Jurnal Penelitian dan Pengukuran Psikologi
Volume 13, Nomor 02, Oktober 2024 http://doi.org/10.21009/JPPP.132.06

ANALISIS BUTIR *SITUASIONAL JUDGEMENT TEST* KOMPETENSI KEPEMIMPINAN KARYAWAN BUMN DENGAN *RASCH MODEL*

Ni Ketut Laksmi Kusuma[1] & Mohammad Abdul Hakim[2]
[1,2] Fakultas Psikologi, Universitas Sebelas Maret

***Email**: laksmikusumaa@student.uns.ac.id*

Abstract

Employee competency assessment plays a crucial role in organizations, particularly in PT X, a state-owned enterprise in Indonesia's transportation sector. PT X developed a leadership competency test using a Situational Judgement Test (SJT), which is expected to objectively measure competencies through realistic work scenarios. The present study involved 2,368 PT X employees from job level 1. Item analysis was conducted on 48 items measuring 7 leadership competencies at competency level 1, aiming to enhance test quality through improvement or elimination of unsuitable items. Results showed that 15 items were answered correctly by >50% of respondents, while 33 items were answered correctly by <50% of respondents. Although most items demonstrated good difficulty levels ($-2 \geq b \geq +2$), some items such as DLE 1.3.3, DLE 1.2.2, and DEX 1.2.2 exhibited suboptimal difficulty levels. The most accurate measurements were found in several items, including SOR 1.2.1, SOR 1.3.1, SOR 1.1.1, and SOR 1.1.2, while DEX 1.2.2 showed less accurate measurement. Fit evaluation (Infit & Outfit) across all items indicated appropriate values (0.5 – 1.5) for the measured competencies, confirming test reliability. The Wright Map revealed that the DEX competency could measure the full range of abilities; the DLE competency captured average to high abilities; Strategic Orientation (SOR), Developing Organizational Capabilities (DOC), and Leading Change (LCH) competencies captured average abilities; while Global Business Savvy (GBS) and Managing Diversity (MDI) competencies captured average and below-average abilities. This study concludes that the SJT demonstrates good item quality and can be reliably used for employee competency assessment at PT X.

Keywords: : Item Analysis, Situational Judgement Test, Leadership Competency, Competency Assessment, Assessment BUMN Employee

Abstrak

Asesmen kompetensi SDM memiliki peran krusial dalam organisasi, khusunya pada PT X, sebuah BUMN di sektor transportasi Indonesia. PT X mengembangkan tes kompetensi kepemimpinan melalui *Situational Judgement Test* (SJT), diharapkan dapat secara objektif mengukur kompetensi dengan skenario pekerjaan realistis. Penelitian ini melibatkan 2.368 karyawan PT X dari kelompok jabatan level 1. Analisis butir dilakukan pada 48 aitem yang mengukur 7 kompetensi kepemimpinan pada level kompetensi 1, dengan tujuan meningkatkan mutu tes melalui perbaikan atau penghapusan butir yang tidak sesuai. Ditemukan bahwa sebanyak 15 aitem dijawab >50% responden, sedangkan 33 aitem dijawab benar oleh <50% responden. Meskipun sebagian besar aitem memiliki tingkat kesukaran yang baik ($-2 \geq b \geq +2$), beberapa aitem seperti DLE 1.3.3, DLE 1.2.2, dan DEX 1.2.2 yang memiliki tingkat kesukaran kurang baik. Pengukuran paling akurat ditemukan pada beberapa aitem, seperti SOR 1.2.1, SOR 1.3.1, SOR 1.1.1, dan SOR 1.1.2, sementara DEX 1.2.2 menunjukkan pengukuran yang kurang akurat. Evaluasi kecocokan (*Infit & Outfit*) pada seluruh aitem menunjukkan nilai yang sesuai (0,5 – 1,5) dengan kompetensi yang diukur, menegaskan keandalan tes. *Wright Map* menunjukkan kompetensi DEX mampu mengukur keseluruhan abilitas; kompetensi DLE mampu memotret abilitas responden rata-rata hingga tinggi; kompetensi *Strategic Orientation* (SOR), *Developing Organizational Capabilities* (DOC), dan *Leading Change* (LCH) memotret abilitas rata-rata; kompetensi *Global Business Savvy* (GBS) dan *Managing Diversity* (MDI) memotret abilitas pada tingkat rata-rata dan di bawah rata-rata. Penelitian ini menyimpulkan bahwa tes SJT ini memiliki kualitas butir yang baik dan dapat diandalkan untuk asesmen kompetensi PT X secara berkelanjutan.

Kata kunci: Analisis Butir, Situational Judgement Test, Kompetensi Kepemimpinan, Asesmen Kompetensi, Asesmen Karyawan BUMN

1. Pendahuluan

Asesmen kompetensi adalah salah satu cara untuk menilai kemampuan dan potensi sumber daya manusia (SDM) dalam suatu organisasi. Asesmen kompetensi berperan penting dalam membantu organisasi meraih SDM berkualitas, meningkatkan produktivitas, prestasi kerja, serta bersaing dengan efektif dalam era perkembangan teknologi dan bisnis yang dinamis (Labola, 2019). Salah satu contoh penerapan asesmen kompetensi di Indonesia adalah PT X, sebuah perusahaan Badan Usaha Milik Negara (BUMN) yang bergerak di bidang transportasi. PT X telah menjalankan asesmen kompetensi untuk karyawan dan calon anggota direksi sesuai dengan Permen BUMN No Per 04/MBU/10/2019 tentang Kamus Kompetensi ASN di Lingkungan Kementerian BUMN, serta Permen BUMN No 11/MBU/07/2021 tentang Persyaratan, Tata Cara Pengangkatan, dan Pemberhentian Calon Anggota Direksi BUMN (Kementerian BUMN, 2019; 2021). Pada umumnya, uji kompetensi dilakukan menggunakan pendekatan *Assessment Center* (AC). Akan tetapi, pelaksanaan AC membutuhkan sumber daya yang besar, mencakup waktu, sarana dan prasarana, melibatkan banyak assessor dan biaya besar, serta menuntut pegawai meluangkan waktu dan mengikuti proses yang kompleks.

Alternatif dari pendekatan AC adalah sebuah tes berbentuk simulasi dengan metode *Situational Judgement Test* (SJT). SJT merupakan sebuah tes terstandar yang mengukur kompetensi pegawai berdasarkan serangkaian situasi dan kasus yang mensimulasikan sebuah pekerjaan, dan meminta *testee* memilih tindakan yang paling optimal. Studi meta analisis menunjukkan validitas operasional 0,26 untuk SJT dan perkiraan validitas rata-rata di berbagai meta-analisis 0,29 untuk AC (Sackett et al., 2021). Uji validitas prediktif pada calon dokter (N = 196) menunjukkan SJT mampu memprediksi kinerja secara keseluruhan (r = 0.37), dan *job knowledge* (r = 0.36) (Lievens & Pattersons, 2011). Penelitian yang sama juga menunjukkan korelasi antara SJT dan AC adalah 0.47. Berbagai temuan ini mengindikasikan bahwa SJT dapat menjadi alternatif metode AC dalam mengukur kompetensi pegawai, sehingga tes kompetensi dapat dilakukan secara lebih efisien.

