https://doi.org/10.34142/2709-7986.2026.31.2.27

PLACEMENT TEST EVIDENCE FOR DIFFERENTIATED TEACHING ENGLISH FOR SPECIFIC PURPOSES IN A FIRST-YEAR LIFE-SCIENCE COHORT

ДАНІ ВСТУПНОГО ТЕСТУВАННЯ НА КОРИСТЬ ДИФЕРЕНЦІЙОВАНОГО ВИКЛАДАННЯ АНГЛІЙСЬКОЇ МОВИ ПРОФЕСІЙНОГО СПРЯМУВАННЯ ПЕРШОКУРСНИКАМ ПРИРОДНИЧОГО ПРОФІЛЮ

Received: September 01, 2026 **Accepted:** October 12, 2026 **Published:** October 15, 2026

Volodymyr KSHEVETSKIY



Ph.D. in Philology, Associate Professor, Department of Foreign Languages for Natural Science Faculties, Yuriy Fedkovych Chernivtsi National University, Chernivtsi, Ukraine.

✉ **E-mail:** *v.kshevetskyi@chnu.edu.ua*
https://orcid.org/0000-0001-9769-4931

How to Cite this Article (APA Style):
Kshevetskiy, V. (2026). Placement test evidence for differentiated teaching English for specific purposes in a first-year life-science cohort. *Educational Challenges, 31*(2), 423–436. https://doi.org/10.34142/2709-7986.2026.31.2.27

ABSTRACT

Purpose. *In Ukrainian universities, first-year English for Specific Purposes (ESP) classes are usually formed from academic groups defined by degree programme, not by language level. The study aimed to measure how much general English proficiency varies inside such groups and what this implies for teaching.*

Methodology. *A cross-sectional quantitative design used anonymised routine placement data from 93 first-year students of the Institute of Biology, Chemistry and Bioresources, Chernivtsi National University, who took a 25-item multiple-choice general English test in 2026. The analysis combined item statistics, KR-20 reliability, variance decomposition for six intact groups (n = 88), a permutation test, grouping simulations and a comparison of three score tiers.*

Мета. *В українських університетах групи для вивчення англійської мови професійного спрямування (ESP) на першому курсі зазвичай формують за освітньою програмою, а не за рівнем володіння мовою. Дослідження мало на меті виміряти, наскільки різниться рівень загальної англійської всередині таких груп, і з'ясувати, що це означає для диференційованого навчання.*

Методологія. *У поперечному кількісному дослідженні використано знеособлені дані вступного тестування 93 студентів першого курсу Інституту біології, хімії та біоресурсів ЧНУ імені Юрія Федьковича, які у 2026 році виконали тест із загальної англійської мови з 25 завдань множинного вибору. Аналіз охоплював показники завдань, надійність KR-20, декомпозицію дисперсії для шести груп (n = 88), пермутаційний тест, моделювання поділу на групи та порівняння трьох підгруп.*

Placement test evidence for differentiated teaching English for specific purposes in a first-year life-science cohort © 2026 by Volodymyr KSHEVETSKIY is licensed under **Creative Commons Attribution-NonCommercial 4.0 International**. To view a copy of the license, visit http://creativecommons.org/licenses/by-nc/4.0/

Results. *Scores ranged from 24 to 100 (M = 69.9, SD = 22.2; KR-20 = .88). Near-chance, lower, middle and upper bands held 5.4%, 28.0%, 35.5% and 31.2% of students. Group membership explained 11.8% of score variance, and the existing groups were no more homogeneous than random allocation (p = .066). Score ranges within groups were 56 to 76 points. Level-based sets would reduce the within-unit standard deviation from 21.9 to 5.1 points, and three tiers inside each group to 11.6 points. Lower-tier students scored below chance on 11 of the 25 items; upper-tier students reached at least 79% on every item. Grammar and lexis subscores correlated strongly (ρ = .73) and did not differ significantly.*

Conclusions. *In this cohort, the academic group is not a unit of language level. Tiered tasks within each group, with separate support for near-chance students and ESP extension work for the strongest, fit the data and research evidence better than regrouping across programmes; this recommendation still needs longitudinal or experimental testing. A short general placement test is reliable enough to form tiers but not to diagnose specific needs.*

