ISSN xxxx-xxxxx (Electronic)

Journal of Mathematical Psychometrics and Measurement Science

Volume 1 Number 1, 10-25 (2026)
https://doi.org/10.63203/xxxxx

Article

Evaluasi Karakteristik Butir Soal Tes Numerikal Differential Aptitude Test melalui Analisis Psikometrik

Lira Erwinda[1], Mira Marlina Idaini[2], Riris Miladul[3], Destri Astrianingsih[4], Yuda Syahputra[5]
[1234] Unversitas Bina Bangsa, Indonesia
[5] Universitas Indraprasta PGRI, Indonesia

Received : Apr 18, 2026
Revised : May 23, 2026
Accepted : May 23, 2026

Keywords:

Psychometric evaluation
Rasch model
Differential aptitude test
Numerical ability

ABSTRAK

Studi ini mengevaluasi karakteristik psikometrik subtes numerik dari Tes Bakat Diferensial menggunakan Model Rasch untuk menentukan kualitas dan kelayakan instrumen dalam mengukur kemampuan numerik. Desain evaluasi psikometrik kuantitatif digunakan yang melibatkan siswa sekolah menengah pertama yang dipilih melalui pengambilan sampel acak sederhana. Data dikumpulkan menggunakan tes bakat numerik dikotomis 40 item dan dianalisis menggunakan Winsteps versi 5.6. Temuan menunjukkan bahwa instrumen tersebut menunjukkan kualitas psikometrik keseluruhan yang baik, dengan reliabilitas yang memuaskan, indeks pemisahan yang dapat diterima, dan konsistensi internal yang memadai. Sebagian besar item sesuai dengan model Rasch, menunjukkan bahwa sebagian besar item tes berfungsi dengan tepat dalam mengukur kemampuan numerik. Distribusi kesulitan item berkisar dari sangat mudah hingga sangat sulit, menunjukkan bahwa instrumen tersebut mencakup berbagai tingkat kemampuan. Peta item-orang menunjukkan bahwa tes tersebut umumnya selaras dengan tingkat kemampuan sebagian besar responden, meskipun jumlah item yang sangat sulit terbatas. Analisis Fungsi Item Diferensial mengungkapkan bahwa sebagian besar item adil di seluruh kelompok, dengan hanya sejumlah kecil yang menunjukkan potensi bias. Secara keseluruhan, subtes numerik dari Tes Bakat Diferensial dianggap cukup layak untuk penilaian pendidikan, meskipun beberapa item memerlukan revisi dan evaluasi lebih lanjut.

Corresponding Author:

Lira Erwinda
Universitas Bina Bangsa
Email: lira.uniba.bk@gmail.com

Pendahuluan

Pengukuran psikologis merupakan komponen fundamental dalam progress pengambilan keputusan di bidang pendidikan, psikologis, maupun pengembangan sumber daya manusia. Instrumen pengukuran digunakan untuk memperoleh informasi mengenai karakteristik individu, seperti kemampuan, bakat, kompetensi, maupun potensi yang dapat menjadi dasar dalam penempatan akademik, seleksi pendidikan, perencanaan karier, hingga evaluasi performa (Baluku et al., 2021). Oleh karena itu, kualitas instrumen pengukuran menjadi aspek yang sangat penting, karena keputusan yang dihasilkan sangat bergantung pada akurasi, reliabilitas, validitas, dan keadilan alat ukur yang digunakan (Syahputra et al., 2025). Standar internasional dalam pengukuran pendidikan dan psikologis juga menekankan bahwa suatu tes harus memenuhi prinsip validitas, reliabilitas, fairness, serta ketepatan interpretasi skor agar hasil pengukuran dapat dipertanggung jawabkan secara ilmiah maupun praktis (Kilgus et al., 2021).

Lebih lanjut, terdapat berbagai jenis tes yang digunkan dalam pengukuran psikologis yaitu test intelegensi, minat, bakat, kepribadian dan sebagainya (Nalbandyan et al., 2026). Masing-masing memiliki kegunaan yang

digunakan sesuai dengan kebutuhan. Misalnya dalam konteks asesmen kemampuan, menggunakan tes bakat (*aptitude test*). Tes bakat (*aptitude test*) merupakan salah satu instrument yang banyak digunakan untuk mengidentifikasi potensi dalam berbagai domain kemampuan tertentu (Hapsari & Hidayat, 2024). Tes tersebut telah digunakan oleh 150 psikolog professional dari lima negara bagian di Amerika Serikat, dan tes tersebut terbukti mampu menunjukkan bakat individu secara baik (Donoso et al., 2010). Tes bakat dirancang bukan hanya untuk menggambarkan kemampuan saat ini, tetapi juga untuk memprediksi kemungkinan performa akademik maupun profesional pada masa mendatang (Santoso et al., 2022). Hal ini terbukti, bahwa 42% perusahan di Amerika telah menggunakan tes bakat yang digunakan untuk seleksi karyawan (Goldstein et al., 2020).

Selain itu, salah satu instrument *aptitude asesmen* yang telah lama digunakan secara luas adalah Differential Aptitude Test (DAT), yang mengukur berbagai dimensi kemampuan, termasuk kemampuan verbal, abstrak, mekanik, penalaran, dan numerical (Goble, 1998). DAT telah digunakan dengan melibattkan 2.118 siswa remaja di Yogyakarta, Indonesia (Setiawati et al., 2018). Subtes numerikal dalam DAT memiliki peran penting karena dirancang untuk mengukur kemampuan individu dalam memahami hubungan kuantitatif, melakukan penalaran berbasis angka, serta menyelesaikan masalah numerik yang relevan dengan tuntutan akademik maupun pekerjaan tertentu (Cupani & Cortez, 2016). Kemampuan numerikal sendiri sering dikaitkan dengan keberhasilan dalam bidang pendidikan, khususnya pada area yang menuntut kemampuan analitis dan pemecahan masalah kuantitatif (Reinhold et al., 2020).

Namun, penggunaan suatu instrumen psikologis, khususnya tes bakat, tidak dapat hanya bergantung pada reputasi historis atau penggunaan yang luas tanpa evaluasi empiris yang berkelanjutan. Karakteristik psikometrik suatu instrument dapat berubah ketika digunakan pada populasi, konteks budaya, bahasa, maupun priode waktu yang berbeda (Peixoto et al., 2021). International Test Commission menekankan bahwa instrumen yang digunakan lintas konteks harus melalui proses evaluasi yang memastikan kesetaraan fungsi pengukuran dan ketepatan interpretasi skor (Hernández et al., 2020). Selain itu, perkembangan karakteristik peserta tes, perubahan konteks pendidikan, serta dinamika kemampuan generasi pengguna juga dapat memengaruhi performa butir soal dari waktu ke waktu.

Pada level teknis, kualitas suatu tes sangat dipengaruhi oleh kualitas masing-masing butir penyusunnya. Butir yang terlalu mudah atau terlalu sulit, memiliki daya pembeda rendah, maupun mengandung *item-writing flaws* dapat menurunkan kualitas psikometrik instrument (Tarigan & Fadillah, 2022). Selain itu, ketidaksesuaian butir dengan konstruk yang diukur juga dapat memengaruhi validitas dan akurasi interpretasi hasil tes (Solichin, 2017). Dalam konteks subtes numerikal DAT, evaluasi terhadap karakteristik butir menjadi penting untuk memastikan bahwa setiap item benar-benar mampu merepresentasikan kemampuan numerikal yang dimaksud serta berfungsi secara optimal pada populasi yang menjadi sasaran pengukuran (Ajmi et al., 2024).

Selanjutnya, penelitian terdahulu menunjukkan bahwa pentingnya psikometrik terhadap instrument pengukuran untuk memastikan kualitas butir soal dan ketepatan inteprestasi hasil asesmen (Alkhatib et al., 2020; Kumar et al., 2021). Analisis psikometrik memungkinkan peneliti mengidentifikasi karakteristik item, seperti tingkat kesukaran, daya pembeda, kesesuaian model, reliabilitas, serta potensi bias pengukuran. Dalam konteks pengukuran pendidikan dan psikologi, evaluasi karakteristik butir menjadi langkah esensial karena kualitas setiap item secara langsung memengaruhi kualitas keseluruhan instrumen (Safitri & Baihaqi, 2026). Selain itu, dalam konteks kemampuan numerika, penelitian Hamid (2025) menunjukkan bahwa kemampuan numerical berkaitan erat dengan performa akademik, khususnya pada domain yang menuntut penalaran kuantitatif dan pemecahan masalah matematis. Penelitian tersebut mengindikasikan bahwa instrument pengukuran kemampuan numerikal perlu memiliki kualitas psikometrik yang baik agar interpretasi hasil tes benar-benar mencerminkan kemampuan aktual peserta.