Sekarang ini berbagai perusahaan, khususnya di lingkungan BUMN seperti PT X, mulai mengembangkan tes kompetensi menggunakan metode *Situational Judgement Test* (SJT). Metode SJT ini diharapkan dapat mengukur kompetensi karyawan secara objektif karena berdasar pada skenario pekerjaan yang realistis dan kontekstual, sehingga dapat menggambarkan situasi yang sering dihadapi oleh karyawan ketika bekerja (Lievens & Patterson, 2011). PT X, Perusahaan BUMN bidang transportasi, telah mengembangkan Tes Kompetensi Kepemimpinan berdasarkan kamus kompetensi bekerjasama dengan Fakultas Psikologi Universitas Sebelas Maret. Pengembangan alat tes kompetensi ini telah melalui uji validitas isi oleh ahli psikometri dan assessor SDM secara independen. Untuk menjamin properti psikometri yang optimal pada alat tes kompetensi, diperlukan analisis lebih lanjut guna mengevaluasi kualitas butir dan kemampuan tes dalam mengukur kompetensi yang telah ditentukan (Fitrianawati, 2017).

Analisis butir merupakan evaluasi yang bertujuan meningkatkan kualitas pengukuran kompetensi dengan memberikan hasil yang valid dan reliabel (Arikunto, 2008). Metode ini berperan dalam mengoptimalkan keandalan skor dan mengurangi jumlah butir yang tidak berfungsi dengan baik dalam pengembangan tes (Courville, 2004). Dengan demikian, melalui analisis butir tidak hanya berguna untuk meningkatkan mutu tes, tetapi juga memberikan informasi diagnostik terkait kemampuan karyawan yang sudah maupun belum memiliki kompetensi yang diukur, akhirnya mendukung perbaikan atau penghapusan butir yang tidak sesuai untuk memperbaiki instrumen tes secara keseluruhan.

Dalam melakukan analisis butir, salah satu pendekatan yang dapat digunakan adalah *Item Response Theory* (IRT). IRT memiliki berbagai variasi parameter logistik (disebut PL), salah satunya adalah 1PL yang dikembangkan oleh Georg Rasch yakni *Rasch Model* (Olsen, 2003). Pemodelan Rasch bertujuan untuk mengembangkan ukuran objektif dan telah memenuhi syarat pengukuran objektif (Krabbe, 2017). *Rasch Model* memiliki akurasi yang tinggi, sehingga mendorong kemungkinan adanya bias yang terlihat (Ramdhani et al., 2018). Dalam penelitian ini, Peneliti akan menggunakan *Rasch Model* untuk mengungkap kualitas butir soal melalui analisis butir SJT pada kompetensi kepemimpinan karyawan BUMN.

Penelitian ini bertujuan untuk menguji kualitas butir soal SJT pada kompetensi kepemimpinan yang dikembangkan untuk karyawan BUMN PT X. Secara khusus, penelitian ini menganalisis pola jawaban terhadap item sesuai dengan tingkat kesulitan (kelompok kompetensi tinggi dan kelompok kompetensi rendah). Penelitian terdahulu menyarankan penggunaan *Rasch Model* untuk menganalisis butir *Situational Judgement Test* (Musid et al, 2023). Kemudian, penelitian lain juga menunjukkan bahwa *Rasch Model* dapat digunakan dalam analisis butir tes (Fernanda & Hidayah, 2020; Azizah & Wahyuningsih, 2020).

Secara teoritis, penelitian ini memberikan kebaruan (*novelty*) karena belum adanya analisis butir soal SJT yang mengukur kompetensi kepemimpinan karyawan, terutama yang secara khusus ditujukan pada karyawan BUMN menggunakan *Rasch Model*. Oleh karena itu, penelitian ini bertujuan untuk melakukan analisis butir SJT terhadap instrumen pengukuran kompetensi kepemimpinan karyawan BUMN dengan *Rasch Model*.

2. Metode Penelitian

Partisipan

Penelitian ini melibatkan karyawan PT X dengan kelompok jabatan level 1 sebagai populasi. Pendekatan *non probability* digunakan dengan metode *total population sampling,* yaitu, melibatkan semua individu dalam populasi yang memiliki karakteristik tertentu. Dalam penelitian ini, karakteristik tersebut adalah karyawan PT X yang masuk dalam kelompok jabatan level 1 wajib mengikuti asesmen kompetensi 2023. Jumlah sampel yang digunakan sebanyak 2.368 peserta. Jumlah tersebut didasarkan pada pertimbangan jumlah peserta asesmen, semakin besar ukuran sampel, maka hasil stabilitas pengukuran akan lebih baik dan akurat (Sumintono, 2013).

Pengumpulan Data

Penelitian ini menggunakan alat ukur SJT untuk mengukur 7 kompetensi kepemimpinan pada karyawan BUMN yang dikembangkan oleh PT X. Alat ukur tersebut terdiri dari 48 item mengukur 7 kompetensi kepemimpinan pada level kompetensi 1. Proses pengambilan data dilaksanakan oleh PT X melalui pelaksanaan asesmen kompetensi pada sistem CBT yang dikembangkan oleh PT X. Pengumpulan data dilakukan di kantor pusat dan kantor cabang yang tersebar pada 16 lokasi di Indonesia

Analisis Data

Teknik analisis data dalam penelitian ini menggunakan metode *Rasch Model* untuk menganalisis setiap butir kompetensi. Analisis data menggunakan aplikasi Jamovi 2.3.28 dengan modul *SnowRMM* akan menghasilkan skor *Item Mean*, *Measure*, *Standard Error* (SE) *Measure*, nilai *Outfit & Infit,* dan persebaran *Wright Map*. Skor *Item Mean* adalah nilai rata-rata butir. Skor *Measure* mengukur tingkat kesukaran item/tes (Seol, 2020). *SE Measure* menandakan seberapa stabil atau tidak stabil perkiraan kemampuan individu yang diukur dalam satuan *logit* (Katz, 2021). Nilai *Outfit* mengukur ketepatan butir antara respons aktual individu dan prediksi *Rasch Model*. Sedangkan, nilai *Infit* menunjukkan kesesuaian antara data dan model pada tingkat keterampilan individu (Linacre, 2002). Salah satu keistimewaan *Rasch Model* adalah menghasilkan peta yang menggambarkan sebaran kemampuan peserta dan sebaran tingkat kesukaran soal dengan skala yang sama yang disebut *Wright Map* (juga disebut *person-item-maps*) (Sumintono, 2017).

