

When the Best Studies Show the Least: A Systematic Review and Typological Synthesis of Differentiated Instruction for Diverse English Learners in Inclusive Education

Mary Anne C. Sedanza

College of Education, Leyte Normal University, Tacloban City, Philippines

maryanne.sedanza@lnu.edu.ph

HOW TO CITE:

Mary Anne C. Sedanza (2026).
When the Best Studies Show the
Least: A Systematic Review and
Typological Synthesis of
Differentiated Instruction for
Diverse English Learners in
Inclusive Education.
International Journal of Special
Education, 41(8s), 559-570

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the
terms and conditions
of the Creative Commons Attribution
(CC BY)
license (<http://creativecommons.org/licenses/by/4.0/>).

ABSTRACT:

Differentiated instruction is widely promoted as an inclusive response for diverse English learners, including those with special educational needs, yet syntheses rarely classify the evidence by how differentiation is operationalized, for whom, and under what design conditions. This study reports a systematic review with a configurative, typological synthesis of 56 controlled effect sizes drawn from 45 studies across 18 countries (2011–2024; 16,773 learners), following PRISMA 2020. A six-facet classification: learner/exceptionality, differentiation dimension, differentiation format, outcome domain, design rigor, and context, was developed and applied to every effect size. The map reveals a field that is broad but uneven: process and content adjustments dominate (55 and 41 of 56), while product and the learning environment are rarely addressed (20 and 13); reading accounts for 32 of 56 outcomes, while productive skills are severely underrepresented. Crossing the typology with design rigor exposes an inverse-rigor gradient: all 13 very-large effects ($g \geq 1.20$) come from quasi-experimental studies (mean $g = 1.03$), whereas the 13 randomized effects average near zero (mean $g = 0.10$), with the null and negative results concentrated among young, struggling learners in well-resourced settings. Classifying, rather than averaging, reframes differentiation as a conditional, not a universal, good, and yields an evidence gap map for inclusive practice and future rigorous trials involving learners with identified support needs.

Keywords: differentiated instruction; special educational needs; inclusive education; systematic review; typology; risk of bias; educational equity.

1. Introduction

Few ideas in contemporary inclusive education are as widely endorsed as differentiated instruction. The premise is intuitive and humane:

because no two learners arrive with the same readiness, language, or history, a single lesson built for an imagined average student fits few learners well, especially learners with disabilities, additional-language needs, giftedness, or

intersecting support needs—the very learners whom inclusive schooling exists to serve. On that premise, differentiation has moved from aspiration to expectation, written into teacher preparation, professional guidance, and standards in many school systems (Tomlinson, 2017; Gibbs & McKay, 2021).

For a practice this widely required, the evidence is surprisingly hard to summarize. Reviews report that differentiation can raise achievement, but they typically group heterogeneous practices under a single label and report a single average effect (Deunk et al., 2018; Hu, 2024; Asriadi et al., 2023; Weganofa & Mujiono, 2025). An average, however, is only useful when the things being averaged are alike. When the same word covers a teacher regrouping a class by reading level, a software package adjusting item difficulty, and a tiered set of writing tasks—delivered to gifted adolescents in one study and struggling first-graders in the next—the mean tells a teacher very little about what to do on Monday. The problem is not that studies are lacking; it is that the field has not yet organized them.

This review takes organization as its task. Rather than pooling the literature into one number, this review classifies it. A companion meta-analysis of the same corpus reported the overall effect and the contextual conditions that move it (Sedanza & Abatayo, 2026, unpublished); the present review asks a prior and different question—what, exactly, has the field studied, and for whom? Configurative reviews of this kind, which arrange rather than aggregate evidence, are well suited to fields where a construct is contested and the practices it names are diverse (Gough et al., 2012). The output is not a verdict on whether differentiation “works,” but a map: a typology of the evidence and a picture of where it is concentrated and where it runs out. This paper does not re-estimate the overall effectiveness of differentiated instruction; it reuses the eligible corpus to classify the evidence base and to identify coverage gaps of particular relevance to inclusive and special education.

Three questions guide the synthesis. First, how can the controlled evidence on differentiated instruction be classified into a coherent, replicable typology? Second, what does that typology reveal about coverage—which learners, practices, and outcomes are well studied, and which are neglected? Third, what happens when the typology is crossed with the methodological quality of the studies themselves? The third question matters more than it may appear. A long line of work in educational measurement shows that design features shape reported effects, with quasi-experiments yielding systematically larger estimates than randomized trials (Cheung & Slavin, 2016). If that pattern holds within the differentiation literature, then a map that ignores rigor will mislead exactly where the stakes for exceptional learners are highest.