Lebih dalam, perkembangan pendekatan psikometrik modern telah mendorong penggunaan model analisis yang lebih komprehensif dibanding pendekatan klasik. Jika *Classical Test Theory* (CTT) berfokus pada skor total dan statistik item dasar, pendekatan modern seperti *Rasch Model* maupun *Item Response Theory* (IRT) memungkinkan evaluasi yang lebih rinci terhadap fungsi masing-masing item (Boone & Staver, 2020). Pendekatan Rasch dan IRT memungkinkan evaluasi mendalam terhadap kualitas item, termasuk analisis *item fit*, reliabilitas, tingkat kesulitan, dan *Differential Item Functioning* (DIF) (Andrich & Marais, 2019).

Meskipun demikian, kajian yang secara khusus mengevaluasi kualitas psikometrik subtes numerikal pada Differential Aptitude Test (DAT) masih relatif terbatas. Sebagian penelitian terkait DAT lebih banyak berfokus pada

penggunaan instrumen sebagai alat asesmen bakat, adaptasi instrumen, atau pengujian validitas secara umum, dibanding evaluasi mendalam terhadap karakteristik butir pada subtes tertentu (Setiawati et al., 2018). Selain itu, sebagian besar instrumen aptitude test dikembangkan dalam konteks budaya dan populasi yang berbeda dengan konteks penggunaannya saat ini, sehingga diperlukan evaluasi ulang untuk memastikan kesesuaian fungsi pengukuran pada populasi lokal (Gökçe et al., 2021; Huang et al., 1997).

Di Indonesia, penggunaan instrumen psikologis impor dalam konteks asesmen pendidikan maupun seleksi masih cukup umum, termasuk instrumen aptitude assessment. Namun, bukti empiris mengenai performa psikometrik butir instrumen tersebut pada populasi Indonesia masih terbatas (Suwartono & Santoso, 2016; Tarigan & Fadillah, 2021). Padahal, karakteristik peserta tes, latar belakang pendidikan, konteks budaya, serta perkembangan kemampuan generasi pengguna dapat memengaruhi performa butir soal. Instrumen yang tidak dievaluasi secara empiris pada populasi target berisiko menghasilkan interpretasi skor yang kurang akurat, bahkan berpotensi menimbulkan bias dalam pengambilan keputusan asesmen (Bagaskara et al., 2025; Novieany et al., 2021).

Selanjutnya, dapat diketahui bahwasannya terdapat hal yang perlu diperhatikan. Masih terbatasnya penelitian yang secara khusus mengevaluasi karakteristik psikometrik subtes numerikal pada *Differential Aptitude Test* (DAT), karena penelitian terkait DAT masih berfokus pada penggunaan instrument sebagai alat asesmen bakar, adaptasi instrument, maupun pengujian validitas secara umum (Setiawati et al., 2018). Selanjutnya, kajian evaluative pada konteks Indonesia masih terbatas (Suwartono & Santoso, 2016; Tarigan & Fadillah, 2021). Terkahir, penggunaan pendekatan psikometrik modern untuk mengevaluasi instrumen aptitude klasik masih relatif terbatas dibanding penggunaannya pada instrumen pendidikan atau asesmen akademik lainnya (Andrich & Marais, 2019; Boone et al., 2013). Kesenjangan ini menunjukkan bahwa adanya kebutuhan untuk melakukan evaluasi empiris terhadap kualitas butir instrument secara mendalam.

Oleh karena itu, penelitian ini penting karena hasil aptitude assessment sering digunakan sebagai dasar dalam pengambilan keputusan akademik, penempatan pendidikan, maupun pertimbangan pengembangan karier. Jika instrumen yang digunakan memiliki kualitas psikometrik yang kurang memadai, maka hasil pengukuran yang diperoleh berpotensi tidak merepresentasikan kemampuan aktual individu secara akurat (Gonzalez et al., 2023; Palermo, 2022). Oleh karena itu, evaluasi karakteristik butir soal pada subtes numerikal DAT diperlukan untuk memastikan kualitas instrumen, meningkatkan akurasi interpretasi hasil asesmen, serta mendukung penggunaan alat ukur yang lebih valid dan reliabel dalam praktik psikologis.

Berdasarkan latar belakang tersebut, penelitian ini bertujuan untuk mengevaluasi karakteristik psikometrik butir soal pada subtes numerikal Differential Aptitude Test (DAT) melalui analisis psikometrik guna mengidentifikasi kualitas item, tingkat kesesuaian butir, serta kelayakan instrumen dalam mengukur kemampuan numerikal pada populasi penelitian.

Metode

Penelitian ini menggunakan pendekatan kuantitatif dengan desain evaluative psikometrik (*psychometric evaluation study*). Pendekatan ini digunakan untuk mengevaluasi kualitas butir soal pada subtes numerical *Differential Aptitude Test* (DAT) berdasarkan karakteristik psikometriknya. Evaluasi dilakukan untuk menilai kelayakan instrumen dalam mengukur kemampuan numerikal melalui analisis parameter butir, sehingga dapat memberikan gambaran mengenai kualitas pengukuran yang dihasilkan.

Partisipasi dalam penelitin ini adalah seluruh siswa seluruh siswa SMP Negeri 67 Jakarta. Teknik pengambilan sampel dalam penelitian ini menggunakan probability sampling dengan metode simple random sampling, yaitu setiap anggota populasi memiliki peluang yang sama untuk terpilih sebagai sampel penelitian (Sugiyono, 2023). Populasi penelitian terdiri atas siswa kelas VII, VIII, dan IX, kemudian dipilih sebanyak 71 peserta secara acak untuk menjadi sampel penelitian.

Instrument dalam penelitian ini adalah subtest numerikal dari Differential Aptitude Test (DAT) yang dikembangkan oleh (Bennett et al., 1990). DAT merupakan battery tes bakat yang dirancang untuk mengukur berbagai kemampuan spesifik individu, termasuk kemampuan verbal, abstrak, mekanik, spasial, dan numerikal. Subtes numerikal secara khusus dirancang untuk mengukur kemampuan individu dalam memahami hubungan kuantitatif, melakukan penalaran berbasis angka, serta menyelesaikan persoalan numerik. Instrument dalam

penelitian ini terdiri dari 40 item pernyataan dengan alternatif respons dikotomus, yaitu jawaban benar diberi skor 1 dan jawaban salah diberi skor 0, sehingga data yang dihasilkan sesuai untuk dianalisis menggunakan Model Rasch.

Data penelitian diperoleh dari hasil administrasi subtes numerikal DAT kepada partisipan sesuai prosedur pelaksanaan tes yang berlaku. Sebelum analisis dilakukan, data respons peserta diperiksa terlebih dahulu untuk memastikan kelengkapan jawaban dan konsistensi data. Respons yang tidak lengkap atau tidak memenuhi kriteria analisis dikeluarkan dari dataset penelitian. Selanjutnya, penelitian ini menjaga kerahasiaan identitas partisipan dengan menggunakan data secara anonim dan hanya untuk kepentingan akademik. Seluruh proses analisis dilakukan berdasarkan prinsip etika penelitian psikologi, termasuk menjaga keamanan data dan kerahasiaan informasi peserta.

Analisis data dalam penelitian ini dilakukan menggunakan pendekatan analisis psikometrik berbasis Model Rasch untuk mengevaluasi karakteristik butir soal pada subtes numerikal *Differential Aptitude Test* (DAT). Model Rasch dipilih karena mampu memberikan estimasi yang lebih objektif terhadap karakteristik item dan kemampuan responden dibandingkan pendekatan tes klasik, serta memungkinkan evaluasi terhadap kesesuaian butir, tingkat kesulitan item, reliabilitas instrumen, dan potensi bias pengukuran (Bond et al., 2021; Boone et al., 2014). Analisis dilakukan menggunakan perangkat lunak Winsteps versi 5.6, yang secara luas digunakan dalam analisis pengukuran berbasis Rasch untuk data dikotomus maupun politomus (Linacre, 2022).

Evaluasi karakteristik psikometrik instrumen dilakukan melalui beberapa parameter analisis. Pertama, item fit statistics digunakan untuk mengidentifikasi kesesuaian masing-masing butir soal terhadap model Rasch. Kesesuaian item menunjukkan sejauh mana respons peserta terhadap suatu butir sesuai dengan ekspektasi model pengukuran. Kriteria kelayakan item ditentukan berdasarkan nilai *Outfit Mean Square* (MNSQ) dalam rentang 0,5–1,5, nilai *Z-standard* (ZSTD) dalam rentang −2,0 hingga +2,0, serta nilai *Point Measure Correlation* yang positif, sebagaimana direkomendasikan dalam analisis Rasch (Boone et al., 2014; Sumintono, B., & Widhiarso, 2015).

