E-ISSN: 2579-9258
P-ISSN: 2614-3038
Jurnal Cendekia: Jurnal Pendidikan Matematika
Volume 10, Nomor 03, Agustus-November 2026, pp. 2088- 2103

Interferensi Jenis Besaran: Mengapa Definisi Variabel yang Dinyatakan dalam Soal Gagal Mengatur Pemodelan Siswa

Rian Efendi[1✉], Rina Ananta Sumawardani Sitepu[2], Marthinus Yohanes Ruamba[3], Histia Sisilia Margareta[4]

[1,2,3,4] Program Studi Pendidikan Matematika, Fakultas Keguruan dan Ilmu Pendidikan, Universitas Cenderawasih, Jl. Kamp Wolker Yabansai, Jayapura, Papua 99351
rianefendi07@gmail.com

Abstract

Indonesian contextual mathematics items often define the variables for students (e.g., "let x be the price of one noken (a woven bag)") to ease modelling, yet whether students use them is untested. We tested this, traced calculation errors, and checked whether the scoring method changes conclusions about interpretation. In this case study, 17 grade 9 students at a state Islamic school in Jayapura completed six multi-step items on linear equation systems in Papuan contexts; seven were interviewed. Two modelling (formulate) items both defined the variables. When the item sought prices among price data, almost all students succeeded (mean 1.76 out of 2). When it sought weights among mostly price data, almost all failed (mean 0.18), and many confused weights with prices. We call this referent-type interference: the dominant kind of number overrides the written definition. Calculation (employ) errors came mainly from subtracting large numbers, and only one student checked the solution in both equations. Conventional scoring against the key rated interpretation (interpret) as very weak (0.35). Conditional scoring, which judges answers against students' own earlier values, raised this to 1.00, but the entire increase came from identical, probably copied, answers. Stated definitions therefore do not guarantee that students can build a model themselves. Item writers should include items whose quantity sought differs in kind from the dominant numbers, and teachers should report both scores while ensuring independent work.
Keywords: conditional scoring, contextual assessment, mathematical literacy, mathematical modelling, referent-type interference

Abstrak

Soal kontekstual di Indonesia sering langsung menuliskan arti variabel, misalnya "misalkan x adalah harga satu noken", agar siswa mudah membuat model, tetapi belum diketahui apakah siswa benar-benar memakainya. Penelitian ini menguji hal itu, menelusuri kesalahan menghitung, dan memeriksa pengaruh cara penskoran terhadap penilaian kemampuan menafsirkan. Studi kasus ini melibatkan 17 siswa kelas IX sebuah madrasah negeri di Kota Jayapura yang mengerjakan enam soal bertahap tentang sistem persamaan linear dua variabel berkonteks Papua; tujuh di antaranya diwawancarai. Dua soal menyusun model (formulate) sama-sama menuliskan arti variabel. Saat yang dicari harga di antara data harga, hampir semua siswa berhasil (rerata 1,76 dari 2). Saat yang dicari berat di antara data yang didominasi harga, hampir semua gagal (rerata 0,18) dan banyak yang mencampuradukkan berat dengan harga. Gejala ini disebut interferensi jenis besaran: jenis angka yang mendominasi soal mengalahkan arti variabel yang tertulis. Kesalahan menghitung (employ) terutama berasal dari pengurangan bilangan besar, dan hanya satu siswa memeriksa jawaban pada kedua persamaan. Penskoran biasa berdasarkan kunci jawaban menilai kemampuan menafsirkan (interpret) sangat rendah (0,35). Penskoran bersyarat, yang berpatokan pada angka hasil siswa sendiri pada langkah sebelumnya, menaikkannya menjadi 1,00, tetapi seluruh kenaikan itu berasal dari jawaban identik yang kemungkinan disalin. Jadi, menuliskan arti variabel dalam soal tidak menjamin siswa mampu menyusun model sendiri. Penyusun soal perlu menyertakan soal yang besaran dicarinya berbeda jenis dari angka yang dominan, sedangkan guru perlu melaporkan kedua skor dan memastikan siswa bekerja mandiri.
Kata kunci: asesmen kontekstual, interferensi jenis besaran, literasi matematika, pemodelan matematis, penskoran bersyarat

Copyright (c) 2026 Rian Efendi, Rina Ananta Sumawardani Sitepu, Marthinus Yohanes Ruamba, Histia Sisilia Margareta

✉ Corresponding author: Rian Efendi
Email Address: rianefendi07@gmail.com (Jl. Kamp Wolker Yabansai, Jayapura, Papua 99351)
Received 04 October 2026, Accepted 18 October 2026, Published 30 October 2026
DoI: https://doi.org/10.31004/cendekia.v10i3.5593

PENDAHULUAN

Literasi matematika dalam kerangka PISA mencakup kemampuan memformulasikan situasi secara matematis (*formulate*), menggunakan konsep dan prosedur (*employ*), serta menafsirkan dan mengevaluasi hasil (*interpret*) (OECD, 2023a). Ketiga proses itu berantai, sehingga kegagalan pada satu tahap merambat ke tahap berikutnya dan skor yang hanya mencatat jawaban akhir tidak dapat menunjukkan letak kegagalan (Galbraith & Stillman, 2006). Capaian literasi matematika siswa Indonesia masih rendah (OECD, 2023b, 2026), dan persoalannya lebih tajam di Papua, tempat capaian numerasi rendah, sumber daya belum merata (Efendi et al., 2025), dan angka putus sekolah tinggi (Khairani et al., 2025).

Kajian mutakhir menempatkan hambatan utama pada peralihan dari situasi ke model matematis. Siswa terutama kesulitan mengenali informasi yang relevan dan menentukan besaran yang dicari (Schukajlow et al., 2023), matematisasi dapat gagal terlepas dari kompetensi pada tahap sesudahnya (Cevikbas et al., 2022), pembangunan model dunia nyata menjadi langkah pembatas (Krawitz et al., 2022), dan pemahaman teks berkontribusi terpisah dari perhitungan terhadap keberhasilan soal cerita (Strohmaier et al., 2026). Di Indonesia, kesalahan siswa pada soal PISA berkonteks paling banyak terjadi pada tahap memahami dan mentransformasi masalah (Wijaya et al., 2014); pada SPLDV, siswa sering tidak memisalkan variabel dan tidak memeriksa kembali jawaban (Pradana & Murtiyasa, 2020), serta terhambat saat menerjemahkan pernyataan verbal ke bentuk aljabar (Wahyuni et al., 2025).

Arus utama penelitian di Indonesia menjawab persoalan itu dengan tugas berkonteks budaya. Kajian berbasis etnomatematika dan pendidikan matematika realistik melaporkan bahwa konteks yang akrab menopang keterlibatan dan penalaran siswa (Darmawijoyo et al., 2025; Prahmana & D'Ambrosio, 2020). Asumsi bahwa ketidakakraban konteks adalah hambatan yang cukup dihilangkan justru diperdebatkan di tingkat internasional: Rubel & McCloskey (2021) mempertanyakan dunia siapa yang diwakili sebuah tugas berkonteks, sedangkan Muskens et al. (2024) melaporkan bahwa muatan dunia nyata menurunkan skor siswa dari keluarga berstatus sosial ekonomi rendah.