3. Hasil dan Diskusi

Dalam analisis *Rasch Model*, dua syarat utama yang harus dipenuhi untuk memastikan validitas dan reliabilitas pengukuran adalah unidimensi dan independensi lokal. *Rasch Model* mengasumsikan bahwa semua item mengukur satu dimensi atau konstruk yang sama, seperti kompetensi kepemimpinan. Untuk memastikan hal ini, uji unidimensi dapat dilakukan menggunakan uji *Martin-Löf,* yang tersedia di paket uji *Rasch* di Jamovi (Linacre, 2011). Berdasarkan uji unidimensionalitas dengan *Martin-Löf* (Tabel 1) dapat disimpulkan bahwa DLE, DEX, DOC, dan LCH menunjukkan unidimensionalitas yang baik ($p > 0.05$), sementara item-item pada GBS, SOR, dan MDI menunjukkan kecenderungan multidimensionalitas ($p < 0.05$). Tetapi di sisi lain, butir-butir SJT kompetensi kepemimpinan disusun berdasarkan indikator perilaku *(key behaviors)* per jenis kompetensi yang ditetapkan dalam kamus kompetensi BUMN PT X. Kamus kompetensi tersebut lebih menekankan pada seperangkat perilaku yang diharapkan pada pegawai pada setiap jenjang jabatan, tanpa terlalu menekankan ada atau tidaknya faktor laten tunggal yang melatarbelakanginya. Oleh karena itu, dalam analisis selanjutnya peneliti lebih fokus pada *fitness* per butir tes.

Tabel 1. Analisis *Unidimensionalitas Per Jenis Kompetensi*

Kompetensi	*Nilai Martin-Löf*	*df*	*p*
Digital Leadership (DLE)	25.6	19	0.143
Global Business Savvy (GBS)	46.9	7	< .001
Strategic Orientation (SOR)	35.8	19	0.011
Driving Execution (DEX)	11	8	0.2
Developing Organizational Capabilities (DOC)	4.93	8	0.766
Leading Change (LCH)	6.62	8	0.578
Managing Diversity (MDI)	21.2	7	0.003

Respons terhadap setiap item harus independen dari respons item lainnya setelah kemampuan responden dipertimbangkan. Korelasi antara item yang melanggar independensi lokal dapat mengurangi akurasi pengukuran (Marais & Andrich, 2008). Sementara itu, *Differential Item Functioning* (DIF) atau bias item tidak dibahas dalam artikel ini, karena membutuhkan analisis terpisah untuk mengidentifikasi kelompok yang

mungkin menciptakan bias, seperti gender, asal daerah, atau bidang kerja. Kajian mendalam terkait kelompok berpotensi bias akan diperlukan di masa mendatang.

Responden penelitian berasal dari berbagai kantor cabang yang tersebar di Pulau Kalimantan (12%), Pulau Jawa (41%), Kepulauan Nusa Tenggara dan Bali (30%), Pulau Sulawesi (10%), Kepulauan Maluku (2%), dan Pulau Papua (4%).

Data penelitian kemudian dianalisis per kompetensi dengan komponen butir yang terungkap, sebagai berikut:

Item Mean, Measure & SE Measure

Tabel 2 diurutkan sesuai dengan *item measure* (tingkat kesulitan) per dimensi dari paling sulit hingga mudah.

Tabel 2. Analisis *Item Mean, Measure & SE Measure*

Item	*Item Mean*	*Measure*	*SE Measure*	*Item*	*Item Mean*	*Measure*	*SE Measure*
Digital Leadership (DLE)				*Strategic Orientation (SOR)*			
DLE 1.3.3	0.0984	2.114	0.0751	SOR 1.2.3	0.367	0.58448	0.0485
DLE 1.2.2	0.2152	1.091	0.0557	SOR 1.2.1	0.399	0.05625	0.0469
DLE 1.1.1	0.3164	0.49	0.0498	SOR 1.2.2	0.444	0.42486	0.0478
DLE 1.2.3	0.3344	0.395	0.0492	SOR 1.1.3	0.475	0.20444	0.0471
DLE 1.3.2	0.4527	-0.192	0.047	SOR 1.3.1	0.487	0.00121	0.0469
DLE 1.1.3	0.5312	-0.566	0.047	SOR 1.1.1	0.503	-0.07797	0.0469
DLE 1.3.1	0.6185	-0.994	0.0484	SOR 1.1.2	0.514	-0.12858	0.0469
DLE 1.2.1	0.6236	-1.019	0.0485	SOR 1.3.3	0.524	-0.17485	0.047
DLE 1.1.2	0.6804	-1.319	0.0504	SOR 1.3.2	0.667	-0.88984	0.0497
Global Business Savvy (GBS)				*Driving Execution (DEX)*			
GBS 1.2.1	0.183	1.1643	0.0626	DEX 1.2.2	0.0653	2.3193	0.091
GBS 1.2.2	0.267	0.5227	0.0553	DEX 1.1.3	0.1888	0.9292	0.0589
GBS 1.2.3	0.301	0.3004	0.0536	DEX 1.2.3	0.3315	0.013	0.0499
GBS 1.1.1	0.346	0.02	0.0519	DEX 1.1.1	0.3987	-0.3477	0.0483
GBS 1.1.2	0.404	-0.3202	0.0506	DEX 1.1.2	0.5011	-0.8721	0.0477
GBS 1.1.3	0.642	-1.6873	0.0532	DEX 1.2.1	0.7128	-2.0418	0.0537
Managing Diversity (MDI)				*Leading Change (LCH)*			
MDI 1.1.2	0.121	1.374	0.0694	LCH 1.2.2	0.220	0.623	0.0549
MDI 1.1.1	0.164	0.965	0.0618	LCH 1.2.1	0.224	0.596	0.0546
MDI 1.2.3	0.213	0.591	0.0567	LCH 1.1.2	0.269	0.314	0.0517
MDI 1.1.3	0.263	0.255	0.0532	LCH 1.2.3	0.358	-0.176	0.0485
MDI 1.2.1	0.536	-1.238	0.049	LCH 1.1.3	0.438	-0.58	0.0474
MDI 1.2.2	0.661	-1.947	0.0527	LCH 1.1.1	0.477	-0.777	0.0473
Developing Organizational Capabilities (DOC)							
DOC 1.2.3	0.263	0.9264	0.0526	DOC 1.1.1	0.442	-0.0372	0.0472
DOC 1.2.2	0.316	0.6167	0.05	DOC 1.2.1	0.592	-0.7787	0.048
DOC 1.1.3	0.402	0.1631	0.0477	DOC 1.1.2	0.614	-0.8903	0.0485

Berdasarkan tabel 2, *Item Mean* menggambarkan nilai rata-rata butir. Pada alat tes ini, data yang terkumpul adalah data dikotomi, sehingga skor rata-rata butir (*item mean*) adalah 0.5. Hal ini berarti, 50% responden menjawab benar atau 50% responden menjawab salah. Jika *item mean* <0.5, maka sebagian responden menjawab salah atau artinya sebagian besar responden menjawab salah (nilai 0). Berdasarkan hasil analisis, sebanyak 15 aitem dijawab benar oleh lebih dari 50% responden, sedangkan 33 aitem dijawab benar oleh kurang dari 50% responden.