Kedua, tingkat kesulitan butir (item difficulty) dianalisis berdasarkan nilai logit yang dihasilkan oleh model Rasch. Nilai logit digunakan untuk menunjukkan posisi relatif tingkat kesukaran setiap item dalam satu skala pengukuran linear. Semakin tinggi nilai logit suatu item, maka semakin tinggi tingkat kesulitan item tersebut bagi responden, sedangkan nilai logit yang lebih rendah menunjukkan item yang lebih mudah dikerjakan (Bond et al., 2021).

Ketiga, analisis reliabilitas dan separation index dilakukan untuk mengevaluasi konsistensi pengukuran instrumen serta kemampuan instrumen dalam membedakan tingkat kemampuan responden maupun tingkat kesulitan item. Reliabilitas person menunjukkan konsistensi respons peserta, sedangkan reliabilitas item menunjukkan kestabilan estimasi karakteristik item. Selain itu, nilai separation index digunakan untuk mengetahui sejauh mana instrumen mampu mengelompokkan responden berdasarkan tingkat kemampuan yang berbeda (Linacre, 2024).

Keempat, Wright Map (person-item map) digunakan untuk memvisualisasikan distribusi kemampuan responden dan tingkat kesulitan item dalam satu skala logit yang sama. Analisis ini memungkinkan peneliti untuk mengevaluasi kesesuaian antara tingkat kemampuan peserta dengan tingkat kesulitan butir soal, sehingga dapat diketahui apakah instrumen telah mampu mengukur kemampuan target secara optimal (Bond et al., 2021).

Selanjutnya, Differential Item Functioning (DIF) juga dianalisis untuk mendeteksi potensi bias item berdasarkan karakteristik kelompok tertentu, seperti jenis kelamin atau tingkat kelas. DIF digunakan untuk mengidentifikasi apakah suatu item memberikan keuntungan atau kerugian yang tidak proporsional kepada kelompok tertentu meskipun memiliki tingkat kemampuan yang setara. Analisis ini penting untuk memastikan fairness instrumen dalam pengukuran psikologis dan pendidikan (Linacre, 2024; Zumbo, 2007).

Hasil dan Pemabahasan

Hasil penelitian ini disajikan berdasarkan psikometrik berbasi Model Rasch untuk mengevaluasi kualitas pengukuran pada subtest numerical *Differential Aptitude Test* (DAT). Evaluasi dilakukan melalui analisis *summary*

statistics, *item fit statistics*, tingkat kesulitan butir, *Wright Map*, serta *Differential Item Functioning* (DIF) guna menilai kelayakan instrumen dalam mengukur kemampuan numerikal.

Analisis psikometrik dilakukan terhadap respons peserta pada 40 item subtes numerikal DAT menggunakan Model Rasch melalui Winsteps versi 5.6. Hasil summary statistics menunjukkan kualitas umum instrumen dari aspek reliabilitas dan kemampuan diskriminasi item maupun responden. Akan di sajikan pada tabel 1 dibawah ini

Tabel 1. Summary Statistics Model Rasch

Statistik	Value
Jumlah item	40
Jumlah responden	71
Item reliability	0.94
Person reliability	0.81
Item separation	3.85
Person separation	2.07
Cronbach Alpha	0.83
SEM	2.62

Berdasarkan tabel 1 menunjukkan hasil analisis *summary statistics* menggunakan Model Rasch, diperoleh gambaran umum mengenai kualitas psikometrik subtes numerikal *Differential Aptitude Test* (DAT) yang dianalisis terhadap 40 butir soal dengan melibatkan 71 responden. Hasil analisis menunjukkan bahwa nilai *item reliability* sebesar 0,94, yang mengindikasikan tingkat reliabilitas item yang sangat baik. Nilai tersebut menunjukkan bahwa instrumen memiliki konsistensi yang tinggi dalam mengestimasi tingkat kesulitan butir, sehingga kualitas item dalam instrumen ini dapat dikatakan stabil dan mampu membedakan tingkat kesulitan antarbutir secara akurat. Sementara itu, nilai *person reliability* sebesar 0,81 menunjukkan reliabilitas yang baik, yang berarti instrumen memiliki konsistensi yang memadai dalam membedakan kemampuan responden berdasarkan pola jawaban yang diberikan.

Selanjutnya, nilai *item separation* sebesar 3,85 menunjukkan bahwa instrumen mampu mengelompokkan butir soal ke dalam beberapa tingkat kesulitan yang berbeda secara jelas. Nilai ini mengindikasikan bahwa subtes numerikal DAT memiliki variasi tingkat kesulitan item yang cukup baik, sehingga mampu merepresentasikan rentang kemampuan numerikal yang beragam. Adapun nilai *person separation* sebesar 2,07 menunjukkan bahwa instrumen cukup mampu membedakan responden berdasarkan tingkat kemampuan numerikal yang dimiliki. Dengan demikian, instrumen ini memiliki kemampuan diskriminatif yang baik dalam mengidentifikasi kelompok responden dengan kemampuan yang berbeda.

Selain itu, nilai Cronbach Alpha sebesar 0,83 menunjukkan tingkat konsistensi internal instrumen yang baik, sehingga secara umum subtes numerikal DAT memiliki reliabilitas yang memadai sebagai alat ukur kemampuan numerikal. Nilai ini mengindikasikan bahwa interaksi antara responden dan item berjalan secara cukup konsisten dalam proses pengukuran. Sementara itu, nilai *Standard Error of Measurement* (SEM) sebesar 2,62 menunjukkan tingkat kesalahan pengukuran yang relatif rendah, sehingga estimasi skor yang dihasilkan oleh instrumen ini dapat dianggap cukup akurat. Secara keseluruhan, hasil *summary statistics* menunjukkan bahwa subtes numerikal DAT memiliki kualitas psikometrik yang baik dan layak untuk digunakan dalam pengukuran kemampuan numerikal pada populasi penelitian.

Selanjutnya analisis unidimensionalitas dilakukan untuk mengevaluasi sejauh mana instrumen mengukur satu konstruk utama yang sama, yaitu kemampuan numerikal. Dalam Model Rasch, unidimensionalitas dapat dievaluasi melalui *Principal Component Analysis of Residuals* (PCAR), dengan melihat proporsi varians yang mampu dijelaskan oleh model pengukuran (*raw variance explained by measures*) serta besarnya varians residual yang tidak terjelaskan pada kontras pertama (*unexplained variance in 1st contrast*). Menurut Sumintono & Widhiarso, (2015), instrumen dikatakan memiliki indikasi unidimensionalitas yang memadai apabila nilai *raw variance explained by measures* melebihi 20%. Akan ditampilkan pada tabel 2 sebagai berikut.

Berdasarkan tabel 2, diperoleh nilai *raw variance explained by measures* sebesar 32,2%, yang menunjukkan bahwa instrumen mampu menjelaskan proporsi varians yang cukup baik dalam merepresentasikan konstruk utama yang diukur. Nilai tersebut telah melampaui batas minimum yang direkomendasikan oleh Sumintono &

Widhiarso (2015), sehingga secara umum subtes numerikal *Differential Aptitude Test* (DAT) menunjukkan kecenderungan mengukur satu konstruk utama, yaitu kemampuan numerikal.

Tabel 2. Analisis Unidimensional

Indikator	Nilai
Raw variance explained by measures	32,2%
Unexplained variance in 1st contrast	4,30 (7,3%)
Unexplained variance in 2nd contrast	3,05 (5,2%)
Unexplained variance in 3rd contrast	2,70 (4,6%)
Unexplained variance in 4th contrast	2,59 (4,4%)
Unexplained variance in 5th contrast	2,20 (3,7%)

Namun demikian, hasil analisis juga menunjukkan bahwa nilai *unexplained variance in 1st contrast* sebesar 4,30 *eigenvalues* (7,3%), yang mengindikasikan masih adanya kemungkinan dimensi sekunder atau variasi lain di luar konstruk utama yang belum sepenuhnya dijelaskan oleh model. Kondisi ini dapat menunjukkan bahwa meskipun instrumen memiliki kecenderungan unidimensional, struktur pengukuran belum sepenuhnya optimal. Secara keseluruhan, hasil ini menunjukkan bahwa subtes numerikal DAT memiliki tingkat unidimensionalitas yang cukup memadai untuk digunakan dalam analisis Rasch, meskipun masih terdapat ruang untuk evaluasi lebih lanjut terhadap struktur internal instrumen.