2. Method

2.1 Design and reporting

This review was conducted as a systematic review with a configurative, typological synthesis and is reported in line with the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA 2020; Page et al., 2021). In a configurative synthesis, the analytic goal is to arrange studies into categories that illuminate the structure of a field, rather than to derive a single pooled estimate (Gough et al., 2012). The protocol, coding scheme, and dataset are archived at the Open Science Framework (<https://osf.io/c2v8f>).

2.2 Corpus and eligibility

The synthesis draws on the same corpus assembled for the companion meta-analysis of differentiated instruction in English language teaching (Sedanza & Abatayo, 2026). Eligibility was defined in PICOS terms: English learners at any educational level; a differentiation intervention across one or more of Tomlinson’s (2017) four dimensions; a control or comparison condition; at least one cognitive academic-language outcome; and a controlled experimental or quasi-experimental design reporting sufficient statistics to compute an effect size. Six databases

(Google Scholar, ERIC, EBSCOhost, Wiley Online Library, ScienceDirect, and Harzing's Publish or Perish) were searched for work published between 2011 and 2024, supplemented by citation-pearl growing and expert consultation. Screening followed the four PRISMA phases. The process yielded 56 effect sizes from 45 studies conducted in 18 countries and involving 16,773 learners. Because the present paper classifies rather than pools these studies, it is reported as a distinct, complementary output and has been disclosed as such. The characteristics and risk-of-bias rating of all 56 effect sizes are reported in Table 1.

2.3 The six-facet classification

Each effect size was classified along six facets developed iteratively from the data and from established theory (Figure 1). Facet 1, learner/exceptionality, recorded the diversity focus of the sample (special educational needs, linguistic, cultural, socioeconomic, migration-related, gender-related, giftedness, and other support-related). Facet 2, differentiation dimension, coded which of Tomlinson's (2017) elements—content, process, product, or learning environment—were adjusted. Facet 3, differentiation format, distinguished traditional, integrated, modern, and emergent designs. Facet 4, outcome domain, recorded the academic-language skill measured. Facet 5, design rigor,

separated randomized and true-experimental studies from quasi-experimental ones, and carried out the study-level risk-of-bias appraisal. Facet 6, context, captured region, English-teaching setting (English as a foreign vs. second language), and educational level. All 56 effect sizes were independently double-coded. The two coding files were compared facet by facet, and all discrepancies were resolved through discussion to consensus. After clarification of the coding rules, final agreement was complete across the categorical facets.

2.4 Risk of bias and analysis

Each study was appraised with a domain-based framework adapted from the Cochrane Risk-of-Bias tool and ROBINS-I, covering group formation, baseline comparability, outcome measurement, incomplete data, and selective reporting. Analysis was descriptive and structural: this review tabulated coverage across facets, constructed cross-classifications and an evidence-gap map, and examined the association between design rigor and effect magnitude. Cell-level Hedges' g values are reported as descriptive summaries of the classification and are unweighted; the random-effects pooled estimate for this corpus is reported separately in the companion meta-analysis. All figures are direct outputs of the coded dataset.

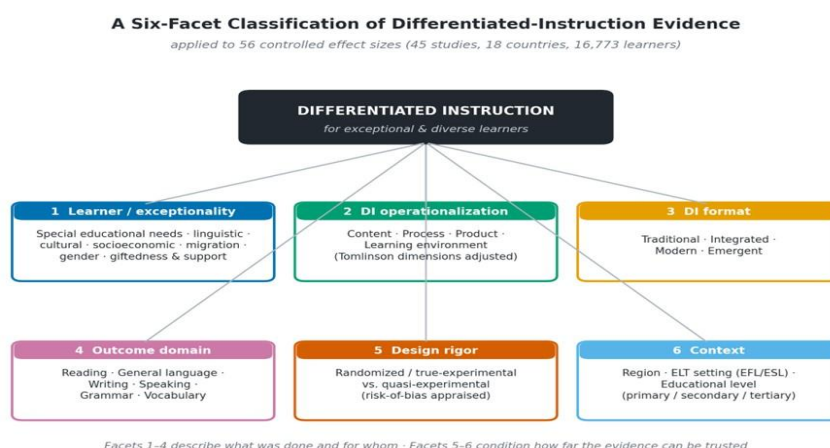


Figure 1. The six-facet classification framework applied to the corpus. Facets 1–4 describe what was done and for whom; facets 5–6 condition how far each finding can be trusted.