Item Measure menunjukkan tingkat kesukaran aitem (Seol, 2020). Tabel *Item Measure* adalah nilai *logit* dari tiap butir soal. Nilai *logit* tertinggi menunjukkan tingkat kesulitan yang tinggi pada soal tersebut di masing-masing dimensinya (Kurniawan, 2018). Semakin tinggi skor pada *Item Measure*, maka soal tersebut semakin sulit, begitupun sebaliknya. Sebagai contoh, aitem DLE 1.3.3 adalah aitem yang memiliki tingkat kesulitan paling tinggi di antara aitem lain pada kompetensi DLE. Selanjutnya, pada tabel per dimensi, jika semakin ke bawah adalah urutan aitem semakin mudah.

Menurut Hambleton & Swaminathan (1985), indeks tingkat kesukaran butir yaitu $-2 \geq b \geq +2$. Soal dikatakan memiliki tingkat kesukaran yang baik jika tidak kurang dari -2 dan tidak lebih dari +2. Berdasarkan hasil analisis di atas dapat disimpulkan bahwa hampir keseluruhan aitem memiliki tingkat kesukaran yang baik. Namun, DLE 1.3.3 dan DEX 1.2.2 dinilai terlalu sulit. Sedangkan DEX 1.2.1 memiliki tingkat kesukaran yang kurang baik karena dinilai terlalu mudah.

Nilai *SE Measure* mencerminkan stabilitas atau fluktuasi perkiraan kemampuan individu dalam satuan *logit*. Selain itu, *SE Measure* mengindikasikan akurasi perkiraan dari *Item Measure* yang mencerminkan tingkat kesulitan aitem. Semakin besar nilai *SE Measure*, menunjukkan bahwa perkiraan pengukuran aitem tersebut kurang akurat (Katz, 2021). Dalam analisis ini, ditemukan bahwa aitem dengan pengukuran paling akurat terdapat pada aitem SOR 1.2.1, SOR 1.3.1, SOR 1.1.1, SOR 1.1.2 yang memiliki nilai *SE Measure* sebesar 0,0469. Sebaliknya, aitem dengan *nilai SE Measure* tertinggi adalah DEX 1.2.2, menunjukkan bahwa pengukuran aitem tersebut kurang akurat.

Item Fitness

Tabel 3. Analisis *Item Fitness (Infit & Outfit)*

Item	*Infit*	*Outfit*	*Item*	*Infit*	*Outfit*
Digital Leadership (DLE)			*Strategic Orientation (SOR)*		
DLE 1.1.1	0.983	1.010	SOR 1.1.1	0.994	1.001
DLE 1.1.2	0.961	0.947	SOR 1.1.2	1.102	1.128
DLE 1.1.3	0.998	0.997	SOR 1.1.3	0.938	0.918
DLE 1.2.1	0.981	0.959	SOR 1.2.1	1.050	1.067
DLE 1.2.2	1.074	1.178	SOR 1.2.2	0.959	0.940
DLE 1.2.3	0.987	0.980	SOR 1.2.3	1.037	1.061
DLE 1.3.1	0.957	0.940	SOR 1.3.1	1.075	1.089
DLE 1.3.2	1.023	1.030	SOR 1.3.2	0.904	0.869
DLE 1.3.3	1.032	1.193	SOR 1.3.3	0.931	0.921
Global Business Savvy (GBS)			*Driving Execution (DEX)*		
GBS 1.1.1	1.082	1.110	DEX 1.1.1	0.985	0.983
GBS 1.1.2	0.979	0.952	DEX 1.1.2	0.961	0.966
GBS 1.1.3	1.092	1.249	DEX 1.1.3	0.987	1.022
GBS 1.2.1	1.158	1.268	DEX 1.2.1	1.004	0.996
GBS 1.2.2	0.824	0.787	DEX 1.2.2	1.038	1.309
GBS 1.2.3	0.854	0.824	DEX 1.2.3	1.021	1.036
Managing Diversity (MDI)			*Leading Change (LCH)*		
MDI 1.1.1	1.012	1.053	LCH 1.1.1	1.009	1.003
MDI 1.1.2	0.996	1.012	LCH 1.1.2	0.984	0.967
MDI 1.1.3	1.054	1.091	LCH 1.1.3	0.968	0.961
MDI 1.2.1	0.983	0.984	LCH 1.2.1	1.033	1.028
MDI 1.2.2	0.988	0.999	LCH 1.2.2	0.987	0.991
MDI 1.2.3	0.949	0.957	LCH 1.2.3	1.027	1.034
Developing Organizational Capabilities (DOC)					
DOC 1.1.1	1.007	1.012	DOC 1.2.1	1.037	1.045
DOC 1.1.2	0.970	0.982	DOC 1.2.2	0.988	0.960
DOC 1.1.3	0.993	0.992	DOC 1.2.3	1.004	1.012

Batas nilai *infit* dan *outfit* dikatakan *fit* berkisar antara 0,5 - 1,5 (Bond & Fox, 2015). Rentang skor *infit* kompetensi DLE adalah 0,961 - 1,074, sedangkan skor *outfit* berkisar 0,940 - 1,193. Skor *infit* kompetensi GBS memiliki rentang 0,824 - 1,158, dengan skor *outfit* antara 0,787 - 1,268. Aitem pada kompetensi SOR menunjukkan kecocokan yang baik, tercermin dari rentang nilai *infit* sebesar 0,904 - 1,102 dan *outfit* sebesar 0,869 - 1.128. Kompetensi DEX tidak memiliki *item misfit*, dengan nilai *infit* berkisar 0,961 - 1.038 dan *outfit* 0,966 - 1.309. Rentang *infit* pada kompetensi MDI adalah 0,949 - 1,054, serta nilai *outfit* 0,957 - 1,053. Kompetensi LCH memiliki rentang *infit* 0,968 - 1,033 dan *outfit* 0,961 - 1,034. Analisis aitem kompetensi DOC menunjukkan nilai *infit* antara 0,970 - 1,037 dan *outfit* 0,960 - 1.309.

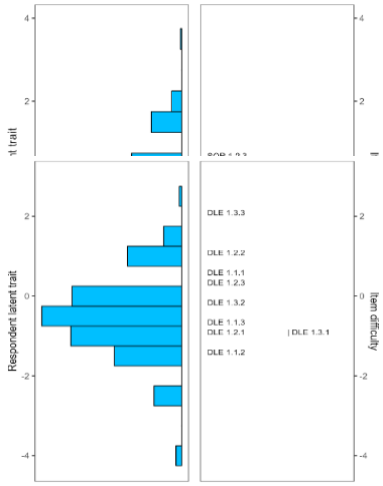
Berdasarkan Tabel 3, semua aitem dinilai *fit*, menandakan bahwa setiap aitem bermanfaat dan baik dalam mengukur kompetensi kepemimpinan yang dimiliki karyawan BUMN PT X. Berdasarkan hasil analisis, keseluruhan aitem memiliki *fit* yang baik mencerminkan bahwa setiap aitem sesuai dengan kompetensi yang ingin diukur dan mendukung keandalan tes (Rost & von Davier, 1994).