Penelitian aljabar menjelaskan mengapa transformasi itu sulit. Pembacaan huruf sebagai label objek, bukan wadah bagi nilai, dominan dan bertahan meskipun telah diajarkan (Klingbeil & Moons, 2025); penerjemahan pernyataan verbal ke simbol gagal secara sistematis (Castro et al., 2022); dan kesalahan pembalikan (*reversal error*) lebih dijelaskan oleh pencocokan urutan kata daripada makna besaran (Soneira et al., 2021) . Perilaku simbolik yang tampak benar pun dapat bertumpu pada pola visual, bukan makna (Kirshner, 1989). Pada tingkat yang lebih umum, pemula menggolongkan masalah menurut ciri permukaan, bukan struktur dalam (Chi et al., 1981), dan ciri permukaan tugas memprediksi capaian soal cerita lebih kuat daripada struktur matematisnya (Vessonen et al., 2024).

Persoalan lain menyangkut cara skor dibaca. Penelitian soal tipe PISA di Indonesia yang memakai penskoran konvensional cenderung menempatkan *interpret* sebagai proses terlemah (Gustiningsi et al., 2026; Hamidy & Jailani, 2019). Padahal, pada butir berantai, kegagalan *interpret* dapat merupakan rambatan dari nilai keliru pada tahap *employ*. Butir yang dirangkai terbukti

melanggar asumsi independensi lokal sehingga penskoran per butir dapat menyesatkan (Eaton et al., 2020) , dan kerangka diagnostik mutakhir dibangun untuk menemukan letak kegagalan dalam penyelesaian bertingkat (Säfström et al., 2024; Sidenvall et al., 2026). May et al. (2023) menemukan bahwa penskoran dikotomis dan kredit parsial berkorelasi kuat, tetapi pertanyaan yang berbeda, yaitu apakah respons dinilai terhadap kunci atau terhadap keluaran siswa sendiri, belum diuji pada butir literasi berantai.

Dari peta kajian tersebut tampak tiga celah. Pertama, penelitian tentang variabel membahas cara siswa membaca simbol, bukan apa yang terjadi ketika soal telah menetapkan makna simbol itu. Padahal, asesmen kontekstual di Indonesia lazim menyatakan definisi variabel di dalam soal, misalnya "misalkan *x* adalah harga satu ...", agar hambatan yang dianggap bukan sasaran pengukuran tersingkir; akibatnya, butir semacam itu mungkin tidak lagi mengukur *formulate*. Kedua, belum ada kajian yang menguji konvensi tersebut ketika jenis besaran yang dicari berbeda dari jenis bilangan yang mendominasi stimulus, padahal kategorisasi berdasarkan ciri permukaan memprediksi bahwa definisi tertulis dapat kalah oleh tampilan halaman. Ketiga, penelitian proses literasi di Indonesia belum melaporkan penskoran bersyarat berdampingan dengan penskoran konvensional, apalagi menguji kepekaannya terhadap respons yang tidak independen. Kebaruan penelitian ini terletak pada (1) identifikasi interferensi jenis besaran sebagai mode kegagalan konvensi definisi variabel, (2) pemanfaatan kontras alami dua butir yang sama-sama menyatakan definisi variabel dalam satu paket asesmen, dan (3) pelaporan ganda skor *interpret* yang disertai analisis kepekaan terhadap penyalinan.

Penelitian ini menjawab tiga pertanyaan: (1) bagaimana capaian *formulate* berbeda di antara dua butir yang sama-sama menyatakan definisi variabel tetapi berbeda jenis besaran yang dicari; (2) mekanisme apa yang mencirikan proses *employ* dan bagaimana kesalahannya merambat; dan (3) sejauh mana penskoran konvensional dan bersyarat memberi gambaran berbeda tentang kemampuan *interpret*. Hasilnya diharapkan bermanfaat bagi penyusun butir asesmen kontekstual, guru yang memakai skor untuk remediasi, dan peneliti asesmen literasi di daerah dengan sumber daya terbatas.

METODE

Desain, Latar, dan Partisipan

Penelitian ini merupakan studi kasus kualitatif dengan data skor pendukung yang memadukan analisis pekerjaan tertulis dan wawancara asesmen dinamis. Asesmen dinamis dipilih karena bantuan berjenjang memperlihatkan apa yang dapat dilakukan siswa secara mandiri sekaligus apa yang menjadi mungkin setelah satu hambatan disingkirkan (Elliott et al., 2018). Penelitian dilaksanakan di MTs Negeri Kota Jayapura, Distrik Muara Tami, satu-satunya distrik Kota Jayapura yang berbatasan darat dengan Papua Nugini; kesulitan pemecahan masalah siswa di sekolah ini telah terdokumentasi (Efendi & Kulimbang, 2025). Partisipan adalah seluruh 17 siswa kelas IX D pada semester ganjil 2026/2027 (sensus kelas) yang diberi kode S1 sampai S17. Tujuh siswa (S1 sampai S7) dipilih untuk diwawancarai secara purposif dengan variasi maksimum berdasarkan kejelasan jejak tulisan, kekhasan

kesalahan, kelengkapan tiga proses, kemampuan menjelaskan, dan adanya kasus menyimpang. Izin diberikan oleh kepala madrasah, persetujuan tertulis diperoleh dari orang tua atau wali, kesediaan diperoleh dari siswa, dan seluruh nama diganti kode sebelum analisis.

Instrumen dan Validasi

Instrumen terdiri atas enam butir uraian SPLDV yang dibangun atas dua stimulus berlatar niaga Papua (Tabel 1). Stimulus A berupa komik enam panel tentang lapak noken rajut dan ukiran kayu mini dengan dua catatan pembelian. Stimulus B berupa bacaan tentang belanja ikan mujair dan udang di Pasar Hamadi dengan tiga kendala: berat total tepat 4 kg, belanjaan berupa campuran, dan uang Rp340.000,00. Tugas 1 menunjuk butir 1 sampai 3 dan Tugas 2 menunjuk butir 4 sampai 6. Karena SPLDV termasuk kurikulum kelas VIII, asesmen ini menguji konten tahun sebelumnya.