3. Results

3.1 The landscape of the evidence base

The 56 effect sizes describe a field that is geographically and pedagogically lopsided (Figure 2). More than half originate in Asia (31 of 56), with North America contributing most of the remainder (18) and Europe and Africa together

accounting for only seven. Foreign-language settings outnumber second-language settings (35 vs. 21), and the studies cluster at the secondary (23) and primary (21) levels, with tertiary and adult contexts thinner (12). The picture is not of a mature, evenly developed evidence base but of a young and unevenly developed one.

The landscape of the evidence base (k = 56 effect sizes)

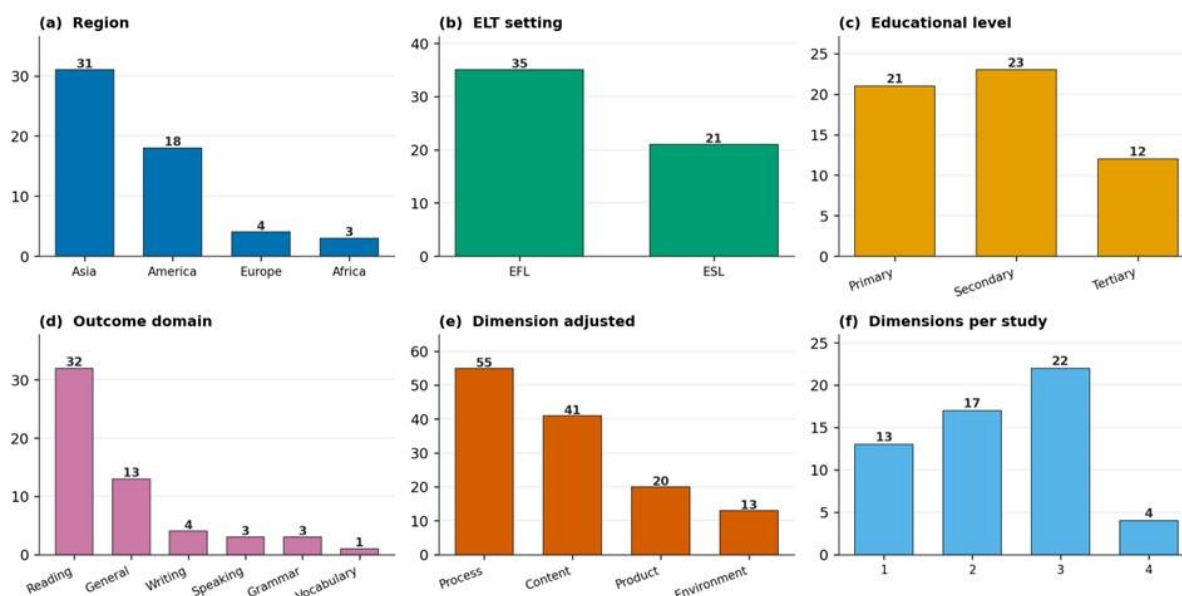


Figure 2. The landscape of the evidence base (k = 56). Distributions across region, English-teaching setting, educational level, outcome domain, adjusted differentiation dimension, and number of dimensions modified per study.

3.2 How differentiation is operationalized

The label “differentiation” conceals a narrow practice. Of Tomlinson’s four dimensions, process was adjusted in almost every study (55 of 56) and content in most (41), but product appeared in only 20, and the learning environment in just 13. Few studies were genuinely multidimensional as the theory envisions: 13 modified a single dimension, and only 4 addressed all 4 dimensions. Although differentiation is often described as comprehensively adaptive, in practice, most studies varied how students work through largely unchanged material. This matters in particular for exceptional learners, for whom the unexamined dimensions—what is produced and the

environment in which learning happens—are often the ones that determine access.

3.3 What gets measured, and for whom

Outcomes are equally concentrated. Reading accounts for 32 of the 56 effect sizes and general language ability for a further 13, leaving writing (4), speaking (3), grammar (3), and vocabulary (1) barely represented. The productive skills on which communicative competence depends are severely underrepresented. On the learner facet, more than a third of the corpus (21 of 56) focused on learners with special educational needs, with the remainder distributed thinly across linguistic, cultural, socioeconomic, migration-related, gender-related, and giftedness dimensions. The evidence-gap map (Figure 3)

makes the pattern visible: dense, dark cells where reading meets process and content adjustment, and empty cells—marked with crosses—wherever productive skills meet the product or environment dimensions. Panel (b) shows the same tilt: mean effects rise as settings become

more foreign-language and less well resourced (e.g., Saleh, 2021; Grain et al., 2022; Magableh & Abdullah, 2020; Yavuz, 2020; Sapan & Mede, 2022), foreshadowing the rigor analysis that follows.