Tabel 3. Item Fit Statistic (n=40)

No Item	Outfit MNSQ	Outfit ZSTD	Pt Measure Corr	Status
1	0.91	-0.26	0.32	Fit
2	0.67	-0.58	0.33	Fit
3	1.13	0.45	0.22	Fit
4	1.17	1.34	0.32	Fit
5	3.20	1.51	-0.09	Tidak Fit
6	1.24	1.83	0.22	Fit
7	0.91	-0.52	0.40	Fit
8	1.62	2.03	0.15	Tidak Fit
9	1.70	1.90	-0.07	Tidak Fit
10	0.92	-0.34	0.35	Fit
11	0.86	-1.16	0.53	Fit
12	0.92	-0.63	0.48	Fit
13	1.13	1.06	0.29	Fit
14	1.10	0.73	0.24	Fit
15	1.38	1.84	0.07	Fit
16	1.18	0.70	0.32	Fit
17	1.70	1.36	0.08	Tidak Fit
18	0.96	-0.30	0.43	Fit
19	0.77	-0.69	0.39	Fit
20	1.70	2.00	-0.09	Tidak Fit
21	0.75	-1.25	0.56	Fit
22	0.97	-0.11	0.48	Fit
23	0.85	-0.89	0.53	Fit
24	0.93	-0.57	0.47	Fit
25	0.80	-1.56	0.57	Fit
26	0.77	-1.96	0.56	Fit
27	0.69	-2.43	0.66	Tidak Fit
28	0.87	-0.99	0.47	Fit
29	0.90	-0.64	0.44	Fit
30	1.39	1.17	0.03	Fit
31	0.77	-1.65	0.58	Fit
32	0.78	-0.93	0.53	Fit
33	1.26	1.11	0.19	Fit
34	1.19	1.43	0.26	Fit
35	0.81	-1.37	0.54	Fit
36	0.84	-1.22	0.49	Fit

37	1.09	0.53	0.37	Fit
38	0.50	-0.66	0.35	Fit
39	0.85	-0.63	0.46	Fit
40	1.02	0.19	0.39	Fit

Selanjutnya, analisis *item fit statistics* dilakukan untuk mengevaluasi kesesuaian masing-masing butir soal terhadap Model Rasch. Kesesuaian butir menunjukkan sejauh mana respons peserta terhadap suatu item sesuai dengan ekspektasi model pengukuran, sehingga dapat digunakan untuk mengidentifikasi item yang berfungsi secara optimal maupun item yang berpotensi bermasalah.

Berdasarkan tabel 2, analisis *item fit statistics* dilakukan untuk mengevaluasi kesesuaian masing-masing butir soal terhadap Model Rasch berdasarkan indikator *Outfit Mean Square* (MNSQ), *Outfit Z-standard* (ZSTD), dan *Point Measure Correlation*. Item dinyatakan sesuai (*fit*) apabila memenuhi kriteria nilai *Outfit MNSQ* pada rentang 0,5–1,5, *Outfit ZSTD* pada rentang −2,0 hingga +2,0, serta memiliki nilai *Point Measure Correlation* positif. Berdasarkan hasil analisis terhadap 40 butir soal subtes numerikal *Differential Aptitude Test* (DAT), diperoleh sebanyak 34 butir soal yang memenuhi kriteria kesesuaian dengan Model Rasch, sedangkan 6 butir soal teridentifikasi tidak sesuai (*misfit*), yaitu item P5, P8, P9, P17, P20, dan P27.

Butir-butir yang teridentifikasi tidak fit menunjukkan adanya ketidaksesuaian antara pola respons peserta dengan ekspektasi Model Rasch. Ketidaksesuaian tersebut terlihat dari nilai *Outfit MNSQ* yang berada di luar rentang ideal, nilai *Outfit ZSTD* yang melebihi batas yang ditentukan, maupun nilai *Point Measure Correlation* yang negatif. Secara khusus, item P5 menunjukkan tingkat misfit paling tinggi dengan nilai *Outfit MNSQ* sebesar 3,20 dan *Point Measure Correlation* negatif (-0,09), yang mengindikasikan bahwa item tersebut kemungkinan tidak berfungsi secara optimal dalam mengukur konstruk kemampuan numerikal. Secara umum, hasil ini menunjukkan bahwa sebagian besar butir pada subtes numerikal DAT telah sesuai dengan Model Rasch, meskipun beberapa item memerlukan evaluasi lebih lanjut atau revisi agar kualitas pengukuran instrumen menjadi lebih optimal.

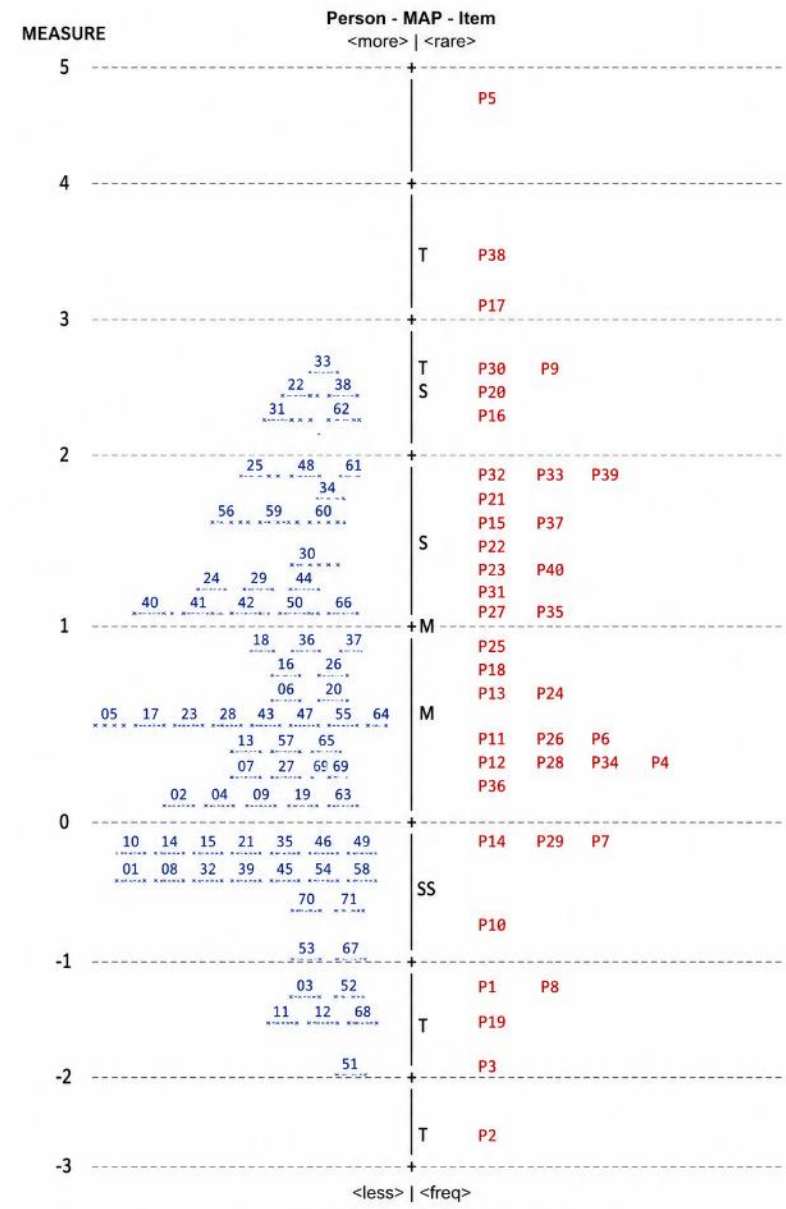
Tabel 4. Tingkat Kesulitan Butir (Item fit; N: 36)

Item	Measure (Logit)	Kategori
P38	2.79	Sangat Sulit
P30	1.44	Sulit
P16	1.22	Sulit
P32	0.93	Sedang
P33	0.93	Sedang
P39	0.84	Sedang
P21	0.76	Sedang
P15	0.59	Sedang
P37	0.59	Sedang
P22	0.51	Sedang
P40	0.43	Sedang
P23	0.36	Sedang
P31	0.21	Sedang
P35	0.14	Sedang
P25	0.00	Sedang
P18	-0.07	Mudah
P13	-0.21	Mudah
P24	-0.27	Mudah
P11	-0.47	Mudah
P6	-0.54	Mudah
P26	-0.54	Mudah
P4	-0.61	Mudah
P12	-0.61	Mudah
P28	-0.61	Mudah
P34	-0.67	Mudah
P36	-0.81	Mudah
P14	-0.94	Mudah
P29	-0.94	Mudah
P7	-1.01	Mudah
P10	-1.45	Mudah

P1	-1.86	Sangat Mudah
P19	-2.15	Sangat Mudah
P3	-2.50	Sangat Mudah
P2	-2.93	Sangat Mudah

Setelah dilakukan analisis kesesuaian butir terhadap Model Rasch, tahap selanjutnya adalah mengevaluasi tingkat kesulitan masing-masing item yang memenuhi kriteria *fit*. Analisis ini bertujuan untuk mengetahui distribusi tingkat kesukaran butir soal dalam subtes numerikal *Differential Aptitude Test* (DAT), sehingga dapat memberikan gambaran mengenai kemampuan instrumen dalam mengukur responden pada berbagai tingkat kemampuan numerikal.