Wright Map

Wright Map atau *person-item map* menggambarkan hierarki antara kemampuan karyawan dengan tingkat kesulitan pada butir-butir skala yang sama (Boone, 2016; Engelhard Jr., 2013; Linacre, 2002). *Wright Map* terbagi menjadi dua area, area kiri menggambarkan distribusi abilitas responden menjawab soal, sedangkan area kanan mengilustrasikan sebaran tingkat kesukaran butir (Blanc & Rojas, 2018). Hasil pada area kanan sama dengan *Item Measure*, hanya saja *Item Measure* menyajikan informasi secara kuantitatif, sedangkan *Wright Map*

menggambarkan persebarannya. Kelompok karyawan berkompetensi tinggi terlihat pada bagian kiri atas, sedangkan semakin ke bawah mencerminkan kelompok karyawan berkemampuan yang lebih rendah. Aitem dengan tingkat kesulitan tinggi berada di bagian kanan atas, sementara aitem dengan tingkat kesulitan rendah ditempatkan di bagian kanan bawah

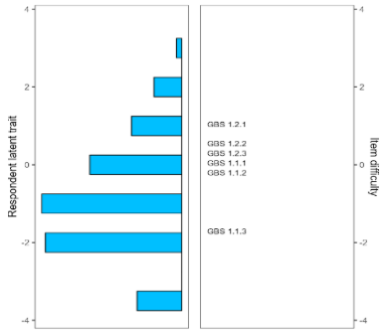
Gambar 1. *Wright Map* Kompetensi DLE



Berdasarkan distribusi *Wright Map* pada Gambar 1., area kanan terlihat bahwa 4 aitem kompetensi DLE (DLE 1.3.3, DLE 1.2.2, DLE 1.1.1, DLE 1.2.3) memiliki tingkat kesukaran di atas rata-rata, sementara 5 aitem lainnya (DLE 1.3.2, DLE 1.1.3, DLE 1.2.1, DLE 1.3.1, DLE 1.1.2) memiliki tingkat kesukaran di bawah rata-rata. *Wright Map* ini menunjukkan bahwa sebagian besar aitem kompetensi DLE hanya memotret abilitas rata-rata hingga tinggi. Namun, tidak ada aitem yang bisa memotret abilitas rendah.

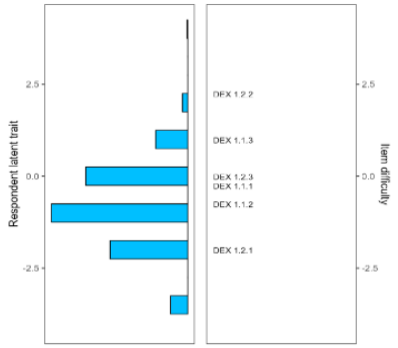
Gambar 2. *Wright Map* Kompetensi GBS

Wright Map pada Gambar 2., menunjukkan bahwa sebagian besar aitem kompetensi GBS hanya memotret abilitas rata-rata dan di bawah rata-rata. Namun, tidak ada aitem yang bisa memotret abilitas rendah dan abilitas tinggi.



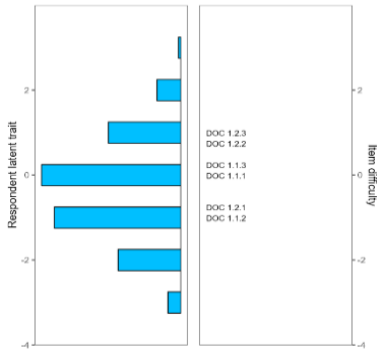
Gambar 3. *Wright Map* Kompetensi SOR

Wright Map pada Gambar 3., menunjukkan bahwa sebagian besar aitem kompetensi SOR hanya memotret abilitas rata-rata. Namun, tidak ada aitem yang bisa memotret abilitas rendah dan abilitas tinggi.



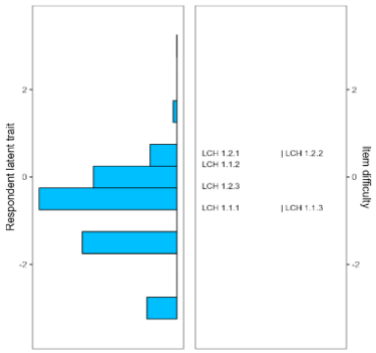
Gambar 4. *Wright Map* Kompetensi DEX

Wright Map pada Gambar 4., *Wright Map* ini menunjukkan bahwa sebagian besar aitem kompetensi DEX mampu memotret abilitas rendah, rata-rata, dan tinggi.



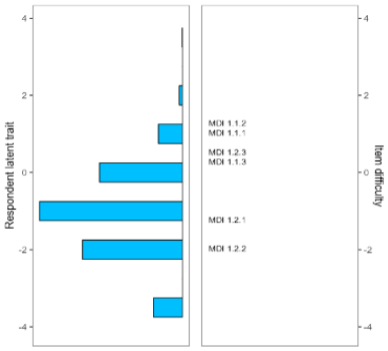
Gambar 5. *Wright Map* Kompetensi DOC

Wright Map pada Gambar 5., *Wright Map* ini menunjukkan bahwa sebagian besar aitem kompetensi DOC hanya memotret abilitas rata-rata. Namun, tidak ada aitem yang bisa memotret abilitas rendah dan abilitas tinggi.



Gambar 6. *Wright Map* Kompetensi LCH

Wright Map pada Gambar 6., *Wright Map* ini menunjukkan bahwa sebagian besar aitem kompetensi LCH mampu memotret abilitas rata-rata. Namun, tidak ada aitem yang bisa memotret abilitas rendah dan tinggi.



Gambar 7. *Wright Map* Kompetensi MDI

Wright Map pada Gambar 7., *Wright Map* ini menunjukkan bahwa sebagian besar aitem kompetensi MDI hanya memotret abilitas rata-rata dan di bawah rata-rata. Namun, tidak ada aitem yang bisa memotret abilitas rendah dan abilitas tinggi.

4. Pembahasan

Penelitian ini bertujuan untuk menguji butir *Situational Judgement Test* (SJT) Kompetensi Kepemimpinan yang disusunkan berdasarkan kamus kompetensi BUMN PT X dengan sampel pegawai level jabatan 1. Sebelumnya, tes ini telah melalui uji validitas konten untuk memastikan kesesuaian butir dengan definisi dan indikator perilaku per kompetensi akan tetapi perlu pengujian lebih lanjut untuk mengetahui kualitas psikometris butir-butir tes tersebut. Secara keseluruhan, penelitian ini menyimpulkan bahwa butir-butir SJT yang dikembangkan berdasarkan kamus kompetensi menunjukkan properti psikometris yang baik. Dengan kata lain, SJT terbukti efektif digunakan untuk mengukur kompetensi kepemimpinan pegawai.