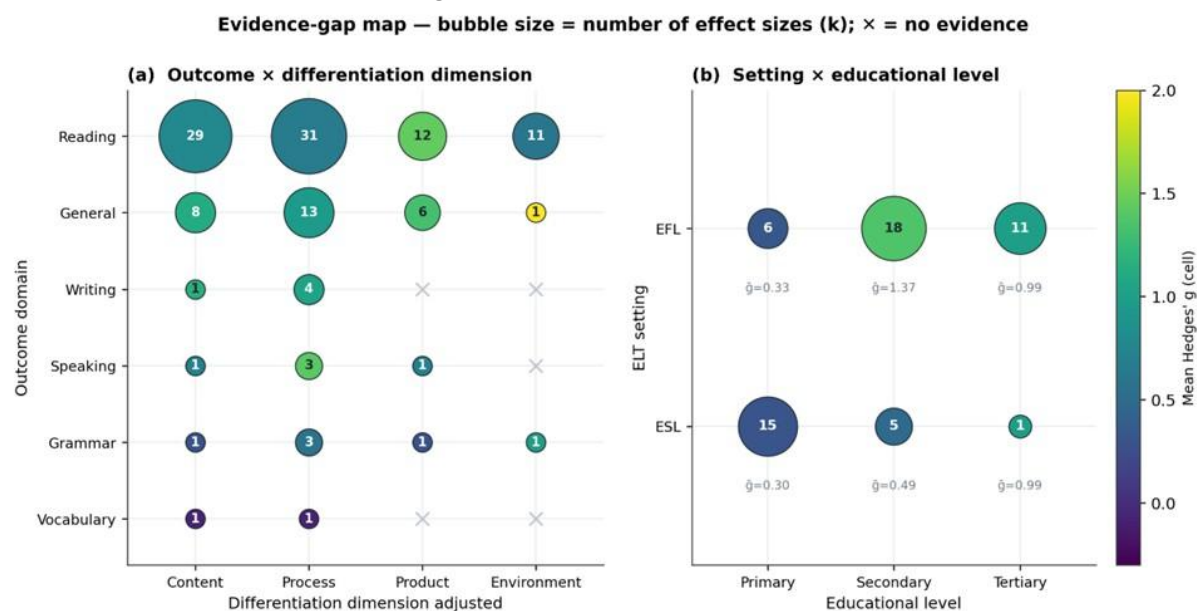


Figure 3. Evidence-gap map. (a) Outcome domain × differentiation dimension and (b) English-teaching setting × educational level. Bubble size is the number of effect sizes; color is the mean Hedges' g of the cell; crosses mark combinations with no evidence.

3.4 The inverse-rigor gradient

The most consequential pattern emerges only when the typology is crossed with design quality (Figure 4). The corpus contains 13 randomized or true-experimental effect sizes and 43 quasi-experimental ones. The two groups show sharply different patterns. Every one of the 13 very-large effects ($g \geq 1.20$) that anchor the upper end of the distribution—up to $g = 3.12$ in a quasi-experimental study from Egypt (Bahgat et al., 2019)—comes from a quasi-experimental design; not a single randomized study reaches that threshold. The quasi-experimental effects average $g = 1.03$, while the randomized effects average $g = 0.10$, barely distinguishable from no

effect. The single most negative result in the corpus ($g = -0.47$) comes from a true-experimental study. The null and negative randomized findings cluster among young, struggling readers in well-resourced North American and European classrooms (e.g., Al Otaiba et al., 2011; Reis et al., 2011; Little et al., 2014; Peters et al., 2022)—among the learners for whom differentiation is frequently recommended. This gradient is consistent with the broader observation that quasi-experiments yield larger effects than randomized trials (Cheung & Slavin, 2016), and it suggests that the headline benefit of differentiation may partly reflect the design conditions under which the largest effects were produced.