Berdasarkan tabel 4, menunjukkan bahwa tingkat kesulitan butir dilakukan berdasarkan nilai *measure* (logit) yang dihasilkan dari Model Rasch. Semakin tinggi nilai logit suatu item, maka semakin tinggi tingkat kesulitan item tersebut, sedangkan nilai logit yang lebih rendah menunjukkan item yang lebih mudah dikerjakan oleh responden. Berdasarkan hasil analisis terhadap item yang memenuhi kriteria kesesuaian model, item dengan tingkat kesulitan tertinggi adalah P38 dengan nilai logit sebesar 2,79, sedangkan item termudah adalah P2 dengan nilai logit sebesar -2,93. Secara umum, distribusi tingkat kesulitan menunjukkan bahwa instrumen memiliki variasi tingkat kesukaran yang cukup beragam, mulai dari item yang sangat mudah hingga sangat sulit, sehingga mampu mengukur kemampuan numerikal responden pada berbagai tingkat kemampuan. Adapun visualisasi pada gambar 1 dibawah ini.



Gambar 1. Wright Map (Person-Item Map)

Gambar 1, menunjukkan bahwa Analisis *Wright Map (Person-Item Map)* dilakukan untuk memvisualisasikan distribusi kemampuan responden dan tingkat kesulitan butir soal dalam satu skala pengukuran yang sama berdasarkan Model Rasch. Peta ini memungkinkan evaluasi terhadap kesesuaian antara kemampuan responden dengan tingkat kesulitan item, sehingga dapat diketahui sejauh mana instrumen mampu mengukur kemampuan target secara optimal. Berdasarkan hasil analisis, distribusi tingkat kesulitan item menunjukkan rentang yang cukup luas, yaitu dari -2,93 logit hingga 4,28 logit, yang mengindikasikan bahwa subtes numerikal *Differential Aptitude Test* (DAT) memiliki variasi tingkat kesukaran dari sangat mudah hingga sangat sulit.

Berdasarkan *Wright Map*, item dengan tingkat kesulitan tertinggi adalah P5, yang berada pada posisi 4,28 logit, menunjukkan bahwa item tersebut merupakan butir yang paling sulit dikerjakan oleh responden. Sebaliknya, item dengan tingkat kesulitan terendah adalah P2, dengan nilai -2,93 logit, yang menunjukkan bahwa item tersebut merupakan butir paling mudah. Namun, berdasarkan hasil analisis *item fit statistics*, beberapa item seperti P5, P8, P9, P17, P20, dan P27 teridentifikasi sebagai item *misfit*, sehingga interpretasi terhadap item-item tersebut perlu dilakukan secara hati-hati.

Distribusi kemampuan responden menunjukkan bahwa mayoritas peserta berada pada rentang kemampuan -1 hingga 1 logit, dengan konsentrasi terbesar di sekitar nilai tengah (0 logit). Hal ini menunjukkan bahwa sebagian besar responden memiliki tingkat kemampuan numerikal pada kategori sedang. Jika dibandingkan dengan distribusi item, terlihat bahwa sebagian besar butir soal juga terkonsentrasi pada rentang kesulitan sedang hingga mudah, sehingga secara umum instrumen cukup sesuai dengan kemampuan mayoritas responden.

Meski demikian, *Wright Map* juga menunjukkan adanya beberapa item dengan tingkat kesulitan yang jauh lebih tinggi dibanding kemampuan sebagian besar responden, seperti P38, P17, P30, P9, dan P20, yang mengindikasikan bahwa item-item tersebut relatif sulit dijangkau oleh mayoritas peserta. Kondisi ini menunjukkan adanya ketidakseimbangan distribusi item pada level kemampuan tinggi. Di sisi lain, keberadaan item dengan tingkat kesulitan rendah seperti P1, P19, P3, dan P2 menunjukkan bahwa instrumen juga mampu mengukur responden dengan kemampuan lebih rendah.

Secara keseluruhan, hasil *Wright Map* menunjukkan bahwa subtes numerikal DAT memiliki cakupan tingkat kesulitan item yang cukup luas dan relatif mampu mengukur kemampuan responden pada berbagai level kemampuan. Namun demikian, distribusi item yang lebih terkonsentrasi pada tingkat kesulitan sedang hingga mudah, serta terbatasnya jumlah item pada tingkat kesulitan sangat tinggi, menunjukkan bahwa instrumen masih memiliki ruang untuk pengembangan agar pengukuran terhadap responden dengan kemampuan numerikal tinggi dapat dilakukan secara lebih optimal.

Selanjutnya, melakukan analisis *Differential Item Functioning* (DIF) dilakukan untuk mendeteksi adanya potensi bias pada butir soal berdasarkan perbedaan kelompok responden. Dalam penelitian ini, item dinyatakan mengalami bias apabila memiliki nilai signifikansi (*p-value*) kurang dari 0,05, yang menunjukkan bahwa item tersebut berfungsi secara berbeda antar kelompok meskipun responden memiliki tingkat kemampuan yang setara. Berdasarkan hasil analisis DIF terhadap 40 butir soal subtes numerikal *Differential Aptitude Test* (DAT), diperoleh bahwa sebagian besar item tidak menunjukkan indikasi bias. Sebanyak 38 butir soal memiliki nilai *p-value* di atas 0,05, sehingga dapat dikategorikan sebagai item yang tidak bias. Hal ini menunjukkan bahwa sebagian besar butir soal memiliki tingkat keadilan pengukuran (*fairness*) yang baik dan berfungsi secara konsisten pada kelompok responden yang dibandingkan.

Namun demikian, ditemukan 2 butir soal, yaitu P9 (p = 0,0232) dan P20 (p = 0,0358), yang teridentifikasi mengalami bias karena memiliki nilai signifikansi di bawah 0,05. Temuan ini menunjukkan bahwa kedua item tersebut berpotensi memberikan keuntungan atau kerugian yang tidak proporsional kepada kelompok tertentu, meskipun responden memiliki tingkat kemampuan numerikal yang setara. Secara keseluruhan, hasil analisis DIF menunjukkan bahwa subtes numerikal DAT memiliki tingkat fairness yang cukup baik, meskipun beberapa item memerlukan evaluasi lebih lanjut untuk memastikan keadilan pengukuran yang optimal.

Hasil penelitian menunjukkan bahwa subtes numerikal *Differential Aptitude Test* (DAT) secara umum memiliki kualitas psikometrik yang baik berdasarkan analisis Model Rasch. Hal ini tercermin dari nilai *item reliability* sebesar 0,94, *person reliability* sebesar 0,81, serta *Cronbach Alpha* sebesar 0,83. Dalam perspektif teori pengukuran Rasch, reliabilitas item yang tinggi menunjukkan bahwa hierarki tingkat kesulitan item dapat diestimasi secara stabil, sedangkan reliabilitas person menunjukkan kemampuan instrumen dalam membedakan tingkat kemampuan responden secara konsisten. Menurut Bond et al. (2020), reliabilitas yang tinggi mengindikasikan bahwa instrumen memiliki konsistensi pengukuran yang baik dan mampu menghasilkan estimasi kemampuan yang relatif stabil.

Tabel 5. Analisis Differential Item Functioning (DIF)

No Item	Chi-Square	p-value	Status
P1	3.9871	0.7812	Tidak Bias
P2	3.7111	0.8123	Tidak Bias
P3	2.8077	0.9022	Tidak Bias
P4	4.9242	0.6691	Tidak Bias
P5	7.5153	0.3771	Tidak Bias
P6	4.5152	0.7188	Tidak Bias
P7	7.8738	0.3437	Tidak Bias
P8	7.8478	0.3461	Tidak Bias
P9	16.2121	0.0232	Bias

P10	0.9828	0.9951	Tidak Bias
P11	6.5053	0.4820	Tidak Bias
P12	11.9371	0.1026	Tidak Bias
P13	7.5913	0.3699	Tidak Bias
P14	11.6599	0.1122	Tidak Bias
P15	11.7447	0.1092	Tidak Bias
P16	7.1566	0.4126	Tidak Bias
P17	11.7528	0.1089	Tidak Bias
P18	7.0354	0.4251	Tidak Bias
P19	9.2379	0.2359	Tidak Bias
P20	15.0109	0.0358	Bias
P21	3.4881	0.8364	Tidak Bias
P22	3.2810	0.8578	Tidak Bias
P23	4.7359	0.6921	Tidak Bias
P24	9.1598	0.2413	Tidak Bias
P25	3.3237	0.8535	Tidak Bias
P26	3.8210	0.8001	Tidak Bias
P27	10.4930	0.1622	Tidak Bias
P28	5.7820	0.5653	Tidak Bias
P29	3.8201	0.8002	Tidak Bias
P30	10.2186	0.1764	Tidak Bias
P31	5.8189	0.5609	Tidak Bias
P32	2.6792	0.9130	Tidak Bias
P33	7.2159	0.4066	Tidak Bias
P34	10.2964	0.1723	Tidak Bias
P35	5.0309	0.6561	Tidak Bias
P36	5.1917	0.6365	Tidak Bias
P37	4.4593	0.7256	Tidak Bias
P38	1.7245	0.9735	Tidak Bias
P39	6.9076	0.4384	Tidak Bias
P40	4.3409	0.7397	Tidak Bias

Temuan ini menunjukkan bahwa subtes numerikal DAT cukup memadai sebagai alat ukur kemampuan numerikal pada populasi penelitian. Nilai *item separation* sebesar 3,85 juga menunjukkan bahwa item mampu dikelompokkan ke dalam beberapa strata tingkat kesulitan yang berbeda, sedangkan *person separation* sebesar 2,07 mengindikasikan bahwa instrumen mampu membedakan responden ke dalam beberapa kelompok kemampuan. Secara teoritis, separation yang memadai menunjukkan kapasitas instrumen dalam melakukan diskriminasi pengukuran secara efektif. Dengan demikian, hasil ini sejalan dengan prinsip pengukuran objektif dalam Model Rasch yang menekankan kestabilan estimasi item dan person.