Tabel 1 Pemetaan butir terhadap proses PISA, besaran yang dicari, dan kunci jawaban

Butir	Proses	Tuntutan tugas	Besaran yang dicari	Kunci
1	*Formulate*	Menyusun SPLDV dari dua catatan pembelian; definisi x dan y sebagai harga satuan dinyatakan dalam soal	Harga (sejenis dengan data)	$2x + 3y = 310.000$; $3x + 2y = 340.000$
2	*Employ*	Menentukan harga satu noken dan satu ukiran	Harga satuan	$x = 80.000$; $y = 50.000$
3	*Interpret*	Memutuskan apakah Rp200.000,00 cukup untuk satu noken dan dua ukiran, disertai alasan	Keputusan	Cukup, sisa Rp20.000,00
4	*Formulate*	Menyusun SPLDV belanja; soal menyatakan "gunakan x untuk berat ikan mujair (kg) dan y untuk berat udang (kg)"	Berat (berbeda jenis dari data harga)	$x + y = 4$; $50.000x + 120.000y = 340.000$
5	*Employ*	Menghitung berat tiap belanjaan agar seluruh uang terpakai	Berat	$x = 2$ kg; $y = 2$ kg
6	*Interpret*	Udang tersisa 1 kg: menentukan kuantitas baru, total biaya, dan kembalian	Kuantitas dan uang	3 kg mujair; Rp270.000,00; sisa Rp70.000,00

Catatan. Butir 1 dan 4 menjadi pasangan utama: keduanya menyatakan definisi variabel dan berbeda pada jenis besaran yang dicari. Butir 2, 3, 5, dan 6 bekerja di atas keluaran butir sebelumnya.

Validitas isi ditelaah oleh tiga ahli, yaitu dua dosen pendidikan matematika dan satu guru matematika sekolah tersebut, yang menilai 56 pernyataan pada skala 1 sampai 4. Koefisien Aiken dihitung dengan $V = \Sigma(r - lo)/[n(c - 1)]$, dengan r skor penilai, lo skor terendah, $n = 3$, dan $c = 4$ (Aiken, 1985), lalu dibaca secara deskriptif karena panelnya kecil (Stefana et al., 2025). Koefisien keseluruhan $V = 0{,}70$, dan 50 dari 56 pernyataan mencapai $V \geq 0{,}60$. Dimensi akurasi matematis memperoleh nilai tertinggi ($V = 0{,}83$), sedangkan dimensi rubrik dan penskoran terendah ($V = 0{,}63$; aspek rubrik pada tingkat butir $V = 0{,}48$). Ketiga ahli meminta deskriptor kredit parsial yang lebih teramati, dua ahli secara mandiri meminta aturan *follow-through* untuk butir 2, 3, 5, dan 6, dan dua ahli menandai ketidakkonsistenan rumusan "uang tidak boleh kurang" dan "semua uang terpakai

habis" pada Tugas 2. Pemeriksaan keterbacaan lisan terhadap delapan siswa setingkat dari sekolah lain menunjukkan semua butir terbaca, meskipun pemeriksaan ini bukan uji keterbacaan penuh.

Pengumpulan dan Analisis Data

Guru kelas memberikan asesmen sebagai tugas sumatif pada 5 Agustus 2026 selama 80 menit, dengan guru sebagai pengawas tunggal dan tata duduk berpasangan; buku, catatan, telepon genggam, dan kalkulator tidak diperkenankan. Wawancara dilaksanakan sehari kemudian dan ditelusuri baris demi baris di atas lembar siswa. Bantuan diberikan dalam urutan tetap: membaca ulang baris yang relevan, memeriksa dengan operasi kebalikan, menguraikan perhitungan menjadi dua langkah, dan peragaan langsung. Transkrip verbatim memuat 503 giliran bicara pewawancara.

Pekerjaan tertulis dikodekan dengan skema a priori yang diturunkan dari definisi proses PISA (sepuluh kode, Tabel 4) dan diskor 0 sampai 2 untuk setiap proses menurut dua aturan (Tabel 2). Aturan konvensional menilai terhadap kunci, sedangkan aturan bersyarat (*error carried forward*, ECF) menilai *employ* dan *interpret* terhadap keluaran anteseden siswa sendiri. Aturan bersyarat mengoperasionalkan permintaan *follow-through* para ahli sebelum data dikumpulkan. Pada butir 2 kedua aturan memberi skor identik karena dua siswa yang modelnya keliru (S12, S13) tidak menyelesaikan model mereka sendiri, sehingga perbedaan aturan tampak pada butir 3. Pada Tugas 2 aturan bersyarat tidak diterapkan karena tujuh siswa mengosongkan butir 4 sampai 6 dan tujuh dari sepuluh model yang ditulis tidak koheren secara dimensi. Dua peneliti mengodekan seluruh lembar secara konsensual. Kesahihan dijaga melalui triangulasi sumber (lembar jawaban, transkrip, dan dokumen asesmen guru) dan triangulasi teknik (tulisan dan lisan), dengan analisis model interaktif (Miles et al., 2014).

Tabel 2 Rubrik skor proses menurut aturan konvensional dan aturan bersyarat

Proses dan aturan	**Skor 2**	**Skor 1**	**Skor 0**
Formulate (kedua aturan)	Kedua persamaan benar dengan besaran yang dinyatakan	Satu persamaan benar, atau struktur benar dengan satuan tidak konsisten	Kedua persamaan tidak benar atau kosong
Employ, konvensional	Kedua nilai sesuai kunci	Metode tepat dengan kekeliruan perhitungan	Tidak ada upaya bermakna
Employ, bersyarat	Kedua nilai mengikuti secara benar dari model siswa sendiri	Metode tepat dengan kekeliruan perhitungan	Tidak ada upaya bermakna
Interpret, konvensional	Keputusan benar dengan perhitungan lengkap	Keputusan benar tanpa perhitungan lengkap	Keputusan keliru atau tidak ada keputusan
Interpret, bersyarat	Keputusan dan perhitungan sahih terhadap nilai siswa sendiri	Keputusan konsisten dengan nilai sendiri tanpa perhitungan lengkap	Tidak ada keputusan, atau keputusan tidak konsisten dengan nilai sendiri

Catatan. Kedua aturan hanya berbeda pada rujukan penilaian *employ* dan *interpret*, yaitu kunci resmi atau keluaran anteseden siswa sendiri.

Satu kondisi pelaksanaan membatasi klaim. S3 mengaku melihat pekerjaan teman sebangku, dan perbandingan sistematis menemukan tiga klaster jawaban identik pada sepuluh lembar: klaster A (S6, S7, S12, S13, S14) dan klaster B (S4, S5, S15) pada kalimat jawaban butir 3, serta klaster C (S3, S8) pada persamaan butir 4. Karena itu, klaim tentang mekanisme berpikir disandarkan pada wawancara dan tujuh lembar di luar klaster (S1, S2, S9, S10, S11, S16, S17), sedangkan frekuensi diperlakukan sebagai gambaran kelas. Kepekaan hasil diuji dengan tiga perlakuan: semua lembar dihitung, setiap klaster dihitung satu respons, dan anggota klaster dikeluarkan.

HASIL DAN DISKUSI

Tabel 3 merangkum distribusi skor. Tugas 1 memperlihatkan pola yang lazim: *formulate* tinggi, *employ* menengah, dan *interpret* rendah. Pada Tugas 2 seluruh proses runtuh dan urutannya terbalik (*employ* 0,29; *interpret* 0,24; *formulate* 0,18). Empat belas siswa memperoleh nol pada seluruh Tugas 2, sedangkan tidak seorang pun memperoleh nol pada seluruh Tugas 1.