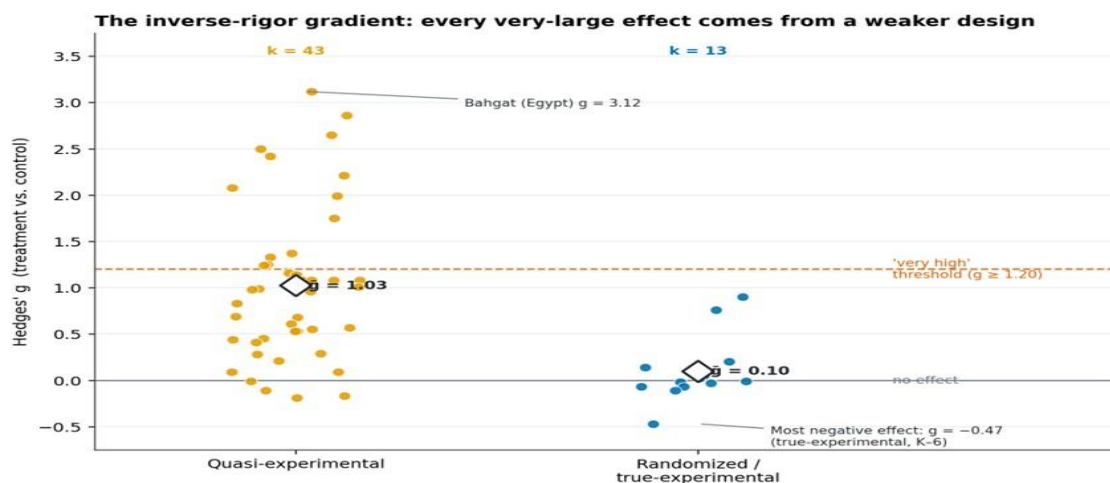


Figure 4. The inverse-rigor gradient. Each point is one effect size; diamonds mark group means. All 13 very-large effects (above the dashed $g = 1.20$ line) come from quasi-experimental studies; the randomized effects cluster around zero.

Two cautions keep this finding honest. First, design rigor is entangled with context: the randomized studies are overwhelmingly set in North American second-language settings, while the very large quasi-experimental effects come from lower-resource foreign-language settings. Rigor and context cannot be fully separated in the existing literature, and the gradient is therefore read as a structural feature of the evidence base

rather than as a clean causal estimate. Second, because the present synthesis classifies rather than pools, the means reported here are descriptive; a formal moderator test of design on the variance-weighted estimates is reported in the companion meta-analysis. Neither caution softens the central point: within this corpus, the more rigorous designs report the most modest gains.

Table 1. Characteristics and risk-of-bias of the 56 included effect sizes, ordered by Hedges' g .

#	g	Study	Country	Set.	Level	Design	Outcome	RoB
1	3.12	Bahgat (2019)	Egypt	EFL	Secondary	Quasi-exp.	Reading	Some concerns
2	2.86	Badawi (2019)	Egypt	EFL	Secondary	Quasi-exp.	Reading	Some concerns
3	2.65	Li (2019)	China	EFL	Tertiary	Quasi-exp.	Speaking	Some concerns
4	2.50	Magableh (2020)	Jordan	EFL	Secondary	Quasi-exp.	Gen. language	Some concerns
5	2.42	Magableh (2020)	Jordan	EFL	Primary	Quasi-exp.	Reading	Some concerns
6	2.21	Saleh (2021)	Bahrain	EFL	Secondary	Quasi-exp.	Reading	Some concerns
7	2.08	Grain (2022)	Iraq	EFL	Tertiary	Quasi-exp.	Gen. language	Some concerns

8	1.99	Labordo (2024)	Philippines	ESL	Secondary	Quasi-exp.	Reading	Some concerns
9	1.75	Magableh (2022)	Jordan	EFL	Secondary	Quasi-exp.	Reading	Some concerns
10	1.37	Pasubillo (2014)	Iran	EFL	Secondary	Quasi-exp.	Reading	Some concerns
11	1.33	Arifani (2020)	Indonesia	EFL	Secondary	Quasi-exp.	Writing	Some concerns
12	1.25	Daidsen (2018)	USA	ESL	Primary	Quasi-exp.	Reading	Some concerns
13	1.24	Aghdam (2024)	Iran	EFL	Tertiary	Quasi-exp.	Writing	Some concerns
14	1.16	Ganapathy (2022)	Malaysia	EFL	Secondary	Quasi-exp.	Writing	Some concerns
15	1.14	Hung (2018)	Vietnam	EFL	Tertiary	Quasi-exp.	Gen. language	Some concerns
16	1.08	Brown (2017)	USA	ESL	Primary	Quasi-exp.	Reading	Some concerns
17	1.08	Yavuz (2020)	Turkey	EFL	Secondary	Quasi-exp.	Gen. language	Some concerns
18	1.08	Masykur (2024)	Indonesia	EFL	Secondary	Quasi-exp.	Gen. language	Some concerns
19	1.08	Tamimi (2024)	Palestine	EFL	Secondary	Quasi-exp.	Gen. language	Some concerns
20	1.01	Magableh (2020)	Jordan	EFL	Secondary	Quasi-exp.	Gen. language	Some concerns
21	0.99	Webb (2016)	China/USA	ESL	Tertiary	Quasi-exp.	Grammar	Some concerns
22	0.98	Aliakbari (2014)	Iran	EFL	Secondary	Quasi-exp.	Reading	Some concerns
23	0.96	Magableh (2021)	Jordan	EFL	Secondary	Quasi-exp.	Reading	Some concerns
24	0.90	Qiao (2023)	China	EFL	Tertiary	RCT / true-exp.	Speaking	Low
25	0.83	Wei (2023)	China	EFL	Tertiary	Quasi-exp.	Gen. language	Some concerns
26	0.76	Matsumura (2012)	USA	ESL	Primary	RCT / true-exp.	Reading	Low