Dari sisi unidimensionalitas, nilai *raw variance explained by measures* sebesar 32,2% menunjukkan bahwa instrumen cukup mampu merepresentasikan konstruk utama yang diukur, yaitu kemampuan numerikal. Sumintono & Widhiarso (2015) menjelaskan bahwa proporsi varians di atas 20% menunjukkan indikasi unidimensionalitas yang dapat diterima dalam analisis Rasch. Namun demikian, nilai *unexplained variance in 1st contrast* sebesar 4,30 menunjukkan kemungkinan adanya dimensi sekunder dalam instrumen. Hal ini menunjukkan bahwa meskipun instrumen cenderung mengukur satu konstruk utama, masih terdapat variasi residual yang dapat disebabkan oleh karakteristik item tertentu atau heterogenitas respons peserta.

Selanjutnya, analisis *item fit statistics* menunjukkan bahwa sebanyak 34 dari 40 butir soal memenuhi kriteria kesesuaian dengan Model Rasch, sementara 6 item teridentifikasi *misfit*, yaitu P5, P8, P9, P17, P20, dan P27. Dalam teori Rasch, item yang fit menunjukkan bahwa pola respons peserta konsisten dengan ekspektasi model, sedangkan item misfit mengindikasikan adanya penyimpangan yang dapat menurunkan akurasi pengukuran. Temuan ini menunjukkan bahwa sebagian besar item pada subtes numerikal DAT telah berfungsi secara memadai dalam mengukur kemampuan numerikal. Namun, keberadaan item misfit mengindikasikan bahwa beberapa butir

belum sepenuhnya optimal. Secara khusus, item P5 menunjukkan tingkat misfit paling tinggi, dengan nilai *Outfit MNSQ* sebesar 3,20 dan *Point Measure Correlation* negatif. Kondisi ini menunjukkan bahwa respons peserta terhadap item tersebut tidak konsisten dengan pola kemampuan yang diharapkan.

Beberapa faktor dapat menjelaskan munculnya item misfit. Pertama, kemungkinan adanya redaksi soal yang ambigu sehingga peserta menafsirkan item secara berbeda. Kedua, item dapat menuntut kemampuan tambahan selain numerikal, seperti pemahaman verbal atau strategi penalaran tertentu, sehingga mengganggu unidimensionalitas konstruk. Ketiga, guessing behavior pada tes pilihan benar-salah juga dapat memengaruhi pola respons. Keempat, heterogenitas kemampuan responden dalam sampel dapat menyebabkan ketidakstabilan respons terhadap item tertentu.

Temuan ini sejalan dengan penelitian psikometrik sebelumnya yang menunjukkan bahwa item misfit sering muncul akibat ketidaksesuaian konstruk, kompleksitas redaksi, atau perilaku respons yang tidak konsisten. Oleh karena itu, item-item tersebut memerlukan evaluasi lebih lanjut sebelum digunakan secara luas dalam konteks asesmen. Beberapa faktor dapat menjelaskan munculnya *item misfit*. Pertama, redaksi soal yang ambigu memungkinkan peserta memberikan interpretasi yang berbeda terhadap item, sehingga respons yang muncul tidak sepenuhnya mencerminkan konstruk yang diukur. Kedua, beberapa item dapat menuntut kemampuan tambahan di luar kemampuan numerikal, seperti pemahaman verbal atau strategi penalaran tertentu, yang berpotensi mengganggu asumsi unidimensionalitas dalam model Rasch. Ketiga, perilaku menebak (*guessing behavior*) pada format tes benar–salah atau pilihan ganda dapat memengaruhi pola respons peserta. Keempat, heterogenitas kemampuan responden dalam sampel juga dapat menyebabkan respons terhadap item tertentu menjadi tidak stabil.

Hasil penelitian ini sejalan dengan (Fährmann et al., 2022; Hsu et al., 2020; Robitzsch, 2022; von Davier & Bezirhan, 2023) yang menyatakan bahwa *item misfit* sering kali dipengaruhi oleh ketidaksesuaian konstruk, kompleksitas redaksi item, serta perilaku respons yang tidak konsisten. Oleh karena itu, item-item yang teridentifikasi *misfit* memerlukan evaluasi dan revisi lebih lanjut sebelum digunakan secara luas dalam konteks asesmen psikologis maupun pendidikan.

Item misfit sering kali dipengaruhi oleh ketidaksesuaian konstruk, kompleksitas redaksi item, serta perilaku respons yang tidak konsisten. Oleh karena itu, item-item yang teridentifikasi *misfit* memerlukan evaluasi dan revisi lebih lanjut sebelum digunakan secara luas dalam konteks asesmen psikologis maupun pendidikan. Namun demikian, terdapat ketimpangan pada level kemampuan tinggi, di mana jumlah item dengan tingkat kesulitan sangat tinggi relatif terbatas. Akibatnya, instrumen mungkin kurang sensitif dalam membedakan responden dengan kemampuan numerikal tinggi. Sebaliknya, item pada kategori mudah hingga sedang lebih dominan, sehingga instrumen lebih optimal untuk populasi dengan kemampuan rendah hingga menengah.

Distribusi kemampuan peserta yang diperoleh dalam penelitian ini dapat dipengaruhi oleh beberapa faktor, seperti karakteristik sampel, tingkat perkembangan kognitif siswa SMP, serta kesesuaian konteks budaya instrumen yang digunakan. Beberapa item dalam Differential Aptitude Test (DAT) kemungkinan dikembangkan berdasarkan karakteristik populasi dan budaya yang berbeda dengan konteks peserta penelitian saat ini, sehingga interpretasi maupun tingkat kesulitan item dapat berubah ketika diterapkan pada populasi lokal. Oleh karena itu, proses adaptasi bahasa dan budaya menjadi aspek penting untuk memastikan bahwa instrumen tetap mampu mengukur konstruk secara akurat, relevan, dan adil sesuai karakteristik responden. Selain itu, evaluasi psikometrik melalui analisis validitas, reliabilitas, dan kesetaraan konstruk perlu dilakukan agar interpretasi hasil asesmen lebih tepat dalam konteks pendidikan Indonesia (Utari & Lestari, 2023).

Selanjutnya, hasil penelitian ini menunjukkan bahwa secara umum subtes numerikal DAT memiliki tingkat fairness yang baik. Dominasi item non-bias menunjukkan bahwa instrumen relatif konsisten dalam mengukur kemampuan numerikal tanpa terlalu dipengaruhi karakteristik kelompok. Namun, keberadaan item bias tetap penting diperhatikan. Bias dapat muncul karena perbedaan pengalaman belajar antar kelompok, konteks budaya, interpretasi bahasa soal, atau familiarity terhadap format item tertentu. Jika item bias tetap dipertahankan, maka terdapat risiko interpretasi hasil yang kurang adil. Temuan ini mendukung pentingnya evaluasi fairness, terutama ketika instrumen digunakan dalam konteks pengambilan keputusan pendidikan atau seleksi.

Secara praktis, hasil penelitian ini menunjukkan bahwa subtes numerikal DAT memiliki potensi yang cukup baik untuk digunakan dalam asesmen kemampuan numerikal pada konteks pendidikan. Mayoritas item menunjukkan kualitas psikometrik yang memadai, reliabilitas instrumen baik, dan tingkat fairness yang cukup

tinggi. Bagi praktisi psikologi pendidikan, hasil ini memberikan dasar empiris bahwa DAT masih relevan digunakan, namun memerlukan evaluasi item tertentu sebelum digunakan sebagai dasar pengambilan keputusan penting. Bagi pengembang instrumen, temuan item misfit dan item bias menjadi dasar untuk revisi atau kalibrasi ulang. Secara teoretis, penelitian ini memperkuat penggunaan pendekatan Rasch sebagai metode evaluasi instrumen psikologis klasik, khususnya aptitude test.