Tabel 3 Distribusi dan rerata skor proses (n = 17, skor maksimum 2)

Tugas	Butir dan proses	Skor 0	Skor 1	Skor 2	Rerata
1	Butir 1, *formulate*	2	0	15	1,76
1	Butir 2, *employ*	0	14	3	1,18
1	Butir 3, *interpret* (konvensional)	14	0	3	0,35
1	Butir 3, *interpret* (bersyarat)	6	5	6	1,00
2	Butir 4, *formulate*	15	1	1	0,18
2	Butir 5, *employ*	14	1	2	0,29
2	Butir 6, *interpret*	15	0	2	0,24

Catatan. Skor Tugas 2 memakai aturan konvensional (lihat bagian Metode).

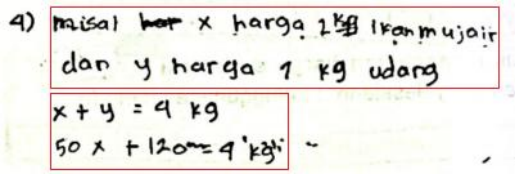
Formulate: Keruntuhan pada Butir Berperancah yang Berbeda Jenis Besaran

Lima belas dari 17 siswa menyusun kedua persamaan butir 1 dengan benar. Dua kesalahan berupa salah salin kuantitas: S12 menulis $2x + 4y = 310.000$ dan $4x + 2y = 340.000$, sedangkan S13 menulis $2x + 2y$ pada kedua baris. Keberhasilan ini tidak sekadar di permukaan, karena setiap siswa yang diwawancarai menjelaskan tanpa diminta bahwa $2x$ menyatakan harga dua noken dan 310.000 adalah total belanja.

Pada butir 4, tujuh siswa mengosongkan jawaban, dan dari sepuluh yang mengerjakan hanya S10 yang menyusun model benar. Kegagalan lainnya spesifik. S4 dan S11 mengganti definisi berat dengan harga, bahkan S11 menulis $x = 50.000$ dan $y = 120.000$, yaitu konstanta yang sudah diketahui. S7, S8, dan S13 menulis definisi "y = berat 1 kg udang" sehingga konstanta melekat pada simbol yang seharusnya berubah. S6 dan S12 mencampur jenis besaran dalam satu sistem, dan tujuh lembar memuat persamaan tak berdimensi, seperti $50x + 120y = 4$ kg atau $50.000x + 120.000y = 170.000$. S4 menjelaskan mekanismenya sendiri:

"Karena di soalnya banyak angka harga per kilogram. Saya langsung berpikir x dan y itu harga. Tetapi setelah saya menulis x + y = 4 kg, sebenarnya yang saya pikirkan berat." (S4)

Lembar S4 memuat kedua model sekaligus (Gambar 1). Baris definisi memberi harga kepada variabel, sedangkan persamaan $x + y = 4$ kg di bawahnya hanya koheren bila variabel menyatakan berat. Definisi dalam soal terbaca, tetapi tidak mengatur tulisan siswa.



Gambar 1 Jawaban S4 pada butir 4. Baris definisi memberi harga kepada x dan y, sedangkan persamaan $x + y = 4$ kg hanya koheren bila x dan y menyatakan berat.

Employ: Galat Aritmetika, Nilai Bersama, dan Absennya Verifikasi

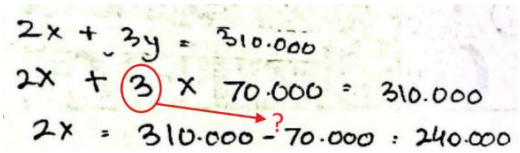
Seluruh siswa memilih metode eliminasi atau substitusi yang tepat pada butir 2, dan tidak ada yang berskor nol. Yang gagal adalah aritmetika bilangan ratusan ribu. Dua belas lembar memuat 930.000 − 680.000 = 350.000, yaitu galat pada langkah meminjam (93 − 68 dihitung 35), yang menghasilkan y = 70.000. Siswa yang gagal pada bentuk satu langkah langsung berhasil begitu pewawancara menguraikannya menjadi 930.000 − 600.000 lalu 330.000 − 80.000. S2, yang berhenti setelah memperoleh nilai keliru, menyelesaikan sistem berstruktur identik dengan bilangan kecil tanpa bantuan.

Empat belas siswa sampai pada y = 70.000, tetapi angka ini tidak dapat dibaca sebagai empat belas peristiwa yang saling bebas. Galat pengurangan yang sama muncul pada empat dari tujuh lembar di luar klaster (S2, S9, S16, S17), sehingga galat itu dapat timbul secara mandiri. Sebaliknya, dua lembar sampai pada 70.000 dengan baris yang tidak dapat diturunkan dari pekerjaan siswa sendiri: S5 salah mengalikan (340.000 × 2 = 660.000) tetapi menulis $5y$ = 350.000, dan S6 menulis 250.000 yang benar di margin tetapi tetap menulis 70.000. Keduanya anggota klaster, sehingga pengambilalihan nilai dari meja sebelah adalah penjelasan yang paling hemat. Wawancara S5 menunjukkan mengapa nilai semacam itu diterima tanpa diperiksa:

"Seperti mengikuti langkah saja, setelah dikurangi harus dapat angka yang bisa dibagi 5. Saya dapat 350.000, dan karena bisa dibagi 5, saya lanjutkan." (S5)

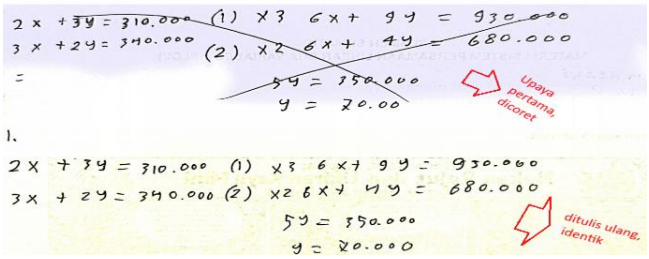
Keterbagian dipakai sebagai pengganti kebenaran, sementara verifikasi hampir tidak ada. Hanya S1 yang mensubstitusikan hasilnya ke kedua persamaan; seorang siswa memeriksa satu persamaan saja, dan dua siswa yang hasilnya benar (S10, S11) tidak memeriksa apa pun. Masalah kedua muncul pada substitusi. Delapan dari 12 siswa yang mencapai tahap itu mengganti suku $3y$ dengan 70.000, bukan 3 × 70.000. Lembar S14 memperlihatkan bentuk paling jelas (Gambar 2): $2x + 3 \times 70.000 = 310.000$ ditulis benar, lalu baris berikutnya menghitung $2x = 310.000 − 70.000$. Keempat pelaku yang diwawancarai (S4 sampai S7) menjelaskan dengan benar bahwa $3y$ berarti harga tiga ukiran, sehingga kesalahan ini tidak bersumber dari ketidaktahuan makna suku. Namun, kedelapan lembar itu

seluruhnya berada dalam klaster A dan B, sehingga frekuensinya menggambarkan sebaran di kelas (Tabel 4).