27	0.69	Phueakphud (2019)	Thailand	EFL	Secondary	Quasi-exp.	Speaking	Some concerns
28	0.68	Soleimani (2012)	Iran	EFL	Tertiary	Quasi-exp.	Gen. language	Some concerns
29	0.61	Ocampo (2018)	Philippines	ESL	Secondary	Quasi-exp.	Reading	Some concerns
30	0.57	Brown (2017)	USA	ESL	Primary	Quasi-exp.	Reading	Some concerns
31	0.55	Brown (2017)	USA	ESL	Primary	Quasi-exp.	Reading	Some concerns
32	0.53	Magableh (2019)	Jordan	EFL	Secondary	Quasi-exp.	Reading	Some concerns
33	0.53	Han (2022)	China	EFL	Tertiary	Quasi-exp.	Gen. language	Some concerns
34	0.45	Melka (2022)	Ethiopia	EFL	Secondary	Quasi-exp.	Grammar	Some concerns
35	0.44	Sapan (2022)	Turkey	EFL	Secondary	Quasi-exp.	Gen. language	Some concerns
36	0.41	ALTAS (2021)	Turkey	EFL	Tertiary	Quasi-exp.	Writing	Some concerns
37	0.29	AlHashmi (2018)	Saudi Arabia	EFL	Tertiary	Quasi-exp.	Grammar	Some concerns
38	0.28	Houtveen (2012)	Netherlands	EFL	Primary	Quasi-exp.	Reading	Some concerns
39	0.21	Gettinger (2012)	USA	ESL	Primary	Quasi-exp.	Reading	Some concerns
40	0.20	Al Otaiba (2011)	USA	ESL	Primary	RCT / true-exp.	Reading	Low
41	0.14	Connor (2011)	USA	ESL	Primary	RCT / true-exp.	Reading	Low
42	0.09	Houtveen (2012)	Netherlands	EFL	Primary	Quasi-exp.	Reading	Some concerns
43	0.09	Han (2022)	China	EFL	Tertiary	Quasi-exp.	Gen. language	Some concerns
44	0.05	Reis (2011)	USA	ESL	Primary	RCT / true-exp.	Reading	Low
45	0.02	Reis (2011)	USA	ESL	Primary	RCT / true-exp.	Reading	Some concerns

46	- 0.01	Al Otaiba (2011)	USA	ESL	Primary	RCT / true-exp.	Reading	Low
47	- 0.01	Brown (2017)	USA	ESL	Primary	Quasi-exp.	Reading	Some concerns
48	- 0.02	Little (2014)	USA	ESL	Secondary	RCT / true-exp.	Reading	Low
49	- 0.03	Little (2014)	USA	ESL	Secondary	RCT / true-exp.	Reading	Low
50	- 0.07	Al Otaiba (2011)	USA	ESL	Primary	RCT / true-exp.	Reading	Low
51	- 0.07	Connor (2011)	USA	ESL	Primary	RCT / true-exp.	Vocabulary	Low
52	- 0.11	Connor (2011)	USA	ESL	Primary	RCT / true-exp.	Reading	Low
53	- 0.11	Tohn (2019)	USA	ESL	Secondary	Quasi-exp.	Gen. language	Some concerns
54	- 0.17	Peters (2022)	Germany	EFL	Primary	Quasi-exp.	Reading	Some concerns
55	- 0.19	Peters (2022)	Germany	EFL	Primary	Quasi-exp.	Reading	Some concerns
56	- 0.47	Avci (2013)	Turkey	EFL	Primary	RCT / true-exp.	Reading	Some concerns

Note. Design: RCT / true-exp. = randomized or true-experimental; Quasi-exp. = quasi-experimental. RoB = overall risk-of-bias rating (domain-level appraisal in Supplementary Table S7); 11 effect sizes were rated low, and 45 had some concerns. Set. = ELT setting (EFL/ESL). Hedges' *g* is the standardized mean difference (positive favors differentiated instruction). Full facet coding (differentiation dimensions and format, and diversity focus) is provided in the Supplementary Material.

