

Upaya Penyempurnaan Alat Ukur: Analisis Validitas PAPI Kostick Item Revisi

Fadillah, Institut Teknologi Bandung, fadillah@itb.ac.id

Medianta Tarigan, Universitas Pendidikan Indonesia, medianta@upi.edu

Abstract

This study aimed to evaluate the psychometric quality of the PAPI Kostick instrument in the Indonesian context through item revision, construct validity testing, and item-level analysis. The sample consisted of **1,289 respondents**. Item revisions were conducted through linguistic and cultural adaptation to ensure semantic clarity and construct equivalence. Construct validity was examined using CFA, which demonstrated a good fit across all dimensions, with RMSEA values below 0.08. Most items were valid, with the highest proportion in Activity (100%). Further evaluation using the two-parameter logistic (2PL) IRT model indicated that 74–85% of items were located at the mid-range of the trait continuum, while discrimination indices were generally strong, with Activity achieving 100% in the normal category, followed by Leadership (89%). Overall, the revised version of PAPI Kostick demonstrated adequate construct validity, balanced item distribution, and satisfactory diagnostic properties, while continuous refinement remains necessary.

Keyword: PAPI Kostick, validity, CFA, IRT

Abstrak

Penelitian ini bertujuan untuk mengevaluasi kualitas psikometris instrumen PAPI Kostick dalam konteks Indonesia melalui revisi item, uji validitas konstruk, dan analisis kualitas butir. Sampel penelitian terdiri dari 1.289 responden. Revisi item dilakukan melalui telaah bahasa dan konteks budaya lokal untuk memastikan kesesuaian makna dengan konstruk. Validitas konstruk dianalisis menggunakan CFA, yang menunjukkan hasil good fit dengan nilai RMSEA < 0,08 pada seluruh dimensi. Mayoritas item dinyatakan valid, dengan proporsi tertinggi pada dimensi Activity (100%). Analisis IRT model 2PL menunjukkan bahwa 74–85% item berada pada posisi menengah pada kontinum konstruk, dengan daya beda normal tertinggi pada dimensi Activity (100%), diikuti Leadership (89%). Secara keseluruhan, hasil ini menunjukkan bahwa instrumen versi revisi memiliki validitas konstruk yang memadai, distribusi item yang seimbang, dan kualitas diagnostik yang baik, meskipun evaluasi berkelanjutan tetap diperlukan.

Keyword: PAPI Kostick, validitas, CFA, IRT

Dalam dunia kerja modern, penggunaan instrumen asesmen psikologis telah menjadi praktik yang lazim dalam proses seleksi dan pengembangan karyawan. Instrumen-instrumen ini digunakan untuk memprediksi potensi kinerja individu, memahami karakteristik kepribadian, serta menilai kesesuaian antara kandidat dengan peran atau budaya organisasi

tertentu (Tang, 2023). Di Indonesia, praktik penggunaan psikotes dalam rekrutmen bukanlah hal yang baru. Salah satu instrumen yang cukup populer dan banyak digunakan adalah PAPI Kostick (Personality and Preference Inventory Kostick), yang dirancang untuk mengevaluasi preferensi kerja dan kepribadian seseorang dalam konteks profesional.

PAPI Kostick pertama kali dikembangkan pada tahun 1960-an oleh Dr. Max Kostick, seorang psikolog asal Amerika Serikat. Tujuan utama pengembangan alat ini adalah untuk menciptakan instrumen asesmen yang tidak hanya praktis, tetapi juga mampu mengevaluasi berbagai aspek penting dalam perilaku kerja secara komprehensif (Furnham & Craig, 1987). Dasar teoretis dari PAPI merujuk pada teori "need-press" yang dikembangkan oleh Henry Murray (1938), yang menekankan interaksi antara kebutuhan internal individu dan tekanan eksternal dari lingkungan. Di Indonesia, PAPI mulai dikenal sejak dekade 1980-an dan mengalami lonjakan popularitas pada akhir 1990-an (Putri et al., 2022). Tes ini mengukur tujuh dimensi utama yang dianggap penting dalam dunia kerja, yakni arah kerja (work direction), kepemimpinan (leadership), aktivitas (activity), sifat sosial (social nature), gaya kerja (work style), temperamen (temperament), dan peran sebagai bawahan (followership) (Jiandy et al., 2017).

Secara konseptual, PAPI hadir dalam dua versi utama, yaitu PAPI-I (Ipsative), yang digunakan untuk pengembangan diri, dan PAPI-N (Normative), yang lebih sesuai untuk tujuan seleksi dan benchmarking antar individu (Forman et al., 2009). Versi normative memberikan skor absolut berdasarkan skala yang seragam bagi semua responden, sehingga memungkinkan analisis komparatif antar individu. Sebaliknya, versi ipsative menghasilkan profil relatif antar dimensi dalam diri individu, namun tidak memungkinkan perbandingan normatif.

Meskipun demikian, dalam praktiknya di Indonesia, versi PAPI-I tetap banyak digunakan dalam konteks seleksi karyawan. Salah satu alasan utama di balik keberlanjutan penggunaan PAPI-I untuk seleksi adalah formatnya yang menggunakan forced-choice, di mana individu harus memilih antara dua pernyataan yang sama-sama positif atau netral. Format ini dianggap lebih tahan terhadap manipulasi jawaban atau faking good dibandingkan dengan format Likert scale pada versi normatif, yang memungkinkan responden memilih opsi yang secara langsung mencerminkan citra diri ideal (Dalal et al.,

2021; Joubert et al., 2015). Dalam konteks seleksi yang kompetitif, keunggulan ini menjadikan PAPI-I tetap menarik bagi banyak praktisi asesmen, karena lebih mampu mengurangi pengaruh bias sosial dalam respons kepribadian.

Namun demikian, ketahanan terhadap faking good dari format forced-choice ini bukan tanpa batas. Dengan semakin mudahnya akses terhadap contoh-contoh item PAPI Kostick di berbagai platform daring, banyak individu kini dapat mempelajari pola-pola jawaban dan memahami preferensi organisasi terhadap tipe kepribadian tertentu. Hal ini secara signifikan mengurangi efektivitas format ipsative sebagai penghalang terhadap manipulasi respons. Penelitian sebelumnya telah menunjukkan bahwa PAPI tetap rentan terhadap manipulasi jawaban, karena banyak individu memiliki pemahaman intuitif tentang cara menyesuaikan respons mereka agar sesuai dengan profil ideal dalam konteks seleksi kerja (Furnham & Craig, 1987; Kuric et al., 2025). Dalam konteks seleksi kerja, hal ini dapat menghasilkan keputusan yang keliru dan berdampak negatif terhadap efektivitas rekrutmen.

