct: A network perspective on psychological phenomena. *New Ideas in Psychology*, *31*(1), 43–53. https://doi.org/10.1016/j.newideapsych.2011.02.007

Seligman, M. E. P. (2006). *Learned optimism: How to change your mind and yur life*. Vintage Books.

Stark, R., & Glock, C. Y. (1968). *American piety: The nature of religious commitment*. University of California Press.

Steinberg, L., & Rogers, A. (2022). Changing the scale: The effect of modifying response scale labels on the measurement of personality and affect. *Multivariate Behavioral Research*, *57*(1), 79–93. https://doi.org/10.1080/00273171.2020.1807305

Syaiful, I. A., & Roebianto, A. (2020). Adapting and examining the factor structure of the self-compassion scale in Indonesian version. *Jurnal Psikologi*, *47*(3), 175–205. https://doi.org/10.22146/jpsi.57608

van Heerden, D. B. J., & Mellenbergh, G. J. (2003). Validity and truth. In H. Yanai, A. Okada, K. Shigemasu, Y. Kano, & J. J. Meulman (Eds.), *New developments in psychometrics* (pp. 1–8). Springer. https://doi.org/10.1007/978-4-431-66886-8_36

Widhiarso, W. (2016). Peranan butir unfavorabel dalam menghasilkan dimensi baru dalam pengukuran psikologi. *Jurnal Psikologi Perseptual*, *1*(1), 40–52. https://doi.org/10.24176/perseptual.v1i1.1078

Willis, G. B. (1999). *Cognitive Interviewing: A "How To" Guide, Reducing Survey Error through Research on the Cognitive and Decision Processes in Surveys* (tech. rep.).

Willis, G. B., Royston, P., & Bercini, D. (1991). The use of verbal report methods in the development and testing of survey questionnaires. *Applied Cognitive Psychology*, *5*(3), 251–26. https://doi.org/10.1002/acp.2350050307

Wilson, B. F., & Peterson, L. S. (1999). Using the NCHS cognitive lab to help design cycle VI of the national survey of family growth. *Proceedings of the Survey Research Methods Section, American Statistical Association*, 997–1002.