Gambar 2 Substitusi S14 pada butir 2. Koefisien 3 ditulis benar pada baris kedua, lalu hilang dari perhitungan pada baris ketiga (3 × 70.000 digantikan 70.000).

Pada Tugas 2, keberhasilan informal muncul tanpa model formal yang benar. S4, yang model butir 4-nya tak berdimensi, menjawab butir 5 dan 6 dengan benar melalui aritmetika: 2 kg dan 2 kg karena 100.000 + 240.000 = 340.000, lalu 3 kg dan 1 kg dengan sisa Rp70.000,00. S1 melakukan hal serupa dan merumuskan sendiri kedua kendala: "beratnya harus 4 kg dan uangnya harus 340.000". Sebaliknya, S10, satu-satunya pemilik model benar, tergelincir saat mengumpulkan suku sejenis. Tiga siswa juga menyadari ada yang keliru dan mengulang pekerjaan dengan hasil berbeda: S11 sampai pada jawaban benar, S7 mengulang prosedur dan kekeliruan yang sama, dan S17 menyalin ulang seluruh pekerjaannya lengkap dengan kekeliruannya (Gambar 3). S7 menyebut batasnya: "Saya merasa ada yang salah ... tetapi saya tidak tahu cara memeriksa bagian mana yang salah." Gejala ini peneliti sebut deteksi tanpa lokalisasi: siswa merasakan adanya galat tetapi tidak memiliki prosedur untuk menemukan letaknya.



Gambar 3 Pekerjaan S17 pada butir 2 yang ditulis dua kali. Blok kedua mengulang blok pertama baris demi baris, termasuk 930.000 − 680.000 = 350.000 dan y = 70.000.

Tabel 4 Kode kesalahan dan frekuensi

Kode	Deskripsi	Frekuensi	Penyebut
C1	Salah salin sistem persamaan pada butir 1	2	17 siswa
C2	Galat pengurangan bilangan besar (930.000 − 680.000 = 350.000)	12	17 siswa
C3	Galat perkalian bilangan besar (340.000 × 2 = 660.000)	2	17 siswa
C4	Galat pembagian bilangan besar (250.000 : 5 = 70.000)	1	17 siswa
C5	Koefisien diabaikan saat substitusi (3y diganti nilai y saja)	8	12 yang mensubstitusi
C6	Ruas kedua persamaan tertukar	3	17 siswa
C7	Tidak ada verifikasi, atau verifikasi pada satu persamaan saja	16	17 siswa

Kode	Deskripsi	Frekuensi	Penyebut
C8	Galat rujukan pada butir 4 (harga untuk berat, atau konstanta melekat pada variabel)	8	10 yang mengerjakan
C9	Galat dimensi pada persamaan biaya (ekspresi rupiah = 4 kg, atau harga satuan dijumlahkan)	7	10 yang mengerjakan
C10	Langkah tidak terturunkan dari baris sebelumnya	2	17 siswa

Catatan. Kode tidak saling eksklusif. Penyebut berbeda karena tujuh siswa mengosongkan butir 4 sampai 6 dan lima siswa berhenti sebelum substitusi. C2 dan C7 juga muncul pada lembar di luar klaster, sedangkan seluruh lembar C5 dan C9 berada dalam klaster.

Interpret: Dua Aturan Skor, Dua Diagnosis

Dengan aturan konvensional, butir 3 menghasilkan rerata 0,35: tiga siswa (S1, S10, S11) berskor 2 dan empat belas berskor 0. Dibaca apa adanya, angka ini menunjukkan kemampuan menafsirkan yang sangat lemah, tetapi pekerjaan tertulis memberi gambaran lain. Klaster B (S4, S5, S15) menyimpulkan bahwa Rp200.000,00 tidak cukup dengan kekurangan Rp60.000,00. Kesimpulan itu keliru terhadap kunci, tetapi tepat terhadap nilai mereka sendiri: dengan noken Rp120.000,00 dan ukiran Rp70.000,00, satu noken dan dua ukiran berharga Rp260.000,00. Mereka mengalikan kuantitas dengan harga satuan, membandingkan total dengan uang, dan menghitung selisihnya; yang cacat adalah masukan dari tahap *employ*. S4 mengenalinya ketika diminta memeriksa: "Cara menghitung totalnya sudah benar, tetapi harga yang saya pakai salah." Klaster A menulis kalimat identik bahwa uangnya kurang, tanpa satu angka pun.

Dengan aturan bersyarat, klaster B memperoleh skor 2 dan klaster A skor 1, sehingga rerata naik menjadi 1,00 dan siswa yang diakui memiliki kemampuan menafsirkan bertambah dari 3 menjadi 11 (Tabel 3). Tidak ada skor yang turun. Aturan ini tidak melonggarkan standar, tetapi memindahkan letak kegagalan dari *interpret* ke *employ*, sesuai dengan kekhawatiran para ahli yang meminta aturan *follow-through* sebelum data dikumpulkan.

Kepekaan Temuan terhadap Respons Berklaster

Seluruh kenaikan skor bersyarat berasal dari klaster A dan B, sehingga besarnya bergantung pada cara klaster diperlakukan (Tabel 5). Bila setiap klaster dihitung satu respons, rerata bersyarat menjadi 0,82 dibanding 0,55 secara konvensional. Bila anggota kedua klaster dikeluarkan, kedua aturan memberi rerata yang sama (0,67), karena enam siswa lain yang nilainya keliru tidak menghasilkan keputusan yang dapat dinilai ulang. Sebaliknya, kontras *formulate* bertahan: ketujuh siswa di luar klaster menyusun model butir 1 dengan benar, tetapi hanya S10 yang berhasil pada butir 4. Jadi, keruntuhan *formulate*, galat pengurangan, dan absennya verifikasi tidak bergantung pada penyalinan, sedangkan besarnya efek penskoran bersyarat bergantung padanya. Yang bertahan dari temuan *interpret* adalah arah diagnosisnya, yang dikonfirmasi oleh wawancara S4.

Tabel 5 Kepekaan hasil *interpret* butir 3 terhadap perlakuan respons berklaster

Perlakuan	n	Rerata konvensional	Rerata bersyarat	Siswa berkredit
Semua lembar dihitung	17	0,35	1,00	3 → 11
Klaster A dan B masing-masing dihitung satu respons	11	0,55	0,82	3 → 5
Anggota klaster A dan B dikeluarkan	9	0,67	0,67	3 → 3

Catatan. Siswa berkredit adalah siswa dengan skor *interpret* minimal 1 (konvensional → bersyarat). Klaster C tidak memuat jawaban butir 3 yang identik.