Selain tantangan manipulasi respons, penerapan PAPI Kostick di Indonesia juga menghadapi persoalan serius dalam hal kualitas kebahasaan dan relevansi budaya. Sebagai instrumen yang berasal dari luar negeri, terjemahan item secara literal dari bahasa Inggris ke bahasa Indonesia sering kali dilakukan tanpa melalui proses adaptasi budaya yang memadai. Hal ini berpotensi menimbulkan bias konstruk dan mengaburkan makna asli item, sehingga responden dapat memberikan interpretasi yang tidak sesuai dengan maksud pengembang instrumen (Swanepoel & Krüger, 2011). Kondisi ini dapat melemahkan validitas hasil asesmen dan menyebabkan kesalahan dalam pengambilan keputusan berbasis data psikologis.

Terlepas dari tingginya penggunaan PAPI Kostick dalam berbagai konteks industri di Indonesia, studi yang secara spesifik mengevaluasi kualitas psikometrisnya masih sangat terbatas. Tidak banyak penelitian yang menelaah secara mendalam sejauh mana versi adaptasi PAPI Kostick saat ini mampu mengukur konstruk kepribadian secara valid dan reliabel di kalangan responden Indonesia. Oleh karena itu, diperlukan upaya sistematis untuk menilai kembali kualitas butir-butir item dalam versi saat ini, dan jika ditemukan kekurangan, perlu dilakukan revisi melalui proses adaptasi bahasa yang lebih kontekstual dan sensitif terhadap budaya lokal.

Berangkat dari permasalahan tersebut, tujuan utama dari penelitian ini adalah untuk mengidentifikasi dan merevisi kualitas item-item dalam instrumen PAPI Kostick versi adaptasi yang digunakan saat ini di Indonesia. Penelitian ini kemudian merevisi item-item yang bermasalah dengan pendekatan adaptasi bahasa dan budaya yang lebih sesuai, sehingga makna item dapat dipertahankan namun tetap relevan. Revisi ini tidak hanya bertujuan memperbaiki pemahaman dan keakuratan interpretasi, tetapi juga untuk meminimalkan potensi faking good, meningkatkan validitas skor asesmen, dan menyediakan item-item baru yang belum tersebar luas di internet, sehingga mengurangi kemungkinan manipulasi oleh peserta yang telah mengakses materi secara daring sebelumnya.

Selanjutnya, untuk mengevaluasi struktur laten dari konstruk yang diukur dalam instrumen ini, akan dilakukan analisis Confirmatory Factor Analysis (CFA) guna menguji kesesuaian antara struktur teoretis dimensi PAPI Kostick dengan data empirik hasil pengukuran. CFA digunakan untuk memastikan bahwa dimensi yang diukur sesuai dengan model yang diharapkan, serta untuk menilai sejauh mana indikator-item memuat konstruk yang tepat sesuai dengan asumsi teoretis. Hasil dari analisis CFA akan memberikan informasi penting mengenai model fit, factor loadings, dan kontribusi masing-masing item terhadap konstruk yang diukur, sehingga dapat memperkuat validitas konstruk dari versi yang telah direvisi.

Selain itu, instrumen yang telah direvisi juga akan dianalisis menggunakan pendekatan Teori Respon Butir (Item Response Theory/IRT) guna memperoleh informasi lebih mendalam terkait karakteristik item, seperti tingkat diskriminasi dan kesesuaian tingkat kesulitan terhadap populasi responden. Penggunaan IRT dalam konteks ini diharapkan dapat memberikan gambaran yang lebih akurat mengenai performa setiap item secara individual, serta mendukung proses kalibrasi item secara statistik agar hasil asesmen kepribadian dapat dipertanggungjawabkan secara ilmiah dan praktis.

Melalui pendekatan tersebut, penelitian ini diharapkan dapat memberikan kontribusi signifikan dalam peningkatan validitas konstruk, validitas isi, dan reliabilitas internal instrumen PAPI Kostick di Indonesia. Hasil penelitian ini juga diharapkan dapat menjadi dasar bagi pengembangan instrumen kepribadian yang lebih adaptif terhadap konteks lokal

dan lebih tahan terhadap manipulasi respons, sehingga tetap relevan dan efektif digunakan dalam proses seleksi serta pengembangan sumber daya manusia di berbagai organisasi.

Metode

Penelitian ini menggunakan pendekatan kuantitatif dengan desain *instrumental development study*, yang bertujuan untuk mengevaluasi kualitas psikometris item-item dalam instrumen PAPI Kostick versi adaptasi bahasa Indonesia. Proses penelitian dilakukan dalam tiga tahap utama: (1) evaluasi kualitas item versi adaptasi saat ini, (2) revisi item berdasarkan hasil evaluasi dan pertimbangan adaptasi bahasa-budaya, dan (3) analisis kualitas item hasil revisi menggunakan *Confirmatory Factor Analysis* (CFA) dan *Item Response Theory* (IRT).

Data dalam penelitian ini dikumpulkan secara daring melalui platform *online test* di <https://psikotes.anargya.id>, dengan total partisipan sebanyak 1289 orang setelah dilakukan *cleaning* data. Pengambilan data dilakukan pada tahun 2024. Tabel 1 berikut ini menyajikan gambaran umum mengenai demografi partisipan.

Tabel 1
Data Demografi

		Total	Percentase
Jenis Kelamin	Laki-Laki	552	43%
	Perempuan	737	57%
Usia	< 20 tahun	112	9%
	20–24 tahun	646	50%
	25–29 tahun	234	18%
	30–34 tahun	124	10%
	35–39 tahun	88	7%
	40+ tahun	85	7%

Instrumen PAPI Kostick versi revisi terdiri atas 90 item yang mengukur tujuh dimensi kepribadian kerja, yaitu *work direction, leadership, activity, social nature, workstyle, temperament*, dan *followership*. Item disusun dalam format *forced choice*, di mana partisipan diminta memilih salah satu dari dua pernyataan yang paling menggambarkan dirinya. Versi revisi ini juga telah mengalami perbaikan bahasa dan penyesuaian makna item agar lebih sesuai dengan konteks budaya serta penggunaan sehari-hari dalam bahasa Indonesia. Rincian aspek yang membentuk masing-masing dimensi disajikan pada Tabel berikut.