Penelitian ini memiliki beberapa keterbatasan. Pertama, jumlah sampel relatif kecil (71 responden), sehingga kestabilan estimasi parameter item masih terbatas. Kedua, sampel hanya berasal dari satu sekolah, sehingga generalisasi hasil ke populasi yang lebih luas perlu dilakukan dengan hati-hati. Ketiga, instrumen yang digunakan merupakan tes klasik yang dikembangkan dalam konteks budaya berbeda, sehingga kemungkinan terdapat pengaruh konteks lokal terhadap performa item. Keempat, analisis DIF dilakukan secara terbatas pada kelompok yang tersedia dalam data, sehingga fairness lintas karakteristik lain belum dapat dievaluasi secara menyeluruh.

Penelitian selanjutnya disarankan menggunakan jumlah sampel yang lebih besar dan lebih heterogen agar estimasi psikometrik menjadi lebih stabil. Pengambilan sampel dari berbagai sekolah atau wilayah juga penting untuk meningkatkan generalisasi temuan. Selain itu, item-item yang teridentifikasi misfit maupun bias perlu direvisi dan diuji ulang untuk memastikan kualitas pengukuran yang lebih baik. Penelitian berikutnya juga dapat membandingkan hasil evaluasi menggunakan Model Rasch dengan pendekatan psikometrik lain seperti IRT atau CFA untuk memperoleh gambaran validitas instrumen yang lebih komprehensif.

Kesimpulan

Penelitian ini bertujuan untuk mengevaluasi karakteristik psikometrik subtes numerikal *Differential Aptitude Test* (DAT) menggunakan pendekatan Model Rasch. Berdasarkan hasil penelitian, secara umum subtes numerikal DAT menunjukkan kualitas psikometrik yang cukup baik dan layak digunakan sebagai instrumen pengukuran kemampuan numerikal pada konteks populasi penelitian. Instrumen ini menunjukkan konsistensi pengukuran yang memadai, kemampuan diskriminasi yang cukup baik dalam membedakan tingkat kemampuan responden, serta kecenderungan mengukur konstruk utama yang relevan, yaitu kemampuan numerikal.

Hasil penelitian menunjukkan bahwa sebagian besar butir soal telah sesuai dengan model pengukuran yang digunakan, sehingga dapat berfungsi secara optimal dalam mengukur kemampuan numerikal. Selain itu, distribusi tingkat kesulitan item menunjukkan variasi yang cukup beragam, mulai dari item yang mudah hingga sulit, yang menunjukkan bahwa instrumen memiliki kemampuan untuk menjangkau responden dengan tingkat kemampuan yang berbeda. Visualisasi melalui *Wright Map* juga menunjukkan bahwa secara umum tingkat kesulitan item relatif sesuai dengan kemampuan mayoritas responden, meskipun masih terdapat keterbatasan dalam menjangkau responden dengan kemampuan numerikal yang lebih tinggi. Di sisi lain, analisis fairness menunjukkan bahwa sebagian besar item telah berfungsi secara adil, meskipun masih ditemukan beberapa butir yang memerlukan evaluasi lebih lanjut karena terindikasi bias.

Temuan ini menegaskan bahwa subtes numerikal DAT masih memiliki potensi sebagai alat ukur kemampuan numerikal dalam konteks asesmen pendidikan, namun belum sepenuhnya optimal. Keberadaan item yang tidak sesuai dengan model pengukuran, indikasi dimensi sekunder, serta potensi bias pada beberapa butir menunjukkan bahwa instrumen masih memerlukan pengembangan lebih lanjut untuk meningkatkan kualitas pengukuran.

Penelitian ini memberikan kontribusi terhadap bidang psikometri, khususnya dalam evaluasi instrumen aptitude test klasik menggunakan pendekatan psikometrik modern. Temuan ini memperkuat pentingnya evaluasi empiris terhadap instrumen psikologis yang telah lama digunakan, terutama ketika diterapkan pada populasi dan konteks budaya yang berbeda. Selain itu, penelitian ini memberikan kontribusi praktis bagi penggunaan instrumen asesmen dalam bidang pendidikan dengan menyediakan dasar empiris terkait kualitas pengukuran subtes numerikal DAT.

Secara praktis, hasil penelitian ini mengimplikasikan bahwa penggunaan subtes numerikal DAT dalam asesmen pendidikan perlu disertai evaluasi kualitas butir secara berkala, terutama jika digunakan sebagai dasar pengambilan keputusan penting seperti penempatan akademik, seleksi, atau perencanaan pendidikan. Revisi terhadap item yang bermasalah diperlukan agar instrumen mampu memberikan hasil pengukuran yang lebih akurat, adil, dan sesuai dengan karakteristik responden.

Penelitian ini memiliki beberapa keterbatasan, antara lain cakupan sampel yang terbatas, konteks penelitian yang hanya dilakukan pada satu lokasi, serta penggunaan instrumen yang dikembangkan dalam konteks budaya

yang berbeda dengan populasi penelitian. Selain itu, evaluasi bias item dalam penelitian ini masih terbatas pada karakteristik kelompok tertentu sehingga belum menggambarkan fairness instrumen secara menyeluruh.

Berdasarkan keterbatasan tersebut, penelitian selanjutnya disarankan untuk melibatkan sampel yang lebih luas dan beragam agar hasil penelitian memiliki daya generalisasi yang lebih baik. Selain itu, butir-butir yang teridentifikasi bermasalah perlu direvisi dan diuji kembali untuk memastikan kualitas pengukuran yang lebih optimal. Penelitian berikutnya juga dapat mengombinasikan pendekatan psikometrik lain untuk memperoleh evaluasi instrumen yang lebih komprehensif.

Referensi

Ajmi, M. Al, Mustakim, S. S., Roslan, S., & Almehrizi, R. (2024). Psychometric characteristics of the numerical ability test for Gulf students. *International Journal of Evaluation and Research in Education* , *13*(4), 2552–2561. https://doi.org/10.11591/ijere.v13i4.28917

Alkhatib, H. S., Brazeau, G., Akour, A., & Almuhaissen, S. A. (2020). Evaluation of the effect of items' format and type on psychometric properties of sixth year pharmacy students clinical clerkship assessment items. *BMC Medical Education*, *20*(1). https://doi.org/10.1186/s12909-020-02107-3

Andrich, D., & Marais, I. (2019). *A Course in Rasch Measurement Theory : Measuring in the Educational, Social and Health Sciences.* Springer Texts in Education. https://doi.org/https://doi.org/10.1007/978-981-13-7496-8

Bagaskara, R. S., Iman, H. I., Assa, N. A., & Yudiana, W. (2025). Properti Psikometri Indonesia Desirable Responding Scale. *Journal of Psychological Science and Profession*, *9*(3), 208–221. https://doi.org/10.24198/jpsp.v9i3.68135

Baluku, M. M., Mugabi, E. N., Nansamba, J., Matagi, L., Onderi, P., & Otto, K. (2021). Psychological Capital and Career Outcomes among Final Year University Students: the Mediating Role of Career Engagement and Perceived Employability. *International Journal of Applied Positive Psychology*, *6*(1), 55–80. https://doi.org/10.1007/s41042-020-00040-w

Bennett, G. K., Seashore, H. G., & Wesman, A. G. (1990). *Differential Aptitude Tests (5th ed.)* (San Antoni). TX: Psychological Corporation.

Bond, T. G., Yan, Z., & Heene, M. (2020). Applying the rasch model: Fundamental measurement in the human sciences. In *Applying the Rasch Model: Fundamental Measurement in the Human Sciences*. Psychology Press. https://doi.org/10.4324/9780429030499

Bond, T. G., Yan, Z., & Heene, M. (2021). *Applying the Rasch Model: Fundamental Measurement in the Human Sciences* (4th ed.). Routledge.

Boone, W. J., & Staver, J. R. (2020). Point Measure Correlation. *Advances in Rasch Analyses in the Human Sciences*, 25–38. https://doi.org/10.1007/978-3-030-43420-5_3

Boone, W. J., Staver, J. R., & Yale, M. S. (2013). *Rasch analysis in the human sciences*. Springer.

Boone, W. J., Stever, J. R., & Yale, M. S. (2014). *Rasch Analysis in the Human Science*. Springer.