Keywords: *EFL proficiency, heterogeneity, higher education, item analysis, tiered tasks, variance decomposition.*

Результати. *Бали коливалися від 24 до 100 (M = 69,9, SD = 22,2; KR-20 = 0,88). До діапазонів випадкового вгадування, нижчого, середнього й вищого рівня належали відповідно 5,4 %, 28,0 %, 35,5 % і 31,2 % студентів. Належність до групи пояснювала 11,8 % дисперсії балів, а наявні групи виявилися не однорідніші за випадково сформовані (p = 0,066). Розмах балів у групах становив від 56 до 76. Групи, сформовані за рівнем, зменшили б стандартне відхилення всередині групи з 21,9 до 5,1 бала, а три підгрупи всередині кожної групи – до 11,6 бала. Студенти нижчої підгрупи відповідали гірше, ніж за випадкового вгадування, в 11 із 25 завдань; студенти вищої підгрупи правильно виконали щонайменше 79 % кожного завдання. Бали за граматику й лексику тісно корелювали (ρ = 0,73) і статистично значущо не різнилися.*

Висновки. *У цій когорті академічна група не є одиницею мовного рівня. Завдання кількох ярусів складності в межах групи, окремий маршрут підтримки для студентів із результатом на рівні вгадування та поглиблені ESP-завдання для найсильніших краще узгоджуються з даними й результатами досліджень, ніж переформування груп між освітніми програмами; цю рекомендацію ще належить перевірити в лонгітюдному або експериментальному дослідженні. Короткий загальний вступний тест достатньо надійний для поділу на підгрупи, але не для діагностики конкретних потреб.*

Ключові слова: *рівень володіння англійською мовою як іноземною, неоднорідність, вища освіта, аналіз тестових завдань, ярусні завдання, декомпозиція дисперсії.*

INTRODUCTION

English has a new legal status in Ukrainian higher education. The Law of Ukraine No. 3760-IX of 4 June 2024 obliges the state to ensure compulsory study of English in higher education institutions in line with each educational programme (Article 8, Part 1, Item 3). The same law ties the improvement of educational standards to the Common European Framework of Reference for Languages (CEFR) (Pro zastosuvannia anhliiskoi movy v Ukraini, 2024; Council of Europe, 2020). For students of non-language programmes, this obligation is usually met by a course of English for Specific Purposes (ESP) in the first years of study.

ESP is defined by its starting point: the course is built around what particular learners need English for (Anthony, 2018; Basturkmen, 2010). Hyland (2022) describes this as the basic idea of the field: learners' needs differ according to their future academic and occupational goals, and the English they learnt at school rarely prepares them for the genres of university study. Needs analysis in ESP therefore considers both the target situation and the learners' present command of the language.

In Ukrainian universities, ESP classes normally follow the academic group, which is formed by programme and specialty. Bolitho and West (2017) found that most universities allocate English teachers to faculties and departments. They also reported that "many teachers complain that they don't cope well with the typical mixed ability

intake in Year 1 of undergraduate study" (p. 87). The same report noted a lack of practical training in mixed-ability teaching and in assessment.

Similar difficulties appear in multilingual tertiary settings in Ukraine, where teachers must address varying levels of proficiency in Ukrainian, Hungarian and English within the same classroom (Huszti & Lechner, 2025). First-year students also face cultural, social and economic challenges of transition to university (Fabeku & Fasanmi, 2024). Much of the Ukrainian ESP literature concentrates on methods, curricula and study hours (Symonenko & Osadchyi, 2019; Tarnopolsky, 2018). The extent of proficiency differences inside intact first-year groups is often mentioned but rarely measured.

Placement tests are the usual source of such measurement. They are meant to place students at the stage of a teaching programme most appropriate to their abilities (Hughes & Hughes, 2020). Their validation literature asks whether a test identifies the students who most need support, whether its content suits its use and whether it is reliable (Wall et al., 1994). Institutional constraints limit what such a test can do (Fulcher, 1997), and decisions based on a single score risk misplacement: Kokhan (2013) estimated about a 40% chance of misplacement when standardised test scores replaced a local placement test.

Green and Weir (2004) raised a further limitation: quick grammar-based placement tests may not provide enough information to guide diagnostic teaching or the sequencing of materials. Placement data are therefore useful for describing the spread of a cohort and for grouping, but should be read with caution when used to prescribe content. Where entry and exit testing is repeated, as in Goundar's (2023) study of first-year students in Fiji, such data can also show how much a cohort moves during the first year.