Tabel 2
Aspek PAPI Kostick

Dimensi	Aspek	Total Pernyataan	Bentuk Pernyataan	Skor Jawaban
<i>Work Direction</i>	1. N (<i>need to finish task</i>)	180	<i>Formed Choice</i>	Terpilih (1)/Tidak terpilih (0)
	2. G (<i>role of hard intense worker</i>)			
	3. A (<i>need to achieve</i>)			
<i>Leadership</i>	1. L (<i>leadership role</i>)			
	2. P (<i>need to control others</i>)			
<i>Activity</i>	3. I (<i>case in decision making</i>)			
	1. T (<i>pace</i>)			
	2. V (<i>vigorous type</i>)			
<i>Social Nature</i>	1. O (<i>need for chance & affection</i>)			
	2. B (<i>need to belong to groups</i>)			
	3. S (<i>social extension</i>)			
	4. X (<i>need to be noticed</i>)			
<i>Workstyle</i>	1. C (<i>organized type</i>)			
	2. D (<i>interest in working with details</i>)			
	3. R (<i>theoretical type</i>)			
<i>Temperament</i>	1. Z (<i>need for change</i>)			
	2. E (<i>emotional restraint</i>)			
	3. K (<i>need to be forceful</i>)			
<i>Followership</i>	1. F (<i>need to support authority</i>)			
	2. W (<i>need for rules & supervision</i>)			

Penelitian ini dilakukan melalui beberapa tahapan sistematis yang bertujuan untuk mengevaluasi dan meningkatkan kualitas psikometris instrumen PAPI Kostick versi adaptasi bahasa Indonesia. Tahapan prosedur secara rinci dijelaskan sebagai berikut:

1) Analisis Awal terhadap Versi Saat Ini

Proses dimulai dengan pengumpulan data menggunakan versi PAPI Kostick yang telah beredar luas dalam versi adaptasi bahasa Indonesia secara kualitatif.

2) Revisi Item Berdasarkan Hasil Analisis dan Telaah Bahasa

Item-item yang menunjukkan kualitas rendah (misalnya terjemahan tidak sesuai dengan bentuk kalimat asli dalam bahasa Inggris, konteks kata yang kurang tepat atau pernyataan item tidak sesuai dengan aspek yang diukur) ditelaah lebih lanjut. Proses revisi dilakukan melalui pendekatan kualitatif dan linguistik dengan langkah-langkah sebagai berikut:

- a. Penelaahan ulang terhadap redaksi bahasa pada setiap item bermasalah untuk mengidentifikasi potensi bias terjemahan, ambiguitas makna, atau ketidaksesuaian konteks budaya lokal.
- b. Perbandingan setiap item bermasalah dengan versi asli dalam bahasa Inggris, untuk meninjau kesetaraan makna dan relevansi konstruk psikologis.
- c. Penyusunan ulang redaksi item melalui proses adaptasi budaya dan bahasa yang lebih kontekstual, melibatkan ahli psikologi industri dan organisasi, ahli bahasa, serta tim validasi isi independen untuk memastikan koherensi semantik dan fungsional.

3) Analisis terhadap Versi Revisi

Setelah proses adaptasi selesai, versi revisi PAPI Kostick digunakan untuk pengumpulan data baru pada populasi yang serupa. Analisis CFA dan IRT kembali dilakukan pada versi revisi ini untuk:

- a. Menguji kembali struktur faktor melalui CFA, guna memastikan bahwa modifikasi item tidak merusak integritas konstruk dan bahkan meningkatkan kesesuaian model secara keseluruhan.
- b. Mengevaluasi ulang karakteristik item melalui IRT, untuk menilai peningkatan parameter diskriminasi dan sebaran tingkat kesulitan. Perbandingan dilakukan antara performa item sebelum dan sesudah revisi guna menentukan efektivitas perubahan yang dilakukan.

Prosedur ini dirancang untuk memastikan bahwa setiap revisi didasarkan pada bukti empiris dan penyesuaian linguistik yang sesuai.

Data yang diperoleh selanjutnya dianalisis untuk menguji validitas konstruk menggunakan metode *Confirmatory Factor Analysis* (CFA). Analisis ini digunakan untuk menentukan sejauh mana seluruh item dalam instrumen PAPI Kostick versi revisi mengukur faktor yang sama. Tahapan pertama yang dilakukan dalam CFA ini yaitu pengujian *Goodness of Fit* (GoF) untuk melihat kecocokan model. Uji kecocokan model dapat dilihat dari nilai RMSEA, GFI, CFI dan TLI. Jika nilai RMSEA yang didapat kurang dari 0,05 maka mengindikasikan bahwa model memiliki tingkat kecocokan yang sangat baik (*close fit*) (Baumgartner & Homburg, 1996). Model akan dikatakan *good fit* jika nilai GFI yang didapat lebih dari 0,9 (S & Mohanasundaram, 2024), serta nilai CFI dan TLI lebih dari 0,8 (Akkus, 2019). Jika model

sudah dinyatakan sesuai, maka analisis dapat dilanjutkan ke tahap uji signifikansi. Sebuah item akan dinyatakan valid apabila memiliki nilai *factor loading* lebih dari 0,5 atau nilai *p-value* kurang dari 0,05 (Ondé & Alvarado, 2020).

Setelah pengujian CFA, maka akan dilanjutkan dengan mengevaluasi kualitas butir soal melalui pendekatan *Item Response Theory* (IRT). *Item Response Theory* (IRT) adalah kumpulan model statistik yang digunakan untuk memahami bagaimana seseorang menjawab soal-soal dalam tes berdasarkan tingkat kemampuan yang dimilikinya (Foster et al., 2017). Dalam penelitian ini digunakan model dua parameter logistik (2PL), karena lebih sesuai untuk data dikotomis, seperti pada item-item dalam inventori kepribadian PAPI Kostick. Model ini telah banyak digunakan dalam penelitian kepribadian, khususnya untuk respons dikotomis yang tidak melibatkan unsur menebak atau kecenderungan menjawab setuju secara otomatis. Oleh karena itu, model 2PL dinilai tepat untuk diterapkan pada konteks PAPI Kostick (Chernyshenko et al., 2001).

Model 2PL mengevaluasi dua parameter utama, yaitu daya beda (*discrimination*) dan tingkat kesulitan (*difficulty*). Parameter daya beda mencerminkan sejauh mana suatu item mampu membedakan peserta yang memiliki kemampuan tinggi dan rendah. Sementara itu, parameter tingkat kesulitan (*difficulty*) menggambarkan seberapa besar peluang seseorang dengan tingkat kemampuan tertentu untuk menjawab suatu butir dengan benar (Mutiawani et al., 2022). Beberapa kriteria digunakan untuk menginterpretasikan model IRT. Untuk parameter daya beda, skor antara 0–2 dikategorikan sebagai normal. Sementara itu, untuk parameter tingkat kesulitan, skor kurang dari -2, antara -2 hingga 2, dan lebih dari 2 masing-masing diklasifikasikan sebagai kategori mudah, rata-rata, dan sulit (Baker, 2001).