Perbandingan dengan Penelitian Terdahulu

Temuan *formulate* menegaskan bahwa hambatan utama terletak pada peralihan situasi ke model (Galbraith & Stillman, 2006; Schukajlow et al., 2023; Wijaya et al., 2014), tetapi menambahkan satu kondisi. Ketika besaran yang dicari sejenis dengan data, siswa yang sama berhasil, sejalan dengan laporan bahwa *formulate* bukan proses terlemah siswa Indonesia (Hamidy & Jailani, 2019). Ketika jenisnya berbeda, definisi tertulis kalah oleh tampilan halaman yang dipenuhi harga. Pola ini konsisten dengan kategorisasi berdasarkan ciri permukaan (Chi et al., 1981; Vessonen et al., 2024) yang bekerja pada titik penerjemahan, serupa dengan kesalahan pembalikan yang dipicu urutan kata alih-alih makna besaran (Soneira et al., 2021). Definisi "y = berat 1 kg udang" juga mirip pembacaan huruf sebagai label objek (Klingbeil & Moons, 2025), dengan perbedaan penting: di sini makna huruf telah ditetapkan soal, dan kecenderungan itu tetap bertahan.

Pada *employ*, galat 350.000 merupakan galat prosedural sistematis yang dapat muncul secara mandiri pada banyak siswa (Brown & Burton, 1978), dan keberhasilan setelah perhitungan diuraikan sejalan dengan peran memindahkan hasil antara ke atas kertas untuk meringankan memori kerja (Medrano & Miller-Cotto, 2025). Absennya verifikasi menguatkan laporan bahwa siswa jarang memeriksa kembali jawaban SPLDV (Pradana & Murtiyasa, 2020), padahal substitusi ke kedua persamaan adalah prosedur yang dapat melokalisasi kekeliruan (Schoenfeld, 2016). Koefisien yang ditulis tetapi tidak dihitung pada lembar S14 sesuai dengan pemrosesan simbol yang visual alih-alih semantik (Kirshner, 1989), meskipun penyalinan tidak dapat disingkirkan sebagai penjelasan alternatif. Keberhasilan informal S1 dan S4 mendukung pandangan bahwa penalaran aritmetika tetap membentuk capaian aljabar simbolik (Filloy & Rojano, 1989; Pitta-Pantazi et al., 2025).

Pada *interpret*, penskoran konvensional mereproduksi pola yang dilaporkan penelitian sebelumnya, yaitu *interpret* sebagai proses terlemah (Gustiningsi et al., 2026; Hamidy & Jailani, 2019). Penskoran bersyarat menunjukkan bahwa sebagian kelemahan itu dapat merupakan rambatan dari *employ*, sejalan dengan premis kerangka diagnostik yang mencari letak kegagalan dalam penyelesaian bertingkat (Säfström et al., 2024; Sidenvall et al., 2026). Berbeda dari May et al. (2023), kontras yang diuji di sini menyangkut rujukan penilaian, bukan kehalusan kredit. Analisis kepekaan

menambahkan peringatan yang belum dibahas literatur: karena butir berantai saling bergantung (Eaton et al., 2020), respons yang disalin dapat menggandakan efek aturan bersyarat.

Kontribusi Teoretis dan Implikasi Praktis

Secara teoretis, penelitian ini mengusulkan interferensi jenis besaran sebagai mode kegagalan konvensi definisi variabel: definisi yang dinyatakan dapat terbaca tanpa mengatur pemakaian variabel ketika jenis besaran yang dicari berbeda dari bilangan yang menonjol. Konsekuensinya, butir yang menyediakan definisi variabel mengukur penerjemahan berbantuan, bukan formulasi. Kajian sistematis tentang karakteristik tugas pemodelan belum memuat jenis besaran (Krawitz et al., 2025), dan temuan ini menunjukkan mengapa karakteristik itu perlu diperhitungkan.

Secara praktis, ada tiga implikasi. Pertama, instrumen yang dimaksudkan mengukur *formulate* perlu memuat sedikitnya satu butir yang jenis besaran yang dicarinya berbeda dari bilangan dominan, dan hasilnya dilaporkan terpisah; pasangan butir 1 dan 4 dapat dijadikan contoh. Kedua, skor konvensional dan bersyarat pada butir berantai sebaiknya dilaporkan berdampingan dengan rubrik pada Tabel 2, sejalan dengan rubrik yang menskor fase pemodelan secara terpisah (Toalongo et al., 2022), disertai pemeriksaan independensi respons dan pengawasan yang mencegah penyalinan. Ketiga, substitusi perlu diajarkan sebagai penggantian nilai di dalam suku dengan koefisien dinyatakan sebagai operasi, kesesuaian antara pernyataan verbal dan simbol perlu dijembatani secara eksplisit (Kusmaryono et al., 2024), dan verifikasi pada setiap persamaan perlu diwajibkan karena hanya verifikasi yang memberi siswa kemampuan melokalisasi galat.

Kelebihan dan Keterbatasan Penelitian

Kelebihan penelitian ini adalah sensus satu kelas, triangulasi lembar jawaban dengan wawancara dinamis, aturan bersyarat yang ditetapkan sebelum data dikumpulkan, serta pengungkapan terbuka atas penyalinan beserta analisis kepekaannya. Keterbatasannya ada enam. Pertama, data berasal dari 17 siswa, satu sekolah, dan satu topik, sehingga tidak ada klaim tentang literasi matematika di Papua secara umum. Kedua, kontras butir 1 dan 4 bukan perbandingan terkendali: butir 4 muncul belakangan sehingga kelelahan tidak dapat disingkirkan, stimulusnya prosa, sistemnya lebih kompleks, dan rumusan Tugas 2 tidak konsisten menurut dua ahli. Ketiga, sepuluh lembar berada dalam klaster jawaban identik, kemungkinan besar akibat tata duduk berpasangan dengan satu pengawas; analisis kepekaan membatasi dampaknya tanpa meniadakannya. Keempat, pengodean bersifat konsensual tanpa statistik kesepakatan antarkoder, sehingga Tabel 4 adalah catatan deskriptif yang disepakati. Kelima, wawancara bersifat retrospektif; jeda satu hari membatasi rasionalisasi tanpa meniadakannya (Ericsson & Simon, 1980), sehingga klaim menyangkut cara siswa menjelaskan jejak tulisannya, bukan kognisi saat asesmen berlangsung. Keenam, butir 3 mengoperasionalkan *interpret* secara sempit sebagai penerapan hasil pada keputusan, bukan evaluasi kewajaran hasil, sebagaimana tercermin pada koefisien Aiken aspek proses butir 3 dan 6 ($V = 0{,}67$).