A university can respond to mixed proficiency in two ways: by regrouping students by level across programmes (streaming) or by differentiating instruction within the academic groups. Differentiated instruction adapts content, process and product to students' readiness, interests and learning profiles (Tomlinson, 2014). The evidence on these two responses is not symmetrical. A second-order meta-analysis of about a century of research found that students benefited from within-class grouping ($0.19 \leq g \leq 0.30$) but not from between-class grouping ($0.04 \leq g \leq 0.06$) (Steenbergen-Hu et al., 2016).

In primary education, homogeneous ability grouping on its own had a small negative effect on low-ability students (Deunk et al., 2018). In secondary education, most studies of within-class differentiation reported small to moderate positive effects (Smale-Jacobse et al., 2019). Differentiated instruction also tends to benefit high-achieving students, although teachers rarely use it for them proactively (Ziernwald et al., 2022). In foreign language teaching, tiered assignments and levelled texts improved reading comprehension in mixed-ability EFL classes (Magableh & Abdullah, 2021), and large mixed-ability classes have been described as a resource rather than only a constraint (Winke & Rawal, 2018).

The same literature shows why differentiation is hard to sustain. Definitions of differentiated instruction vary widely across studies (Bondie et al., 2019). Teachers report implementing it only occasionally (Pozas et al., 2020), and its use depends on teachers' self-efficacy and on class size (Suprayogi et al., 2017). Assessment data can help: in maritime ESP, digital benchmarking and formative assessment tools supported the development of cadets' professional English competence (Kononova, 2026). What a department first needs, however, is a clear picture of how large the differences inside its

groups are and where they lie. Most of the evidence cited above comes from school education; quantitative descriptions of intact university ESP groups are scarce.

The **purpose** of this study was to measure the distribution of entry general English proficiency within and between the intact first-year groups of one natural-science institute, and to draw from it implications for grouping and differentiated ESP teaching. The study addressed four research questions:

RQ1. What is the entry general English proficiency profile of the cohort, and how reliable is the placement test for this cohort?

RQ2. How is score variance distributed between and within the intact academic groups?

RQ3. How far would alternative arrangements (level-based sets across groups, or tiers inside each group) reduce the spread of proficiency that a teacher faces?

RQ4. Which item areas separate the score tiers, and what does this imply for tier-specific instruction?

METHODOLOGY

The study used a cross-sectional quantitative design. It was a secondary analysis of anonymised placement test data collected as part of routine teaching at the start of the 2026/2027 academic year.

– Participants

The participants were first-year bachelor's students of the Educational and Scientific Institute of Biology, Chemistry and Bioresources (IBChB) at Yuriy Fedkovych Chernivtsi National University. The institute trains students in biology, ecology, chemistry, biotechnology, food quality and safety, agricultural production, land management and the teaching of biology. English is taught to these students by the Department of Foreign Languages for Natural Science Faculties as a course of English for professional purposes, in the students' academic groups.

The test recorded 94 responses. One student submitted twice; only the first attempt was kept, which left 93 students. Of these, 88 stated their academic group, and they belonged to six groups of 9 to 22 students. The five students without a group number were included in the cohort-level analyses and excluded from the group-level analyses. To protect anonymity, the groups are labelled A–F in this article. Age and gender were not recorded; all participants were first-year entrants of 2026. The same test was used in other units of the university; their results are outside the scope of this article.

– Instrument

The instrument was the 25-item "Test your English – General English" test published online by Cambridge University Press & Assessment (n.d.). Items 1–5 ask the test-taker to choose the best response in a short exchange from three options. Items 6–25 are sentences with a gap and four options, and they target grammar and vocabulary, including collocation. For the analysis, I classified the items into three areas: functional language (items 1–5), grammar (items 6, 7, 8, 9, 11, 12 and 21) and lexis (items 10, 13–20 and 22–25). Items 8 (*as usual*) and 11 (*falling*) lie between grammar and lexis; they were kept in the grammar area.

The test measures general English, not ESP, and its publisher does not provide a CEFR level for a given score. The scores are therefore read here as a ranking within the cohort, not as CEFR levels. Blind guessing on all 25 items would yield an expected score of about 27 points.