Hasil

Analisis Deskriptif

Analisis deskriptif dilakukan untuk memberikan gambaran umum mengenai sampel dalam penelitian ini. Deskripsi data dianalisis menggunakan beberapa ukuran statistik, dengan hasil yang disajikan pada Tabel 3.

Tabel 3
Analisis Deskriptif

Dimensi	Mean	Standard Deviation	Variance	Kurtosis	Skewness	Minimum	Maximum
<i>Work Direction</i>	14,31	2,95	8,71	-0,08	0,008	5	23
<i>Leadership</i>	10,79	4,22	17,81	-0,23	0,48	2	26
<i>Activity</i>	9,50	2,23	4,96	-0,03	-0,04	1	16
<i>Social Nature</i>	18,69	3,84	14,77	-0,45	-0,10	9	27
<i>Workstyle</i>	14,70	3,05	9,30	-0,34	-0,27	6	21
<i>Temperament</i>	12,31	2,77	7,70	-0,10	0,008	3	22
<i>Followership</i>	9,69	2,60	6,76	0,005	-0,31	1	17

Berdasarkan Tabel 3, skor rata-rata tertinggi terdapat pada dimensi *Social Nature* (18,69), diikuti oleh dimensi *Workstyle* (14,70). Rentang respons paling bervariasi terdapat pada dimensi *Social Nature*, dengan nilai varians sebesar 14,77. Hal ini menunjukkan bahwa keberagaman skor pada dimensi tersebut cukup tinggi. Secara keseluruhan, data tergolong homogen karena standar deviasi setiap dimensi lebih kecil daripada nilai rata-ratanya.

Revisi Item Berdasarkan Telaah Bahasa

Proses revisi item dilakukan berdasarkan hasil telaah bahasa dan konteks lokal. Setiap item ditampilkan dalam tiga bentuk, yaitu item asli PAPI Kostick, item versi adaptasi Bahasa Indonesia yang umum digunakan, dan item versi revisi. Contoh ini menunjukkan perubahan redaksi yang bertujuan meningkatkan validitas serta memperjelas bahasa dan relevansinya. Tabel berikut menyajikan perbandingan ketiga versi tersebut.

Tabel 4
Telaah Bahasa pada item PAPI Kostick

Aspek	Item	Redaksi
Need to Achieve (A)	Item Asli	"I like to set myself high standards"
	Item Versi Lama	"Biasanya saya suka bekerja keras"
	Item Versi Revisi	"Saya suka menetapkan standar yang tinggi untuk diri sendiri"
Need for Closeness (O)	Item Asli	"I like to make just one or two close friends"
	Item Versi Lama	"Saya senang bersahabat intim dengan seseorang"
	Item Versi Revisi	"Saya senang berteman dekat dengan satu atau dua teman saja"

<i>Need to be Noticed (X)</i>	Item Asli	<i>"I like to be in the spotlight"</i>
	Item Versi Lama	"Saya suka berkelakar"
	Item Versi Revisi	"Saya merasa senang ketika diperhatikan"
<i>Need for Change (Z)</i>	Item Asli	<i>"I tend to be restless"</i>
	Item Versi Lama	"Saya suka melakukan hal-hal yang baru dan berbeda"
	Item Versi Revisi	"Saya cenderung tidak bisa tenang"

Berdasarkan Tabel 4, sejumlah item perlu disesuaikan karena terjemahan awalnya tidak sepenuhnya selaras dengan makna asli maupun konteks budaya Indonesia. Pada aspek *need for closeness* (O), kata "intim" dalam Bahasa Indonesia memiliki konotasi berbeda sehingga perlu diganti dengan redaksi yang lebih tepat. Pada aspek *need to be noticed* (X), terjemahan awal justru menggambarkan perilaku suka melucu, padahal item asli menekankan preferensi terhadap perhatian publik. Sementara itu, pada aspek *need for change* (Z), versi adaptasi mengandung dua makna sekaligus yang berpotensi membingungkan responden. Penyesuaian dilakukan agar setiap item lebih akurat merepresentasikan konstruk dan tetap relevan secara budaya.

Analisis terhadap Versi Revisi

CFA

Untuk mengevaluasi validitas konstrukt, langkah pertama adalah menguji kecocokan model (*Goodness of Fit*) setiap aspek menggunakan indikator RMSEA, CFI, GFI, dan TLI, yang hasilnya ditampilkan pada tabel berikut.

Tabel 5
Uji Kecocokan Model

Dimensi	Aspek	RMSEA < 0,08	GFI > 0,9	CFI > 0,8	TLI > 0,8
Work Direction	N	0,032	0,992	0,913	0,835
	G	0,036	0,989	0,918	0,876
	A	0,03	0,991	0,889	0,834
Leadership	L	0,044	0,984	0,955	0,94
	P	0,026	0,993	0,927	0,875

	I	0,059	0,979	0,882	0,808
Activity	T	0,044	0,986	0,885	0,827
	V	0,027	0,993	0,933	0,886
Social Nature	O	0,032	0,991	0,885	0,811
	B	0,04	0,987	0,889	0,84
	S	0,063	0,976	0,876	0,821
	X	0,025	0,996	0,956	0,886
Workstyle	C	0,028	0,995	0,943	0,863
	D	0,042	0,984	0,875	0,833
	R	0,014	0,996	0,991	0,982
Temperament	Z	0,035	0,991	0,919	0,854
	E	0,046	0,986	0,897	0,838
	K	0,026	0,993	0,929	0,878
Followership	F	0,033	0,992	0,946	0,903
	W	0,057	0,975	0,862	0,816

Berdasarkan Tabel 5, seluruh nilai RMSEA, GFI, CFI, dan TLI berada dalam batas yang memenuhi kriteria, sehingga model dinyatakan *good fit*. Setelah itu, validitas setiap item diuji melalui CFA, dengan hasil pada tabel berikut.