4. Discussion

4.1 A map, not a verdict

Classifying the literature rather than averaging it changes the question from “Does differentiation work?” to “What has the field actually built evidence about?” The answer is sobering. The

field has concentrated its effort on a narrow slice of the possible: adjusting process and content, measuring reading, in foreign-language secondary classrooms, often without random assignment. Whole regions of the map—productive language skills, the learning environment, the full range of exceptionality dimensions, and rigorous designs in low-resource settings—remain sparsely populated or empty. A practice cannot be said to be evidence-based for a territory that the evidence has never visited.

4.2 An equity reading

For special and inclusive education, the gradient carries a pointed warning. The most rigorous evidence now available concerns young, struggling learners in well-resourced classrooms, and it shows that differentiation has a low effect on outcomes. The largest effects in this corpus

come from weaker designs in lower-resource settings. Read together, these patterns suggest that the confidence with which differentiation is recommended for these learners outruns the quality of evidence behind it. This is not an argument against differentiating; it is an argument against assuming the benefit is uniform, large, and established for the learners who can least afford instruction that misses.

For special and inclusive education specifically, the question is not whether teachers should adapt instruction but whether particular adaptations have been tested for particular needs. The typology shows that the most access-relevant dimensions—what learners are asked to produce and the environment in which they work—remain the least examined (product in 20 of 56 studies, the learning environment in 13), even though these are often decisive for the participation of learners with disabilities and other support needs. The evidence-gap map, therefore, identifies not only where research is thin in general, but where it is thin precisely on the dimensions that inclusive education most depends upon. This review should therefore be read as an inclusive-education synthesis of differentiated-instruction evidence for diverse English learners, not as a complete review of disability-specific intervention research.

4.3 Implications for practice

Three implications follow for practitioners and teacher educators. First, treat differentiation as a set of distinct practices, not a single intervention: a study that adjusts only the process is weak evidence for a program that promises to adapt product and environment as well. Second, weigh the evidence by its design when planning provision for exceptional learners, and be wary of large effects drawn from non-randomized designs. Third, concentrate scarce effort where the map and the rigorous evidence converge—identifying priority learners, grouping by need, and sustaining a small number of well-implemented adaptations—rather than attempting to differentiate every dimension for every student at once.

4.4 Limitations and future directions

This synthesis includes only quantitative, controlled studies that meet the parent corpus's eligibility criteria; qualitative and single-case evidence, which is important in special education, is not represented. The diversity-dimension facet is reported at the corpus level; a full cross-classification of exceptionality against differentiation type awaits the release of the per-study diversity codes. And because rigor and context are confounded, the inverse-rigor gradient should be tested directly. This review calls for three things: primary studies that report who did the adapting, who was prioritized, and which dimensions were modified; rigorous trials in the low-resource, foreign-language settings where effects currently look largest; and randomized evidence on productive skills and on the exceptionality groups the map leaves blank.

5. Conclusion

Differentiated instruction is not one thing, and the evidence for it is not evenly built. By classifying rather than averaging the controlled literature, this review offers special and inclusive education a usable map: a typology of what has been studied, an honest picture of where the evidence is thin, and a clear signal that the strongest designs report the most modest gains. The task ahead is not simply to differentiate more, but to differentiate in ways that are supported by stronger evidence.

Author Declaration

The author confirms that this work is original, has not been published elsewhere, and is not under consideration by another journal. The author declares no conflicts of interest and received no specific funding for this work.

Generative AI Statement

Generative AI tools were used only for language editing and to assist with the visual presentation of author-created figures. All figures are direct outputs of the coded dataset. The author reviewed and verified all content and takes full responsibility for the manuscript.