KESIMPULAN

Penelitian ini menjawab tiga pertanyaan. Pertama, pada dua butir yang sama-sama menyatakan definisi variabel, rerata *formulate* turun dari 1,76 menjadi 0,18 ketika besaran yang dicari berubah dari harga menjadi berat di tengah data harga; hanya satu dari sepuluh siswa yang mengerjakan butir 4 menyusun model benar, dan kontras ini bertahan pada lembar di luar klaster (tujuh dari tujuh benar pada butir 1, satu dari tujuh pada butir 4). Definisi variabel yang dinyatakan dalam soal karena itu mengukur penerjemahan berbantuan, bukan formulasi. Kedua, hambatan *employ* bersifat aritmetika, terutama galat meminjam pada bilangan ratusan ribu (12 dari 17 lembar), diperparah oleh penerimaan bilangan yang habis dibagi sebagai tanda benar dan oleh verifikasi yang hanya dilakukan satu siswa. Ketiga, penskoran bersyarat menaikkan rerata *interpret* butir 3 dari 0,35 menjadi 1,00 dan memindahkan letak kegagalan ke *employ*, tetapi besarnya kenaikan bergantung pada respons berklaster: 0,82 bila klaster dihitung satu respons dan tidak berbeda bila anggota klaster dikeluarkan.

Sejalan dengan keterbatasan tersebut, penelitian lanjutan disarankan untuk (1) menguji jenis besaran secara terkendali dengan urutan butir diseimbangkan serta modalitas stimulus, rumusan, dan kompleksitas dipertahankan; (2) melaksanakan asesmen dengan tempat duduk terpisah dan lebih dari satu pengawas agar respons independen; (3) melibatkan koder independen dan melaporkan statistik kesepakatan; (4) menerapkan pelaporan ganda pada sampel yang lebih besar dan beragam di Papua; dan (5) menguji apakah pelaporan ganda benar-benar mengubah keputusan remediasi guru.

UCAPAN TERIMA KASIH

Peneliti berterima kasih kepada kepala madrasah, guru matematika, dan siswa MTs Negeri Kota Jayapura, serta kepada tiga penelaah ahli. Penelitian ini didanai PNBP BLU Universitas Cenderawasih Tahun 2026 melalui Surat Perjanjian Kerja Penelitian Nomor 116/UN20.2.1/PG/2026. Peneliti menyatakan tidak ada konflik kepentingan. Instrumen dan lembar pengodean anonim dapat diperoleh dari penulis korespondensi atas permintaan yang wajar.

REFERENSI

Aiken, L. R. (1985). Three Coefficients for Analyzing the Reliability and Validity of Ratings. *Educational and Psychological Measurement*, *45*(1), 131–142. https://doi.org/10.1177/0013164485451012

Brown, J. S., & Burton, R. R. (1978). Diagnostic Models for Procedural Bugs in Basic Mathematical Skills*. *Cognitive Science*, *2*(2), 155–192. https://doi.org/10.1207/s15516709cog0202_4

Castro, E., Cañadas, M. C., Molina, M., & Rodríguez-Domingo, S. (2022). Difficulties in semantically congruent translation of verbally and symbolically represented algebraic statements. *Educational Studies in Mathematics*, *109*(3), 593–609. https://doi.org/10.1007/s10649-021-10088-3

Cevikbas, M., Kaiser, G., & Schukajlow, S. (2022). A systematic literature review of the current discussion on mathematical modelling competencies: state-of-the-art developments in conceptualizing, measuring, and fostering. *Educational Studies in Mathematics*, *109*(2), 205–236. https://doi.org/10.1007/s10649-021-10104-6

Chi, M. T. H., Feltovich, P. J., & Glaser, R. (1981). Categorization and Representation of Physics Problems by Experts and Novices*. *Cognitive Science*, *5*(2), 121–152. https://doi.org/10.1207/s15516709cog0502_2

Darmawijoyo, D., Zulkardi, Z., Putri, R. I. I., Hapizah, H., & Syutaridho, S. (2025). How do students use mathematical reasoning to solve PISA-type mathematics problems based on making kite contexts? *Infinity Journal*, *14*(4), 1065–1080. https://doi.org/10.22460/infinity.v14i4.p1065-1080

Eaton, P., Frank, B., & Willoughby, S. (2020). Detecting the influence of item chaining on student responses to the Force Concept Inventory and the Force and Motion Conceptual Evaluation. *Physical Review Physics Education Research*, *16*(2), 020122. https://doi.org/10.1103/PhysRevPhysEducRes.16.020122

Efendi, R., Diana, Y., & Panjaitan, A. T. (2025). Analisis Disposisi Matematis, Kecemasan Matematika, dan Numerasi Siswa di Daerah 3T: Perbatasan RI-PNG. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *9*(2), 498–511. https://doi.org/10.31004/cendekia.v9i2.3967

Efendi, R., & Kulimbang, E. (2025). Pengaruh Self-Regulated Learning dan Gaya Belajar terhadap Kemampuan Pemecahan Masalah Matematika Siswa MTs Negeri Kota Jayapura. *Jurnal Cendekia : Jurnal Pendidikan Matematika*, *10*(1), 25–39. https://doi.org/10.31004/cendekia.v10i1.4601

Elliott, J. G., Resing, W. C. M., & Beckmann, J. F. (2018). Dynamic assessment: a case of unfulfilled potential? *Educational Review*, *70*(1), 7–17. https://doi.org/10.1080/00131911.2018.1396806

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review*, *87*(3), 215–251. https://doi.org/10.1037/0033-295X.87.3.215

Filloy, E., & Rojano, T. (1989). Solving Equations: the Transition from Arithmetic to Algebra. *For the Learning of Mathematics*, *9*(2), 19–25. https://www.jstor.org/stable/40247950

Galbraith, P., & Stillman, G. (2006). A framework for identifying student blockages during transitions in the modelling process. *ZDM*, *38*(2), 143–162. https://doi.org/10.1007/BF02655886

Gustiningsi, T., Pasaribu, F. T., Junita, R., & Ismail, Z. (2026). Investigating students' mathematical literacy: How students solve PISA-like problems in the Jambi context. *Jurnal Elemen*, *12*(2), 631–651. https://doi.org/10.29408/jel.v12i2.34156

Hamidy, A., & Jailani, J. (2019). Kemampuan proses matematis siswa Kalimantan Timur dalam menyelesaikan soal matematika model PISA. *Jurnal Riset Pendidikan Matematika*, *6*(2), 133–149. https://doi.org/10.21831/jrpm.v6i2.26679

Khairani, A., Hakiki, N. I., & Yopan, M. (2025). Analisis Faktor Penyebab Angka Putus Sekolah di Tingkat SD dan SMP di Papua: Kajian Studi Pustaka Berbasis Teori Struktural. *JIIP - Jurnal Ilmiah Ilmu Pendidikan*, *8*(1), 45–51. https://doi.org/10.54371/jiip.v8i1.6510

Kirshner, D. (1989). The Visual Syntax of Algebra. *Journal for Research in Mathematics Education*, *20*(3), 274. https://doi.org/10.2307/749516

Klingbeil, K., & Moons, F. (2025). The omnipresence of the Letter-as-Object misconception: a multilevel Latent Transition Analysis on understanding variables. *Educational Studies in Mathematics*, *120*(3), 455–476. https://doi.org/10.1007/s10649-025-10428-7