Tabel 6
Uji Validitas CFA Sesudah Revisi

Dimensi	Valid	Tidak Valid
Work Direction	26 (96%)	1 (4%)
Leadership	26 (96%)	1 (4%)
Activity	18 (100%)	0 (0%)
Social Nature	33 (92%)	3 (8%)
Workstyle	24 (89%)	3 (11%)
Temperament	26 (96%)	1 (4%)
Followership	15 (83%)	3 (17%)

Berdasarkan Tabel 6, sebagian besar item pada setiap dimensi memenuhi kriteria validitas, dengan proporsi item valid yang tinggi. Dimensi *Activity* menunjukkan validitas sempurna (100%), sedangkan *Work Direction*, *Leadership*, dan *Temperament* masing-masing mencapai 96%. Namun, beberapa item masih tidak valid, terutama pada *Workstyle* (11%) dan *Followership*

(17%). Temuan ini menunjukkan bahwa meskipun kualitas instrumen meningkat setelah revisi, item yang belum valid tetap perlu ditinjau ulang, khususnya terkait kejelasan redaksi dan kesesuaian konteks agar representasi konstruk lebih konsisten.

IRT

Setelah analisis validitas konstruk melalui CFA, kualitas butir dievaluasi lebih lanjut menggunakan *Item Response Theory* (IRT). Analisis ini menilai kemampuan item dalam membedakan responden serta menentukan tingkat kesulitannya. Model yang digunakan adalah 2PL, yang memeriksa daya beda (*discrimination*) dan tingkat kesulitan (*difficulty*). Hasil parameter untuk setiap dimensi yang telah direvisi disajikan pada tabel berikut.

Tabel 7
Parameter IRT 2PL Sesudah Revisi

Dimensi	Difficulty			Discriminan	
	Rendah	Rata-rata	Sulit	Normal	Tidak
Work Direction	4 (15%)	20 (74%)	3 (11%)	23 (85%)	4 (15%)
Leadership	2 (7%)	23 (85%)	2 (7%)	24 (89%)	3 (11%)
Activity	3 (17%)	14 (78%)	1 (6%)	18 (100%)	0 (0%)
Social Nature	4 (11%)	25 (69%)	7 (19%)	29 (81%)	7 (19%)
Workstyle	4 (15%)	19 (70%)	4 (15%)	19 (70%)	8 (30%)
Temperament	5 (19%)	21 (78%)	1 (4%)	21 (78%)	6 (22%)
Followership	6 (33%)	11 (61%)	1 (6%)	11 (61%)	7 (39%)

Berdasarkan Tabel 7, sebagian besar item pada setiap dimensi berada pada tingkat kesulitan sedang, menunjukkan bahwa instrumen mampu mengukur kemampuan responden secara optimal pada populasi target. Proporsi tertinggi terlihat pada *Leadership* (85%), *Temperament* (78%), dan *Work Direction* (74%). Item yang lebih mudah, seperti pada *Followership*, *Activity*, dan *Work Direction*, tetap penting untuk menjangkau responden dengan kemampuan rendah sehingga distribusi tingkat kesulitan menjadi lebih seimbang dan inklusif.

Pada aspek daya beda, sebagian besar item berada pada kategori normal, menandakan kemampuan diskriminasi yang baik. Dimensi *Activity* menunjukkan hasil terbaik (100%), disusul *Leadership* (89%), *Work Direction* (85%), *Social Nature* (81%), *Temperament* (78%), dan

Workstyle (70%). Meskipun *Followership* memiliki proporsi terendah (61%), nilainya masih memadai. Temuan ini menunjukkan bahwa revisi instrumen berhasil meningkatkan kualitas butir, baik dari sisi representasi konstruk maupun kemampuan mendeteksi variasi antar individu.

Pembahasan

Hasil analisis deskriptif memberikan gambaran awal mengenai karakteristik responden dalam penelitian ini. Nilai rata-rata tertinggi ditemukan pada dimensi Social Nature ($M = 18,69$), yang mengindikasikan bahwa secara umum partisipan memiliki kecenderungan kuat dalam aspek sosial, seperti kebutuhan untuk berinteraksi, bekerja sama, dan membangun relasi dengan orang lain. Varians yang juga tinggi pada dimensi ini (14,77) menunjukkan adanya keragaman skor yang relatif besar antar individu, sehingga dapat disimpulkan bahwa meskipun mayoritas responden memiliki orientasi sosial yang tinggi, tingkat intensitasnya berbeda-beda. Hasil penelitian ini sejalan dengan temuan sebelumnya yang menunjukkan bahwa gaya komunikasi masyarakat di Indonesia cenderung memadukan ciri *High Context Culture* (HCC) dan *Low Context Culture* (LCC). Artinya, meskipun mayoritas responden memiliki orientasi sosial yang tinggi, intensitas dan bentuk ekspresinya dapat berbeda sesuai situasi (Ibrahim, 2017).

Dimensi lain yang menonjol adalah *Workstyle* ($M = 14,70$), dengan standar deviasi sebesar 3,05. Temuan ini menegaskan bahwa gaya kerja partisipan, termasuk dalam hal perencanaan dan konsistensi kerja, cenderung relatif baik dengan variasi antar individu yang lebih homogen dibandingkan Social Nature. Sementara itu, dimensi dengan rata-rata terendah adalah *Activity* ($M = 9,50$), yang dapat diinterpretasikan sebagai indikasi bahwa aktivitas spontan maupun dorongan untuk selalu bergerak dinamis tidak terlalu dominan pada sampel penelitian ini.

Secara keseluruhan, data menunjukkan pola distribusi yang cenderung normal, sebagaimana ditunjukkan oleh nilai skewness dan kurtosis yang berada di sekitar nol. Standar deviasi setiap dimensi lebih kecil dibandingkan dengan nilai rata-ratanya, sehingga dapat disimpulkan bahwa data relatif homogen. Hal ini memberikan dasar yang baik untuk analisis

lanjutan, seperti uji validitas konstruk melalui Confirmatory Factor Analysis (CFA), karena asumsi dasar mengenai stabilitas distribusi data telah terpenuhi.

Revisi Item Berdasarkan Telaah Bahasa

Proses revisi item pada PAPI Kostick dalam penelitian ini menegaskan pentingnya penyesuaian redaksi agar sesuai dengan konteks budaya dan bahasa lokal, tanpa mengubah esensi konstruk yang diukur. Hasil telaah menunjukkan bahwa beberapa item dalam versi adaptasi lama mengandung pergeseran makna yang cukup signifikan dibandingkan dengan item asli, sehingga berpotensi menimbulkan ambiguitas interpretasi. Misalnya, pada aspek Need to Achieve (A), terjemahan lama “Biasanya saya suka bekerja keras” lebih menekankan pada etos kerja secara umum, sedangkan item asli “I like to set myself high standards” sesungguhnya menekankan pada penetapan standar pribadi yang tinggi. Perbedaan ini penting karena dapat mengubah konstruk yang sesungguhnya ingin diukur, sehingga dilakukan revisi menjadi “Saya suka menetapkan standar yang tinggi untuk diri sendiri” agar lebih selaras dengan maksud asli.