Uji butir tes dilakukan dengan menggunakan *Rasch Model*. Pada tahap pertama, kami melakukan evaluasi dimensionalitas butir-butir SJT per kompetensi. Hasil uji dengan kriteria Martin-Löf menunjukkan bahwa empat dari tujuh kompetensi memiliki unidimensionalitas di antara butir-butir tesnya. Sementara, tiga kompetensi lainnya yaitu *Global Business Savvy* (GBS), *Strategic Orientation* (SOR), dan *Managing Diversity* (MDI) mengindikasikan multi dimensionalitas. Temuan ini memberikan gambaran bahwa perilaku – perilaku kunci (*key behaviors*) pada ketiga kompetensi ini cenderung tidak mengarah pada sebuah faktor laten tunggal, melainkan terdiri dari dari kelompok-kelompok perilaku yang mengarah pada kompetensi tertentu. Temuan ini dapat menjadi bahan evaluasi bagi perusahaan dalam penyusunan kamus kompetensi, apakah *key behaviors* yang ditetapkan didesain mengukur kompetensi yang bersifat unidimensional (faktor laten Tunggal) atau multidimensional. Akan tetapi dalam konteks pengukuran kompetensi, unidimensionalitas butir tes dapat dikesampingkan berdasarkan pertimbangan bahwa capaian kompetensi lebih menekankan pada ada atau tidaknya perilaku-perilaku kunci pada pegawai.

Selanjutnya, peneliti melakukan uji *item (mis)fitness*. Dalam konteks pengukuran, *item misfit* perlu diwaspadai karena dapat menurunkan keandalan tes, seperti memiliki daya diskriminasi yang buruk (Zubairi & Kassim, 2006). Pada penelitian ini, tidak ditemukan *item misfit* yang dapat menurunkan keandalan tes, sehingga kemungkinan memiliki daya diskriminasi yang baik, meskipun hal tersebut tidak dikaji dalam penelitian ini. Jika suatu aitem ditemukan tidak sesuai dengan *Rasch Model*, hal tersebut dapat menunjukkan adanya masalah dalam konstruksi aitem, seperti daya diskriminasi yang rendah (Zubairi & Kassim, 2006), atau bahkan mengisyaratkan bahwa parameter aitem tersebut mungkin tidak valid, artinya aitem sebenarnya mengukur kemampuan yang berbeda (Reise, 1990).

Wright Map membantu kita menjawab pertanyaan apakah aitem yang kita miliki sesuai dengan populasi yang kita target, sehingga kita memiliki aitem yang dapat mengumpulkan informasi tentang responden dalam keseluruhan rentang persebaran kemampuannya. Sebaliknya, juga membantu memastikan apakah terdapat keseimbangan yang tepat antara jumlah aitem yang mudah, sulit, dan berada di tengah rentang kemampuan, sehingga memastikan bahwa aitem-aitem tersebut relevan dan sesuai dengan sasaran yang ditetapkan (Katz, 2021).

Dalam analisis *Wright Map*, mayoritas aitem pada kompetensi *Driving Execution* (DEX) mampu memotret keseluruhan abilitas yakni rendah, rata-rata, hingga tinggi. Sementara itu, aitem pada kompetensi *Digital Leadership* (DLE) mampu memotret abilitas responden rata-rata hingga tinggi. Di sisi lain, mayoritas aitem pada kompetensi *Strategic Orientation* (SOR), *Developing Organizational Capabilities* (DOC), dan *Leading Change* (LCH) memotret abilitas rata-rata. Aitem pada kompetensi *Global Business Savvy* (GBS) dan *Managing Diversity* (MDI) mencerminkan kemampuan memotret abilitas pada tingkat rata-rata dan di bawah rata-rata.

Temuan dalam penelitian ini menunjukkan bahwa alat tes ini dapat secara meyakinkan digunakan sebagai instrumen dalam asesmen kompetensi. Hasil ini konsisten dengan pernyataan bahwa kualitas instrumen tes dapat

diukur dari kemampuannya memberikan informasi yang akurat mengenai kompetensi karyawan yang diujikan (Azizah & Wahyuningsih, 2020). Penelitian ini memperlihatkan bahwa butir SJT kompetensi kepemimpinan ini memiliki karakteristik psikometris yang baik, mampu mengukur kompetensi yang diinginkan. Hal ini sesuai dengan tujuan SJT, yaitu menangkap respons konstekstual peserta asesmen terhadap situasi pekerjaan yang dapat memprediksi kinerja peserta asesmen di masa mendatang (Lievens & Motowidlo, 2015).

Temuan dalam penelitian ini menunjukkan bahwa alat tes SJT yang digunakan dalam mengukur kompetensi kepemimpinan dapat digunakan secara berkelanjutan sebagai instrumen asesmen kompetensi. Hal ini konsisten dengan pernyataan bahwa kualitas instrumen tes dapat dinilai dari kemampuannya memberikan informasi akurat mengenai kompetensi karyawan (Azizah & Wahyuningsih, 2020). Meskipun demikian, sebaran taraf sukar butir yang tidak sesuai dengan sebaran abilitas responden dapat memengaruhi akurasi pengukuran. Ketidaksesuaian ini mengakibatkan pengukuran yang kurang akurat, terutama jika butir tes terlalu mudah atau terlalu sulit, sehingga instrumen tes tidak mampu secara efektif membedakan kemampuan kepemimpinan responden. Butir yang tidak mengikuti distribusi abilitas responden akan mengurangi reliabilitas tes dan membuatnya kurang informatif dalam mengevaluasi kompetensi. Dalam hal ini, *Wright Map* dapat menunjukkan ketidaksesuaian distribusi kesulitan butir dengan kemampuan responden, yang memerlukan perbaikan atau eliminasi butir agar tes tetap valid. Meski terdapat beberapa tantangan terkait sebaran taraf sukar butir, penelitian ini memperlihatkan bahwa butir SJT memiliki karakteristik psikometris yang baik dan mampu mengukur kompetensi yang diinginkan. Hal ini sejalan dengan tujuan SJT untuk menangkap respons kontekstual peserta asesmen terhadap situasi pekerjaan, yang pada gilirannya dapat memprediksi kinerja mereka di masa depan (Lievens & Motowidlo, 2015).

Keterbatasan dalam penelitian ini mencakup beberapa aspek yang perlu diperhatikan. Pertama, terbatasnya analisis ini yang hanya mengukur level kompetensi 1, tidak mengukur butir SJT yang ditujukan pada level 2. Selain itu, skripsi ini tidak memperdalam pemahaman kemampuan responden secara perorangan (*person analysis*), sehingga terdapat potensi untuk peningkatan dalam merinci dan memahami karakteristik individu dalam konteks analisis butir. Sebagai tambahan, parameter psikometris yang dijelaskan terbatas pada tingkat kesukaran dan *item fitness* saja, tanpa merinci parameter lain yang mungkin memiliki dampak signifikan dalam pengukuran. Oleh karena itu, penelitian ini memberikan ruang untuk pengembangan lebih lanjut dalam mengeksplor parameter lain yang relevan dalam analisis butir dengan *Rasch Model.*

5. Kesimpulan

Penelitian ini membuktikan bahwa butir *Situational Judgement Test* (SJT) kompetensi kepemimpinan karyawan BUMN PT X memiliki kualitas dan keandalan butir yang baik terlihat dari tingkat kesukaran aitem dan *item fitness* yang baik. Hal ini menunjukkan bahwa alat tes ini mampu mengukur karakteristik psikometris yang baik dan dapat digunakan secara berkelanjutan sebagai instrumen dalam asesmen kompetensi PT X.