Krawitz, J., Chang, Y.-P., Yang, K.-L., & Schukajlow, S. (2022). The role of reading comprehension in mathematical modelling: improving the construction of a real-world model and interest in Germany and Taiwan. *Educational Studies in Mathematics*, *109*(2), 337–359. https://doi.org/10.1007/s10649-021-10058-9

Krawitz, J., Schukajlow, S., Yang, X., & Geiger, V. (2025). A Systematic Review of International Perspectives on Mathematical Modelling: Modelling Goals and Task Characteristics. *ZDM – Mathematics Education*, *57*(2–3), 193–212. https://doi.org/10.1007/s11858-025-01683-2

Kusmaryono, I., Aminudin, M., Ubaidah, N., & Chamalah, E. (2024). The bridging understanding of language and mathematical symbols between teachers and students: An effort to increase mathematical literacy. *Infinity Journal*, *13*(1), 251–270. https://doi.org/10.22460/infinity.v13i1.p251-270

May, T. A., Koskey, K. L. K., Bostic, J. D., Stone, G. E., Kruse, L. M., & Matney, G. (2023). Examining how using dichotomous and partial credit scoring models influence sixth-grade mathematical problem-solving assessment outcomes. *School Science and Mathematics*, *123*(2), 54–67. https://doi.org/10.1111/ssm.12570

Medrano, J., & Miller-Cotto, D. (2025). Understanding working memory as a facilitator of math problem-solving: Offloading as a potential strategy. *British Journal of Educational Psychology*, *95*(3), 871–887. https://doi.org/10.1111/bjep.12767

Miles, M. B., Huberman, A. M., & Saldana, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook* (3rd ed.). SAGE Publications, Inc.

Muskens, M., Frankenhuis, W. E., & Borghans, L. (2024). Math items about real-world content lower test-scores of students from families with low socioeconomic status. *Npj Science of Learning*, *9*(1), 19. https://doi.org/10.1038/s41539-024-00228-8

OECD. (2023a). *PISA 2022 Assessment and Analytical Framework*. OECD Publishing. https://doi.org/10.1787/dfe0bf9c-en

OECD. (2023b). *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*. OECD Publishing. https://doi.org/10.1787/53f23881-en

OECD. (2026). *PISA 2025 Results (Volume I): Future-Ready Students*. OECD Publishing. https://doi.org/10.1787/73451bc5-en

Pitta-Pantazi, D., Chimoni, M., & Christou, C. (2025). Revisiting the Relationship of Arithmetical Thinking and Letter-Symbolic Algebra. *International Journal of Science and Mathematics Education*, *23*(3), 689–711. https://doi.org/10.1007/s10763-024-10493-z

Pradana, D. A. Y., & Murtiyasa, B. (2020). Kemampuan siswa menyelesaikan masalah berbentuk soal cerita sistem persamaan linear ditinjau dari kemampuan penalaran. *Pythagoras: Jurnal Pendidikan Matematika*, *15*(2). https://doi.org/10.21831/pg.v15i2.35419

Prahmana, R. C. I., & D'Ambrosio, U. (2020). Learning Geometry and Values from Patterns: Ethnomathematics on the Batik Patterns of Yogyakarta, Indonesia. *Journal on Mathematics Education*, *11*(3), 439–456. https://doi.org/10.22342/jme.11.3.12949.439-456

Rubel, L. H., & McCloskey, A. V. (2021). Contextualization of mathematics: which and whose world? *Educational Studies in Mathematics*, *107*(2), 383–404. https://doi.org/10.1007/s10649-021-10041-4

Säfström, A. I., Lithner, J., Palm, T., Palmberg, B., Sidenvall, J., Andersson, C., Boström, E., & Granberg, C. (2024). Developing a diagnostic framework for primary and secondary students' reasoning difficulties during mathematical problem solving. *Educational Studies in Mathematics*, *115*(2), 125–149. https://doi.org/10.1007/s10649-023-10278-1

Schoenfeld, A. H. (2016). Learning to Think Mathematically: Problem Solving, Metacognition, and Sense Making in Mathematics (Reprint). *Journal of Education*, *196*(2), 1–38. https://doi.org/10.1177/002205741619600202

Schukajlow, S., Krawitz, J., Kanefke, J., Blum, W., & Rakoczy, K. (2023). Open modelling problems: cognitive barriers and instructional prompts. *Educational Studies in Mathematics*, *114*(3), 417–438. https://doi.org/10.1007/s10649-023-10265-6

Sidenvall, J., Boström, E., Granberg, C., Lithner, J., Säfström, A. I., Andersson, C., Palmberg, B., & Palm, T. (2026). Teachers' in-the-moment diagnosis of students' problem-solving difficulties in mathematics. *Frontiers in Education*, *11*. https://doi.org/10.3389/feduc.2026.1831773

Soneira, C., Bansilal, S., & Govender, R. (2021). Insights into the reversal error from a study with South African and Spanish prospective primary teachers. *Pythagoras*, *42*(1). https://doi.org/10.4102/pythagoras.v42i1.613

Stefana, A., Damiani, S., Granziol, U., Provenzani, U., Solmi, M., Youngstrom, E. A., & Fusar-Poli, P. (2025). Psychological, psychiatric, and behavioral sciences measurement scales: best practice guidelines for their development and validation. *Frontiers in Psychology*, *15*. https://doi.org/10.3389/fpsyg.2024.1494261

Strohmaier, A. R., Schons, C., Knabbe, A., Vogel, M., Saado, K., & Obersteiner, A. (2026). The relation between text comprehension and computation in mathematical word-problem solving. *Learning and Instruction*, *101*, 102235. https://doi.org/10.1016/j.learninstruc.2025.102235

Toalongo, X., Trelles, C., & Alsina, Á. (2022). Design, Construction and Validation of a Rubric to Evaluate Mathematical Modelling in School Education. *Mathematics*, *10*(24), 4662. https://doi.org/10.3390/math10244662

Vessonen, T., Dahlberg, M., Hellstrand, H., Widlund, A., Korhonen, J., Aunio, P., & Laine, A. (2024). Task Characteristics Associated with Mathematical Word Problem-Solving Performance Among Elementary School-Aged Children: A Systematic Review and Meta-Analysis. *Educational Psychology Review*, *36*(4), 117. https://doi.org/10.1007/s10648-024-09954-2

Wahyuni, R., Suwanto, F. R., Sthephani, A., & Ahyan, S. (2025). Students' obstacles in solving algebra form problems viewed from mathematical problem-solving ability. *Infinity Journal*, *14*(3), 587–606. https://doi.org/10.22460/infinity.v14i3.p587-606

Wijaya, A., van den Heuvel-Panhuizen, M., Doorman, M., & Robitzsch, A. (2014). Difficulties in solving context-based PISA mathematics tasks: An analysis of students' errors. *The Mathematics Enthusiast*, *11*(3), 555–584. https://doi.org/10.54870/1551-3440.1317