Hal serupa terlihat pada aspek Need for Closeness (O). Penggunaan istilah “intim” dalam versi lama berpotensi disalahartikan dalam konteks budaya Indonesia karena konotasinya cenderung berbeda dari maksud aslinya. Revisi item menjadi “Saya senang berteman dekat dengan satu atau dua teman saja” membuat redaksi lebih natural dalam bahasa Indonesia sekaligus tetap mempertahankan makna konstruk. Dengan demikian, revisi ini tidak hanya memperjelas pemahaman responden, tetapi juga memperkuat validitas isi.

Pada aspek Need to be Noticed (X), terjemahan lama “Saya suka berkelakar” menyimpang jauh dari makna asli “I like to be in the spotlight”. Perbedaan tersebut dapat menggeser fokus konstruk dari kecenderungan mencari perhatian ke perilaku melucu. Oleh karena itu, revisi menjadi “Saya merasa senang ketika diperhatikan” dilakukan agar tetap menangkap esensi kecenderungan individu untuk mencari sorotan publik sesuai konteks budaya lokal.

Contoh terakhir adalah aspek Need for Change (Z), di mana versi lama mengandung makna ganda yang bisa ditafsirkan sebagai kecenderungan suka hal baru maupun sifat tidak tenang. Kondisi ini berpotensi menimbulkan bias respon. Revisi menjadi “Saya cenderung tidak bisa

tenang” membuat item lebih spesifik dan konsisten dengan makna asli, sehingga dapat meningkatkan kejelasan dan kemampuan item dalam merepresentasikan konstruk.

Temuan ini semakin menegaskan pentingnya proses revisi item untuk memastikan kesesuaian makna dengan konstruk yang diukur. Item dalam versi lama yang mengalami pergeseran arti atau mengandung ambiguitas berpotensi menurunkan validitas konstruk karena responden dapat menafsirkan isi pernyataan dengan cara yang berbeda dari maksud aslinya. Melalui revisi redaksi, setiap item diupayakan lebih jelas, relevan secara budaya, dan tetap mempertahankan makna asli yang ingin diukur. Dengan demikian, revisi ini diharapkan mampu meningkatkan keakuratan pengukuran serta memastikan bahwa perbedaan skor responden benar-benar mencerminkan variasi psikologis, bukan bias bahasa atau konteks. Hasil ini sejalan dengan yang direkomendasikan pada penelitian sebelumnya mengenai pentingnya ketepatan redaksi item untuk menjaga validitas konstruk dalam mengkonstruksikan suatu alat ukur (Clark & Watson, 2019; Cook & Beckman, 2006). Item yang ambigu atau mengalami pergeseran makna perlu direvisi, terutama saat instrumen diadaptasi lintas waktu atau populasi, karena jika tidak, dapat melemahkan validitas alat ukur (Revelle & Garner, 2023; Wulff & Mata, 2025).

Validitas Konstruk PAPI Kostick

Analisis CFA digunakan dalam penelitian ini untuk menguji validitas konstruk instrumen. Secara khusus, CFA berfungsi menilai sejauh mana item-item yang dikembangkan memang merefleksikan dimensi teoretis yang mendasarinya. Validitas konstruk ini mencakup dua aspek utama. Pertama, validitas konvergen, yang ditunjukkan oleh besarnya loading faktor tiap butir pada konstruk yang seharusnya diukur. Semakin tinggi nilai loading, semakin kuat butir tersebut dalam merepresentasikan konstruk. Kedua, validitas diskriminan, yang ditunjukkan oleh rendahnya korelasi antar konstruk yang berbeda, sehingga setiap dimensi dapat dikatakan unik dan terpisah. Dengan demikian, CFA bukan hanya menguji kecocokan data empiris terhadap model teoretis yang diajukan, tetapi juga memastikan bahwa instrumen mampu mengukur konstruk sesuai tujuan konseptualnya (Rogers, 2024).

Hasil uji kecocokan model (Goodness of Fit) menunjukkan bahwa seluruh aspek dari masing-masing dimensi instrumen memenuhi kriteria cut off yang telah ditetapkan. Nilai RMSEA seluruhnya berada di bawah 0,08, sedangkan nilai GFI, CFI, dan TLI berada pada rentang

yang mendukung kecocokan model. Hal ini mengindikasikan bahwa struktur faktor yang diasumsikan dalam instrumen konsisten dengan data empiris, sehingga model dapat dinyatakan memiliki tingkat kecocokan yang baik (good fit) (Baumgartner & Homburg, 1996). Temuan ini memberikan dasar yang kuat untuk melanjutkan analisis validitas item pada tingkat yang lebih rinci.

Selain itu, mayoritas item dalam instrumen telah memenuhi kriteria validitas. Dimensi Activity menunjukkan hasil paling konsisten dengan 100% item valid, sedangkan Work Direction, Leadership, dan Temperament juga memperlihatkan tingkat validitas tinggi dengan 96% item valid. Sementara itu, dimensi Social Nature dan Workstyle meskipun menunjukkan validitas yang cukup baik, masih menyisakan sejumlah kecil item yang tidak valid, masing-masing 8% dan 11%. Dimensi Followership menunjukkan proporsi item tidak valid terbesar, yakni 17%, yang mengindikasikan perlunya telaah lebih mendalam terhadap redaksi maupun relevansi item dalam konteks konstruk yang diukur.

Secara umum, hasil ini menunjukkan bahwa proses revisi item yang dilakukan berhasil meningkatkan kualitas instrumen, sebagaimana ditunjukkan oleh tingginya persentase item valid pada sebagian besar dimensi. Namun, temuan adanya item yang masih tidak valid pada beberapa dimensi menegaskan bahwa proses adaptasi instrumen tidak berhenti pada satu tahap revisi saja (Clark & Watson, 2019; Duran & Celik, 2025; Kyriazos & Stalikas, 2018). Peninjauan lanjutan diperlukan untuk memastikan setiap item memiliki kejelasan makna, relevansi budaya, dan konsistensi dengan model teoretis. Dengan demikian, diharapkan instrumen dapat memberikan representasi konstruk yang lebih kuat dan stabil pada pengukuran berikutnya.