Berdasarkan analisis dengan model *Rasch*, ditemukan bahwa sejumlah item memerlukan perhatian khusus guna meningkatkan kualitas instrument pengukuran. Terutama, perlu dipertimbangkan perbaikan atau penghapusan pada item DLE 1.3.3, DEX 1.2.2, dan DEX 1.2.1. Rekomendasi ini bertujuan untuk meningkatkan akurasi instrumen dan memastikan bahwa item-item tersebut dapat optimal dalam mengukur kompetensi yang diinginkan. Perbaikan pada item tersebut diharapkan dapat memberikan hasil yang lebih reliabel dan relevan bagi PT X dalam menggunakan instrument pada asesmen kompetensi. Penelitian selanjutnya disarankan untuk melakukan eksplorasi mendalam terhadap analisis butir, khususnya pada SJT dengan *Rasch Model*, termasuk dalam mengungkap karakteristik individu melalui *person analysis*. Peneliti juga dapat memperdalam analisis dengan mengeksplorasi parameter psikometris lain yang relevan seperti *Item Characteristic Curve* (ICC) untuk mengetahui suatu aitem dapat membedakan antara individu dengan tingkat kemampuan yang berbeda. Peneliti selanjutnya dapat melakukan analisis *Differential Item Functioning* (DIF) guna memastikan bahwa tingkat kesulitan aitem sesuai dengan tujuan pengukuran kompetensi. Hal ini memungkinkan Peneliti untuk lebih mendetail dan menyeluruh dalam memahami analisis butir dengan *Rasch Model*, terutama dalam konteks SJT. Eksplorasi ini dapat memberikan kontribusi berharga dalam pengembangan penelitian di bidang psikometri dan pengukuran.

6. Referensi

Affleck, P., Bowman, M., Wardman, M., Sinclair, S., & Adams, R. (2016). Can we improve on situational judgement tests? *British Dental Journal*, *220*(1), 9-10. https://doi.org/10.1038/sj.bdj.2016.17

Aiken, L. R. (1994). *Psychological Testing and Assessment* (8th ed.). Allyn & Bacon.

Anastasi, A., & Urbina, S. (1997). *Psychological Testing* (7th ed.). Upper Saddle River, NJ: Prectice Hall.

Ang, S., Van Dyne, L., & Rockstuhl, T. (2015). Cultural intelligence: Origins, conceptualization, evolution, and methodological diversity. In M. J. Gelfand, C.-Y. Chiu, & Y.-Y. Hong (Eds.), *Handbook of advances in culture and psychology*, Vol. 5, pp. 273–323). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780190218966.003.0006

Arikunto, S. (2008). *Dasar-Dasar Evaluasi Pendidikan*. Bumi Aksara.

Ashraf, Z. A., & Jaseem, K. (2020). Classical and modern methods in item analysis of test tools. *International Journal of Research and Review*, *7*(5), 397-403.

Azizah, A., & Wahyuningsih, S. (2020). Penggunaan Model Rasch untuk Analisis Instrumen Tes pada Mata Kuliah Matematika Aktuaria. *JUPITEK: Jurnal Pendidikan Matematika*, *3*(1), 45-50. https://doi.org/10.30598/jupitekvol3iss1pp45-50

Blanc, A., & Rojas, A. J. (2018). Use of Rasch Person-Item Maps to Validate a Theoretical Model for Measuring Attitudes toward Sexual Behaviors. *PLOS ONE*, *13*(8), e0202551. https://doi.org/10.1371/journal.pone.0202551

Bond, T.G., & Fox, C.M. (2015). *Applying the rasch model fundamental measurement in the human sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Boone, W. J. (2016). Rasch Analysis for Instrument Development: Why, When, and How? *CBE—Life Sciences Education*, *15*(4), rm4. https://doi.org/10.1187/cbe.16-04-0148

Courville, T. G. (2004). *An Empirical Comparison of Item Response Theory and Classical Test Theory Item/Person Statistics.* Unpublished Ph.D Dissertation, Texas A & M University.

Engelhard Jr., G. (2013). *Invariant Measurement: Using Rasch Models in the Social, Behavioral and Health Sciences*. https://doi.org/10.4324/9780203073636

Fernanda, J. W., & Hidayah, N. (2020). Analisis Kualitas Soal Ujian Statistika Menggunakan Classical Test Theory dan Rasch Model. *Square: Journal of Mathematics and Mathematics Education*, *2*(1), 49. https://doi.org/10.21580/square.2020.2.1.5363

Fitrianawati, M. (2017). *Peran Analisis Butir Soal Guna Meningkatkan Kualitas Butir Soal, Kompetensi Guru dan Hasil Belajar Peserta Didik*. https://publikasiilmiah.ums.ac.id/xmlui/handle/11617/9117

Guenole, N., Chernyshenko, O., Stark, S., & Drasgow, F. (2014). Are Predictions Based on Situational Judgement Tests Precise Enough for Feedback in Leadership Development? *European Journal of Work and Organizational Psychology*, *24*(3), 433-443. https://doi.org/10.1080/1359432x.2014.926890

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1985, December 31). *Item Response Theory*. SpringerLink. https://link.springer.com/book/10.1007/978-94-017-1988-9

Jumini, S., Madnasri, S., Cahyono, E., & Parmin, P. (2023, June). Analisis Kualitas Butir Soal Pengukuran Literasi Sains Melalui Teori Tes Klasik Dan Rasch Model. In *Prosiding Seminar Nasional Pascasarjana* (Vol. 6, No. 1, pp. 758-765). https://proceeding.unnes.ac.id/index.php/snpasca/article/view/2215

Karabatsos, G. (2000). A Critique of Rasch Residual Fit Statistics. *Journal of Applied Measurement, 1*(2), 152–176. https://pubmed.ncbi.nlm.nih.gov/12029176/

Katz, D., Clairmont, A., & Wilton, M. (2021). Chapter 3 the Rasch model | *Measuring what matters: Introduction to Rasch analysis in R. Bookdown.* https://bookdown.org/chua/new_rasch_demo2/viewdifficulty.html

Kementerian BUMN. (2019). *Kamus Kompetensi ASN di Lingkungan Kementerian BUMN*. https://jdih.bumn.go.id/storage/peraturan/PER%2004%20MBU%2010%202019.pdf

Kementerian BUMN. (2021). *Permen BUMN no. PER-11/MBU/07/2021 Tahun 2021*. https://peraturan.bpk.go.id/Details/181405/permen-bumn-no-per-11mbu072021-tahun-2021

Krabbe, P. F. M. (2017). Item Response Theory. *The Measurement of Health and Health Status*, 171–195. https://doi.org/10.1016/B978-0-12-801504-9.00010-6

Krumm, S., Lievens, F., Hüffmeier, J., Lipnevich, A. A., Bendels, H., & Hertel, G. (2015). How "situational" is judgment in situational judgment tests? *Journal of Applied Psychology*, *100*(2), 399–416. https://doi.org/10.1037/a0037674

Kurniawan, U., & Andriyani, K. D. (2018). Analisis Soal Pilihan Ganda dengan Rasch Model. *Statistika*, 6(1), 34-39. https://doi.org/10.26714/jsunimus.6.1.2018.%25p

Labola, Y. A. (2019). Konsep Pengembangan Sumber Daya Manusia Berbasis Kompetensi, Bakat dan Ketahanan dalam Organisasi. *Jurnal Manajemen & Kewirausahaan*, *7*(1), 28-35.