References

- Al Otaiba, S., Connor, C. M., Folsom, J. S., Greulich, L., & Meadows, J. (2011). Assessment data-informed guidance to individualize kindergarten reading instruction: Findings from a cluster-randomized control field trial. *Elementary School Journal*, 111(4), 535–560. <https://doi.org/10.1086/659031>
- Asriadi, M., Hadi, S., Istiyono, E., & Retnawati, H. (2023). Does differentiated instruction affect learning outcome? Systematic review and meta-analysis. *Journal of Pedagogical Research*, 7(1), 18–33. <https://doi.org/10.33902/JPR.202322021>
- Bahgat, R. M. B., Shehata, G. M. S., Ali, A. B. A., & Salem, M. H. S. (2019). The effect of differentiated instruction on developing English reading comprehension skill of sixth graders. *Journal of Educational Sciences*, 38, 85–99. <https://doi.org/10.21608/maeq.2019.140978>
- Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283–292. <https://doi.org/10.3102/0013189X16656615>
- Deunk, M. I., Smale-Jacobse, A. E., de Boer, H., Doolaard, S., & Bosker, R. J. (2018). Effective differentiation practices: A systematic review and meta-analysis of studies on the cognitive effects of differentiation in primary education. *Educational Research Review*, 24, 31–54. <https://doi.org/10.1016/j.edurev.2018.02.002>
- Gibbs, K., & McKay, L. (2021). Differentiated teaching practices of Australian mainstream classroom teachers: A systematic review and thematic analysis. *International Journal of Educational Research*, 109, 101799. <https://doi.org/10.1016/j.ijer.2021.101799>
- Gough, D., Thomas, J., & Oliver, S. (2012). Clarifying differences between review designs and methods. *Systematic Reviews*, 1, 28. <https://doi.org/10.1186/2046-4053-1-28>
- Grain, H. M. J. S., Neamah, N. R., Al-Gburi, G., Abduzahra, A. T., Hassan, A. Y., Kadhim, A. J., Obaid, A. A., & Yahea, S. A. (2022). Differentiated instruction effect on academic achievements of level-2 English students: A case on Iraq public-sector universities. *Eurasian Journal of Applied Linguistics*, 8(2), 87–95. <https://doi.org/10.32601/ejal.911544>
- Hu, L. (2024). Utilization of differentiated instruction in K–12 classrooms: A systematic literature review (2000–2022). *Asia Pacific Education Review*, 25, 507–525. <https://doi.org/10.1007/s12564-024-09931-y>
- Little, C. A., McCoach, D. B., & Reis, S. M. (2014). Effects of differentiated reading instruction on student achievement in middle school. *Journal of Advanced Academics*, 25(4), 384–402. <https://doi.org/10.1177/1932202X14549250>
- Magableh, I. S. I., & Abdullah, A. (2020). Effectiveness of differentiated instruction on primary school students' English reading comprehension achievement. *International Journal of Learning, Teaching and Educational Research*, 19(3), 20–35. <https://doi.org/10.26803/ijlter.19.3.2>
- Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., et al. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- Peters, M. T., Hebbeker, K., & Souvignier, E. (2022). Effects of providing teachers with tools for implementing assessment-based differentiated reading instruction in second grade. *Assessment for Effective Intervention*, 47(3), 157–169. <https://doi.org/10.1177/15345084211014926>
- Reis, S. M., McCoach, D. B., Little, C. A., Muller, L. M., & Kaniskan, R. B. (2011). The effects of differentiated instruction and enrichment pedagogy on reading achievement in five elementary schools. *American Educational Research Journal*, 48(2), 462–501. <https://doi.org/10.3102/0002831210382891>
- Saleh, A. H. A. E. (2021). The effectiveness of differentiated instruction in improving Bahraini EFL secondary school students' reading comprehension skills. *REiLA: Journal of Research and Innovation in Language*, 3(2), 135–145. <https://doi.org/10.31849/reila.v3i2.6816>
- Sapan, M., & Mede, E. (2022). The effects of differentiated instruction (DI) on achievement, motivation, and autonomy among English learners. *Iranian Journal of Language Teaching Research*, 10(1), 127–144. <https://doi.org/10.30466/ijlitr.2022.121125>
-

- Sedanza, M. A. C., & Abatayo, J. A. (2026). Beyond the average: A systematic review and meta-analysis of differentiated instruction for diverse English learners [Unpublished manuscript]. Leyte Normal University. <https://osf.io/c2v8f>
- Tomlinson, C. A. (2017). *How to differentiate instruction in academically diverse classrooms* (3rd ed.). ASCD.
- Weganofa, R., & Mujiono. (2025). A meta-analysis of the effects of differentiated instruction on English language proficiency. *JEES (Journal of English Educators Society)*, 10(2), 85–94. <https://doi.org/10.21070/jees.v10i2.1929>
- Yavuz, A. C. (2020). The effects of differentiated instruction on Turkish students' L2 achievement and student and teacher perceptions. *Eurasian Journal of Applied Linguistics*, 6(2), 313–335. <https://doi.org/10.32601/ejal.776002>