Evaluasi Kualitas Item

Setelah uji validitas konstruk melalui CFA menunjukkan model yang sesuai, analisis dilanjutkan dengan mengevaluasi kualitas butir menggunakan pendekatan Item Response Theory (IRT) model 2PL. Analisis ini memberikan gambaran lebih rinci mengenai bagaimana setiap item berfungsi dalam membedakan responden berdasarkan karakteristik psikologis yang diukur, serta distribusi tingkat kesulitan antar butir.

Hasil penelitian menunjukkan bahwa sebagian besar item berada pada kategori kesulitan sedang (*moderate difficulty*), dengan proporsi dominan pada dimensi Leadership (85%), Temperament (78%), dan Work Direction (74%). Pola ini mengindikasikan bahwa instrumen secara umum telah seimbang dalam mengukur responden dengan tingkat kemampuan rata-rata, yang sesuai dengan karakteristik populasi target. Di samping itu, keberadaan sejumlah item mudah, seperti pada dimensi Followership (33%) dan Activity (17%) yang tetap memberikan kontribusi penting karena memungkinkan instrumen menjangkau responden dengan kemampuan relatif rendah. Distribusi yang relatif proporsional antara item mudah, sedang, dan sulit memperkuat inklusivitas instrumen sehingga mampu menangkap variasi kemampuan secara lebih luas.

Dari sisi daya beda (*discrimination*), temuan menunjukkan bahwa mayoritas item memiliki kualitas yang memadai. Seluruh item pada dimensi Activity berada dalam kategori normal (100%), menunjukkan kemampuan diskriminatif yang sangat baik. Dimensi lain seperti Leadership (89%), Work Direction (85%), dan Social Nature (81%) juga menampilkan proporsi item dengan daya beda normal yang tinggi. Sementara itu, Workstyle (70%) dan Followership (61%) menunjukkan proporsi lebih rendah, namun tetap berada pada kisaran yang dapat diterima. Artinya, meskipun ada beberapa item yang daya bedanya masih kurang optimal, instrumen secara keseluruhan telah mampu membedakan responden dengan cukup efektif.

Hasil ini menegaskan bahwa revisi instrumen tidak hanya memperbaiki aspek bahasa dan konteks budaya, tetapi juga meningkatkan kualitas psikometris dari segi fungsionalitas butir (Cruchinho et al., 2024; Ohrbach et al., 2008). Dengan daya beda yang baik dan distribusi kesulitan yang representatif, instrumen ini dapat dikatakan telah memiliki kemampuan diagnostik yang lebih komprehensif. Temuan tersebut sekaligus memperkuat posisi PAPI Kostick versi revisi sebagai alat ukur yang lebih sesuai digunakan dalam konteks lokal, baik untuk tujuan asesmen individu maupun riset psikologis.

Secara keseluruhan, hasil analisis CFA dan IRT menunjukkan bahwa instrumen versi revisi memiliki validitas konstruk yang memadai, distribusi posisi item pada kontinum konstruk yang relatif seimbang, serta daya beda yang pada umumnya berada dalam kategori baik. Kondisi ini mengindikasikan bahwa instrumen telah berfungsi sesuai dengan tujuan pengukuran dan mampu menangkap variasi karakteristik psikologis secara akurat. Meskipun

demikian, keberadaan sejumlah item dengan kualitas yang masih belum optimal menegaskan perlunya evaluasi dan pengembangan berkelanjutan. Dengan demikian, hasil penelitian ini menegaskan kontribusi proses revisi terhadap peningkatan kualitas instrumen, sekaligus membuka ruang bagi penelitian lanjutan pada konteks dan populasi yang lebih beragam.

Kesimpulan

Penelitian ini menegaskan bahwa proses revisi item pada PAPI Kostick berperan penting dalam memastikan kesesuaian makna dengan konstruk yang diukur, sekaligus meningkatkan kejelasan bahasa dan relevansi budaya. Hasil analisis CFA menunjukkan bahwa struktur faktor instrumen konsisten dengan data empiris, dengan mayoritas dimensi memiliki proporsi item valid yang tinggi. Evaluasi lebih lanjut melalui IRT juga mengonfirmasi bahwa sebagian besar item memiliki daya beda yang memadai serta distribusi posisi yang seimbang pada kontinum konstruk. Hasil penelitian ini juga memberikan beberapa implikasi yang relevan bagi penggunaan instrumen PAPI Kostick dalam konteks asesmen psikologis di Indonesia. Pertama, revisi item yang dilakukan telah menghasilkan butir yang lebih sesuai dengan bahasa dan konteks budaya lokal, sehingga dapat meminimalkan bias interpretasi responden. Hal ini meningkatkan keandalan asesmen dalam pengambilan keputusan di bidang organisasi maupun pendidikan. Kedua, validitas konstruk yang terkonfirmasi melalui CFA dan kualitas butir yang teruji melalui IRT memperkuat keyakinan bahwa instrumen ini dapat digunakan sebagai alat ukur yang cukup akurat untuk memetakan profil psikologis individu. Temuan ini penting bagi praktisi SDM, konselor karier, maupun psikolog industri dan organisasi yang membutuhkan instrumen dengan keandalan untuk mendukung proses seleksi, pengembangan, dan penempatan karyawan. Ketiga, distribusi item yang seimbang pada kontinum konstruk memungkinkan instrumen menjangkau responden dengan tingkat karakteristik yang beragam, sehingga instrumen lebih inklusif dalam menggambarkan variasi psikologis antar individu.

Saran

Berdasarkan hasil penelitian, beberapa hal dapat direkomendasikan yaitu item yang masih menunjukkan kelemahan, khususnya pada dimensi Workstyle dan Followership, perlu ditinjau kembali untuk memastikan kejelasan dan konsistensi konstruk. Penelitian lanjutan

sebaiknya melibatkan sampel yang lebih beragam agar stabilitas validitas dan reliabilitas instrumen semakin teruji. Kemudian yang terakhir, penggunaan instrumen versi revisi dapat dipertimbangkan dalam berbagai konteks asesmen psikologis, dengan tetap diintegrasikan bersama metode lain guna memperoleh gambaran yang lebih komprehensif.