Cupani, M., & Cortez, F. D. (2016). Análisis psicométricos del Subtest de Razonamiento Numérico utilizando el Modelo de Rasch. *Revista de Psicología*, *25*(2), 1–16. https://doi.org/10.5354/0719-0581.2016.44558

Donoso, O. A., Hernandez, B., & Horin, E. V. (2010). Use of psychological tests within vocational rehabilitation. *Journal of Vocational Rehabilitation*, *32*(3), 191–200. https://doi.org/10.3233/JVR-2010-0509

Fährmann, K., Köhler, C., Hartig, J., & Heine, J. H. (2022). Practical significance of item misfit and its manifestations in constructs assessed in large-scale studies. *Large-Scale Assessments in Education*, *10*(1). https://doi.org/10.1186/s40536-022-00124-w

Goble, D. (1998). Using the Differential Aptitude Tests for Selection and Prediction in Vocational Education and Training. *Australian Journal of Career Development*, *7*(1), 20–23. https://doi.org/10.1177/103841629800700107

Gökçe, S., Berberoğlu, G., Wells, C. S., & Sireci, S. G. (2021). Linguistic Distance and Translation Differential Item Functioning on Trends in International Mathematics and Science Study Mathematics Assessment Items. *Journal of Psychoeducational Assessment*, *39*(6), 728–745. https://doi.org/10.1177/07342829211010537

Goldstein, H., Pulakos, E., Passmore, J., & Semedo, C. (2020). The prevalence of cognitive tests in personnel selection. In *The Wiley Blackwell handbook of the psychology of recruitment, selection and employee retention*. https://ebrary.net/106495/psychology/prevalence_cognitive_tests_personnel_selection

Gonzalez, O., Georgeson, A. R., & Pelham, W. E. (2023). How Accurate and Consistent Are Score-Based Assessment

Decisions? A Procedure Using the Linear Factor Model. *Assessment*, *30*(5), 1640–1650. https://doi.org/10.1177/10731911221113568

Hamid, A. (2025). Hubungan antara Kemampuan Numerik dan Efektivitas Pemecahan Masalah Matematis Mahasiswa. *Jurnal Ilmiah Matematika (JIMAT)*, *6*(2), 602–612. https://doi.org/10.63976/jimat.v6i2.1077

Hapsari, A. D., & Hidayat, R. (2024). Assessing the Predictive Power of Aptitude Tests on Academic Achievement of Students in Science and Technology Majors. *Gadjah Mada Journal of Psychology*, *10*(2), 118–130. https://doi.org/10.22146/gamajop.83491

Hernández, A., Hidalgo, M. D., Hambleton, R. K., & Gómez-Benito, J. (2020). International test commission guidelines for test adaptation: A criterion checklist. *Psicothema*, *32*(3), 390–398. https://doi.org/10.7334/psicothema2019.306

Hsu, C. L., Jin, K. Y., & Chiu, M. M. (2020). Cognitive Diagnostic Models for Random Guessing Behaviors. *Frontiers in Psychology*, *11*. https://doi.org/10.3389/fpsyg.2020.570365

Huang, C. D., Church, A. T., & Katigbak, M. S. (1997). Identifying cultural differences in items and traits: Differential item functioning in the NEO Personality Inventory. *Journal of Cross-Cultural Psychology*, *28*(2), 192–218. https://doi.org/10.1177/0022022197282004

Kilgus, S. P., Eklund, K., von der Embse, N. P., Weist, M., Barber, A. J., Kaul, M., & Dodge, S. (2021). Structural Validity and Reliability of Social, Academic, and Emotional Behavior Risk Screener–Student Rating Scale Scores: A Replication Study. *Assessment for Effective Intervention*, *46*(4), 259–269. https://doi.org/10.1177/1534508420909527

Kumar, D., Jaipurkar, R., Shekhar, A., Sikri, G., & Srinivas, V. (2021). Item analysis of multiple choice questions: A quality assurance test for an assessment tool. *Medical Journal Armed Forces India*, *77*, S85–S89. https://doi.org/https://doi.org/10.1016/j.mjafi.2020.11.007

Linacre, J. M. (2022). *A User's Guide to WINSTEPS MINISTEP Rasch-Model Computer Programs*. In winsteps.com.

Linacre, J. M. (2024). *Winsteps Rasch Measurement Computer Program Version 5.6*. Winsteps.com.

Nalbandyan, R., Gilbert, J. B., Franco, V. R., & Domingue, B. W. (2026). Signposts on the Path From Nominal to Ordinal Scales: Moving From a Discrete to a Continuous View. *Educational and Psychological*, *1*(1). https://doi.org/10.1177/00131644261440556

Novieany, E., Satiadarma, M. P., & Idulfilastri, R. M. (2021). Pengujian Validitas Konstruk, Reliabilitas Internal, Dan Analisis Butir (Studi Adaptasi Alat Ukur Skrining Gangguan Bipolar Di Indonesia). *Jurnal Muara Ilmu Sosial, Humaniora, Dan Seni*, *5*(1), 39–46. https://doi.org/10.24912/jmishumsen.v5i1.9500.2021

Palermo, C. (2022). Rater characteristics, response content, and scoring contexts: Decomposing the determinates of scoring accuracy. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.937097

Peixoto, E. M., Zanini, D. S., & de Andrade, J. M. (2021). Cross-cultural adaptation and psychometric properties of the Kessler Distress Scale (K10): an application of the rating scale model. *Psicologia: Reflexao e Critica*, *34*(1). https://doi.org/10.1186/s41155-021-00186-9

Reinhold, F., Hofer, S., Berkowitz, M., Strohmaier, A., Scheuerer, S., Loch, F., Vogel-Heuser, B., & Reiss, K. (2020). The role of spatial, verbal, numerical, and general reasoning abilities in complex word problem solving for young female and male adults. *Mathematics Education Research Journal*, *32*(2), 189–211. https://doi.org/10.1007/s13394-020-00331-0

Robitzsch, A. (2022). Four-Parameter Guessing Model and Related Item Response Models. *Mathematical and Computational Applications*, *27*(6), 95. https://doi.org/10.3390/mca27060095

Safitri, Z., & Baihaqi, M. (2026). Quality Profile of Arabic Final Semester Assessment Items : A Psychometric Analysis. *Al-Lisan: Jurnal Bahasa*, *11*(1), 87–102. https://doi.org/10.30603/al.v11i1.7322

Santoso, A. P. Y., Nanditya, A. D., Rahmawati, A. N., & Al Hasna, A. S. (2022). Efektivitas Penggunaan Tes Dat (Differential Aptitude Test) Pada Pendidikan Di Indonesia. *Flourishing Journal*, *2*(2), 137–145. https://doi.org/10.17977/um070v2i22022p137-145

Setiawati, F. A., Izzaty, R. E., & Hidayat, V. (2018). Evaluasi Karakteristik Psikometrik Tes Bakat Differensial dengan Teori Klasik. *Humanitas*, *15*(1), 46. https://doi.org/10.26555/humanitas.v15i1.7249

Solichin, M. (2017). Mujianto Solichin Universitas Pesantren Tinggi Darul Ulum (Unipdu) Jombang Pendahuluan Kegiatan evaluasi dalam dunia pendidikan merupakan komponen integral dalam program pembelajaran di samping rencana pembelajaran (kurikulum), tujuan pembelajaran , b. *DIRASAT: Jurnal Manajemen & Pendidikan Islam*, *2*(2), 192–213. https://journal.unipdu.ac.id/index.php/dirasat/article/view/879/637

Sugiyono. (2023). *Metode Penelitian Kuatitatif Kualitatif & R&D* (Sutopo, Ed.; 5th ed.). A.

Sumintono, B., & Widhiarso, W. (2015). *Aplikasi Model Rasch untuk Penelitian I lmu-ilmu sosial (edisi revisi)*. Trim

Komunikata.

Suwartono, C., & Santoso, J. B. (2016). Attitudes Toward Psychological Test Use in Indonesia. *ANIMA Indonesian Psychological Journal*, *31*(4), 160–169. https://doi.org/10.24123/aipj.v31i4.575

Syahputra, Y., Rahmat, C. P., & Erwinda, L. (2025). *Instrumentasi Tes dalam Bimbingan dan Konseling*. CV Eureka Media Aksara.

Tarigan, M., & Fadillah. (2021). Properti Psikometri Struktur Intelegensi IST Subtes Verbal (Satzergaenzung, Wortauswahl, dan Analogien) berbahasa Indonesia. *Jurnal Muar Ilmu Sosial, Humaniora, Dan Seni*, *5*(1), 63–72.

Tarigan, M., & Fadillah, F. (2022). Properti Psikometrik Intelligenz Struktur Test Subtes Kemampuan Numerik (Rechenaufgaben dan Zahlen Reihen). *Intuisi : Jurnal Psikologi Ilmiah*, *13*(2), 155–170. https://doi.org/10.15294/intuisi.v13i2.31839

Utari, D., & Lestari, R. (2023). Metode Adaptasi Lintas Budaya Instrumen Kidscreen-27 Di Asia: Integrative Review. *Jambura Journal of Health Sciences and Research*, *5*(2), 474–484. https://doi.org/10.35971/jjhsr.v5i2.18195

von Davier, M., & Bezirhan, U. (2023). A Robust Method for Detecting Item Misfit in Large-Scale Assessments. *Educational and Psychological Measurement*, *83*(4), 740–765. https://doi.org/10.1177/00131644221105819

Zumbo, B. D. (2007). *Three generations of DIF analyses*.