Lievens, F., & Motowidlo, S. J. (2015). Situational Judgment Tests: From Measures of Situational judgment to Measures of General Domain Knowledge. *Industrial and Organizational Psychology*, *9*(1), 3-22. https://doi.org/10.1017/iop.2015.71

Lievens, F., & Patterson, F. (2011). The Validity and Incremental Validity of Knowledge Tests, Low-Fidelity Simulations, and High-Fidelity Simulations for Predicting Job Performance in Advanced-Level High-Stakes Selection. *Journal of Applied Psychology*, 96(5), 927-940. https://doi.org/10.1037/a0023496

Lievens, F., Peeters, H., & Schollaert, E. (2008). Situational Judgment Tests: A Review of Recent Research. *Personnel Review*, *37*(4), 426-441.

Lievens, F., Buyse, T., & Sackett, P. R. (2005). Retest effects in operational selection settings: Development and test of a framework. *Personnel Psychology*, 58(4), 981–1007. https://doi.org/10.1111/j.1744-6570.2005.00713.x

Linacre, J.M. (2002). What Do Infit and Outfit Mean-Square and Standardized Mean?. *Rasch Measurement Transaction*, 16, 878.

Linacre, J.M. (2002). Understanding Rasch Measurement: Optimizing Rating Scale Category Effectiveness. *Journal of Applied Measurement*. 3. 85-106.

Linacre, J. M. (2012). Expected score ICC, IRF (Rasch-half-point thresholds). *Winsteps and Facets: Rasch Analysis + Rasch Measurement Software + 1PL IRT*. https://www.winsteps.com/winman/expectedscoreicc.htm

Linden, W. J., & Hambleton, R. K. (1997). Item Response Theory: Brief History, Common Models, and Extensions. *Handbook of Modern Item Response Theory*, 1-28. https://doi.org/10.1007/978-1-4757-2691-6_1

Muktamiroh, H., Herqutanto, H., Soemantri, D., & Purwadianto, A. (2021). The Potential of Situational Judgement Test as an Instrument of Ethical Competence Assessment: A Literature Review. *Jurnal Pendidikan Kedokteran Indonesia: The Indonesian Journal of Medical Education*, *10*(3), 314. https://doi.org/10.22146/jpki.53735

Musid, N. A., Matore, M. E., & Hamid, H. A. (2023, September 23). Inter-rater reliability for assessing digital leadership situational judgement test linguistic validation using Cohen kappa. *Journal for ReAttach Therapy and Developmental Diversities.* https://www.jrtdd.com/index.php/journal/article/view/1504

Olsen, L. W. (2003). *Essays on Georg Rasch and His Contributions to Statistics*. Københavns Universitet, Økonomisk Institut.

Passi, V., Doug, M., Peile, E., Thistlethwaite, J., & Johnson, N. (2010). Developing medical professionalism in future doctors: A systematic review. *International Journal of Medical Education*, 1, 19-29. https://doi.org/10.5116/ijme.4bda.ca2a

Rasch, G. (1966). An Item Analysis which Takes Individual Differences into Account. *British Journal of Mathematical and Statistical Psychology*, *19*(1), 49–57. https://doi.org/10.1111/j.2044-8317.1966.tb00354.x

Rasch, G. (1960). *Studies in Mathematical Psychology: I. Probabilistic Models for Some Intelligence and Attainment Tests*. Nielsen & Lydiche. https://psycnet.apa.org/record/1962-07791-000

Rost, J., & Von Davier, M. (1994). A Conditional Item-Fit Index for Rasch Models. *Applied Psychological Measurement*, 18(2), 171-182. https://doi.org/10.1177/014662169401800206

Sackett, P. R., Zhang, C., Berry, C. M., & Lievens, F. (2021). Revisiting meta-analytic estimates of validity in personnel selection: Addressing systematic overcorrection for restriction of range. *Journal of Applied Psychology*, 106(7), 1031–1052. https://doi.org/10.1037/apl0000994

Seol, H. (2020). *Item Analysis - Jamovi*. https://forum.jamovi.org/viewtopic.php?f=6&t=1385

Sumintono, B. (2017). Rasch Model Measurement as Tools in Assessment for Learning. Advaces in social science. *Education and Humanities Research*, 173.

Sumintono, B. (2014). Model Rasch untuk penelitian sosial kuantitatif.

Sumintono. (2013). Ukuran Sampel untuk Kalibrasi Aitem. Rasch Model: Riset Kuantitatif. https://deceng3.wordpress.com/2013/08/13/sampel/

Sumintono, B., & Widhiarso, W. (2013). *Aplikasi Model Rasch Untuk Penelitian Ilmu-Ilmu Sosial (Edisi Revisi)*. Trim Komunikata Publishing House.

Walsh, J. L., Woolley, M. R., Brady, M. F., Melick, S. R., & Carretta, T. R. (2021, December). *Air Force Officer Qualifying Test (AFOQT) form T: Psychometric Evaluation of the Situational Judgment Test.* DTIC. https://apps.dtic.mil/sti/citations/AD1157021

Widhiarso, W. (2021). *Panduan Penulisan Situational Judgment Test (SJT)*. Yogyakarta: UPAP Fakultas Psikologi UGM.

Widhiarso, W., Hidayat, R., & Anggoro, W. J. (2018). *Panduan Pengembangan Tes Penilaian Situasional (Situational Judgement Test)*. Yogyakarta: Fakultas Psikologi UGM & Pusat Penilaian Pendidikan Balitbang Kemdikbud.

Widhiarso, W. (2017). *Penerapan Model Rasch untuk Mengevaluasi Tes UKKS dan UKPS*. https://widhiarso.staff.ugm.ac.id/wp/wp-content/uploads/Widhiarso-Penerapan-Model-Rasch-Untuk-Mengevaluasi-Tes-UKKS-Dan-UKPS.pdf

Yukl, G. A. (2002). *Leadership in organizations* (5th ed.). Prentice Hall.

Zubairi, A.M., & Kassim, N.L.A. (2006). Classical and rasch analyses of dichotomously scored reading comprehension test items. Malaysian Journal of ELT Research, 2(1), 1-20. https://www.researchgate.net/publication/254504568