Referensi

- Akkus, A. (2019). Developing a Scale to Measure Students' Attitudes toward Science. *International Journal of Assessment Tools in Education*, 6(4), 706–720. <https://doi.org/10.21449/ijate.548516>
- Baker, F. B. (2001). *The Basics Of Item Response Theory second edition*. ERIC Clearinghouse on Assessment and Evaluation.
- Baumgartner, H., & Homburg, C. (1996). Applications of structural equation modeling in marketing and consumer research: A review. *International Journal of Research in Marketing*, 139–161. <https://doi.org/10.1016/j.ecolecon.2015.08.017>
- Chernyshenko, O. S., Stark, S., Chan, K. Y., Drasgow, F., & Williams, B. (2001). Fitting item response theory models to two personality inventories: Issues and insights. *Multivariate Behavioral Research*, 36(4), 523–562. https://doi.org/10.1207/S15327906MBR3604_03
- Clark, L. A., & Watson, D. (2019). Constructing Validity: New Developments in Creating Objective Measuring Instruments. *Psychol Assess*. <https://doi.org/10.1037/pas0000626.Constructing>
- Cook, D. A., & Beckman, T. J. (2006). Current concepts in validity and reliability for psychometric instruments: Theory and application. *American Journal of Medicine*, 119(2), 166.e7-166.e16. <https://doi.org/10.1016/j.amjmed.2005.10.036>
- Cruchinho, P., López-Franco, M. D., Capelas, M. L., Almeida, S., Bennett, P. M., da Silva, M. M., Teixeira, G., Nunes, E., Lucas, P., & Gaspar, F. (2024). Translation, Cross-Cultural Adaptation, and Validation of Measurement Instruments: A Practical Guideline for Novice Researchers. *Journal of Multidisciplinary Healthcare*, 17(March 2024), 2701–2728. <https://doi.org/10.2147/JMDH.S419714>
- Dalal, D. K., Zhu, X. (Susan), Rangel, B., Boyce, A. S., & Lobene, E. (2021). Improving Applicant Reactions to Forced-Choice Personality Measurement: Interventions to Reduce Threats to Test Takers' Self-Concepts. *Journal of Business and Psychology*. <https://doi.org/10.1007/s10869-019-09655-6>
- Duran, V., & Celik, F. (2025). The 8-Factor reasoning styles scale: development, validation, and psychometric evaluation. *BMC Psychology*, 13, 939. <https://doi.org/https://doi.org/10.1186/s40359-025-03320-9>
- Forman, M., Haugbolle, K., Vogelius, P., & Storgaard, K. (2009). Employees In The Dual Role of Employee And End-User-A New Source of Innovation For Companies? *Proceeding of*

- 5th Nordic Conference on Construction Economics and Organisation, 2(June), 3–220.
- Foster, G. C., Min, H., & Zickar, M. J. (2017). Review of Item Response Theory Practices in Organizational Research: Lessons Learned and Paths Forward. *Organizational Research Methods*, 20(3), 465–486. <https://doi.org/10.1177/1094428116689708>
- Furnham, A., & Craig, S. (1987). Fakeness and correlates of the perception and preference inventory. *Personality and Individual Differences*, 8(4), 459–470. [https://doi.org/10.1016/0191-8869\(87\)90207-8](https://doi.org/10.1016/0191-8869(87)90207-8)
- Ibrahim, I. (2017). Konsep Diri dalam Komunikasi. *Al-Hikmah*. <https://doi.org/10.24260/al-hikmah.v11i2.851>
- Jiandy, R., Wiratama, Y. W., & Hansun, S. (2017). Implementation of C4.5 and PAPI Kostick to predict students potential as organization caretaker. *International Journal of Information Technology, Computer Science and Open Source*, 1(1), 1–8.
- Joubert, T., Inceoglu, I., Bartram, D., Dowdeswell, K., & Lin, Y. (2015). A comparison of the psychometric properties of the forced choice and likert scale versions of a personality instrument. *International Journal of Selection and Assessment*, 23(1), 92–97. <https://doi.org/10.1111/ijsa.12098>
- Kuric, E., Demcak, P., Smrecek, P., & Spilakova, B. (2025). User modeling for detecting faking-good intent in online personality questionnaires in the wild based on mouse dynamics. *Multimedia Tools and Applications*. <https://doi.org/10.1007/s11042-025-20852-9>
- Kyriazos, T. A., & Stalikas, A. (2018). Applied Psychometrics: The Steps of Scale Development and Standardization Process. *Psychology*, 09(11), 2531–2560. <https://doi.org/10.4236/psych.2018.911145>
- Mutiawani, V., Athaya, A. M., Saputra, K., & Subianto, M. (2022). Implementing Item Response Theory (IRT) Method in Quiz Assessment System. *TEM Journal*, 11(1), 210–218. <https://doi.org/10.18421/TEM111-26>
- Ohrbach, R., Granger, C., List, T., & Dworkin, S. (2008). Preliminary development and validation of the Jaw Functional Limitation Scale. *Community Dentistry and Oral Epidemiology*. <https://doi.org/10.1111/j.1600-0528.2007.00397.x>
- Ondé, D., & Alvarado, J. M. (2020). Reconsidering the Conditions for Conducting Confirmatory Factor Analysis. *Spanish Journal of Psychology*, May, 1–15. <https://doi.org/10.1017/SJP.2020.56>
- Putri, C. A., Rofiqoh, E., Wulandari, F. A., & Prastiningrum, F. A. (2022). Implementasi Alat Tes Papi Kostick (Review Sistematis). *Flourishing Journal*, 2(2), 121–129. <https://doi.org/10.17977/um070v2i22022p121-129>
- Revelle, W., & Garner, K. M. (2023). Measurement: Reliability, construct validation, and scale construction. In *Handbook of research methods in social and personality psychology*. <https://www.cambridge.org.libproxy.nie.edu.sg/core/publications/books>
- Rogers, P. (2024). Best practices for your confirmatory factor analysis: A JASP and lavaan tutorial. *Behavior Research Methods*, 56(7), 6634–6654. [https://doi.org/10.3758/s13428-024-](https://doi.org/10.3758/s13428-024-024)

02375-7

- S, S., & Mohanasundaram, T. (2024). Fit Indices in Structural Equation Modeling and Confirmatory Factor Analysis: Reporting Guidelines. *Asian Journal of Economics, Business and Accounting*, 24(7), 561–577. <https://doi.org/10.9734/ajeba/2024/v24i71430>
- Swanepoel, I., & Krüger, C. (2011). Revisiting validity in cross-cultural psychometric test development: A systems-informed shift towards qualitative research designs. *South African Journal of Psychiatry*, 17(1), 10–15. <https://doi.org/10.4102/sajpsychiatry.v17i1.250>
- Tang, S. (2023). *The Psychiatrist: Clinical and The Power and Significance of Psychological Assessment*. 5(5). <https://doi.org/10.4172/tpctj.1000217>
- Wulff, D. U., & Mata, R. (2025). Semantic embeddings reveal and address taxonomic incommensurability in psychological measurement. *Nature Human Behaviour*, 9(5), 944–954. <https://doi.org/10.1038/s41562-024-02089-y>