– **Procedure**

The test was set up as a Google Forms quiz worth 4 points per item (0–100). Students took it online on 8–9 September 2026, during the first weeks of the course, without supervision. Each student saw the score immediately after submitting. Because a repeat attempt could then be informed by the first result, only first attempts were analysed. The answer key reproduced the score recorded by the form for every response.

Scores were grouped into four descriptive bands. The near-chance band (≤ 28 points, at most seven correct answers) covers results that blind guessing could produce. The lower band covers 32–56 points (8–14 correct), the middle band 60–84 points (15–21 correct) and the upper band ≥ 88 points (at most three errors). For the item analysis, the near-chance and lower bands were merged into one lower tier, which gave three tiers of similar size.

The data were pseudonymised before the analysis: names were removed and each email address was replaced by a code. Results are reported only in aggregate, and no individual can be identified from them. The study involved no intervention. Because it used only anonymised data collected for routine teaching, no separate research consent was obtained; the analysis followed the confidentiality principles of the Declaration of Helsinki (2008 revision).

– **Statistical analysis**

For RQ1, I computed descriptive statistics and the Shapiro–Wilk test of normality. Reliability was estimated with the Kuder–Richardson formula 20 (Kuder & Richardson, 1937), and the standard error of measurement (SEM) as $SD \times \sqrt{(1 - \text{KR-20})}$. Item facility ($p$) and the corrected item–total correlation (point-biserial correlation with the sum of the other items) were computed for every item. A Spearman correlation tested whether items became harder towards the end of the test.

For RQ2, differences between the six groups were tested with the Kruskal–Wallis test, with ε^2 as the effect size (Tomczak & Tomczak, 2014), and with a one-way ANOVA, with η^2 as the share of variance between groups. The intraclass correlation ICC(1), the usual index of clustering in educational data (Hedges & Hedberg, 2007), was also computed. The Brown–Forsythe test compared the spread of scores across groups. A permutation test compared the pooled within-group standard deviation of the real groups with 5,000 random allocations of the 88 scores to six groups of the observed sizes.

For RQ3, I compared the pooled within-unit standard deviation of the real groups with that of three alternative arrangements of the same 88 students. These were (a) six level-based sets of the observed sizes, formed by rank order across the institute; (b) two tiers inside each group; and (c) three tiers inside each group. For RQ4, item facility was computed for each tier and each item area. The Wilcoxon signed-rank test, with the rank-biserial correlation as the effect size, compared each student's grammar and lexis subscores (percentage correct), and a Spearman correlation described their relationship. All tests were two-tailed at $\alpha = .05$. The analysis used Python 3.9.6 with pandas 2.3.3, SciPy 1.13.1 and matplotlib 3.9.4.

RESULTS

– **Cohort profile and test reliability (RQ1)**

Table 1 summarises the scores of the 93 students. The distribution was negatively skewed and departed from normality (Shapiro–Wilk $W = .92$, $p < .001$). The median (76) was higher than the mean (69.9), and almost a third of the students scored 88 or more. Five students (5.4%) scored at or below the chance band, and four scored 100.

Table 1

Placement Test Scores of the Cohort (N = 93)

Statistic	Value
Mean (95% CI)	69.9 (65.3–74.5)
Standard deviation	22.2
Median (interquartile range)	76 (52–88)
Range	24–100
Skewness	−0.47
KR-20 reliability	.88
Standard error of measurement	7.6 points
Near-chance band (≤ 28)	5 (5.4%)
Lower band (32–56)	26 (28.0%)
Middle band (60–84)	33 (35.5%)
Upper band (≥ 88)	29 (31.2%)

Note. Scores 0–100 (4 points per item); bands as n (%).

The test was reliable for this cohort (KR-20 = .88). The SEM was 7.6 points, or about two items. Item facility ranged from .42 (item 10) to .92 (items 1 and 2). The functional items were the easiest (p = .83–.92), while the grammar (p = .48–.90) and lexis items (p = .42–.87) covered a wider range. Items became harder towards the end of the test (Spearman ρ = −.52, p = .007). Corrected item–total correlations ranged from .14 (item 13) to .70 (item 21). Four items discriminated weakly (r < .25): items 2, 4, 13 and 14.

– **Variation between and within groups (RQ2)**

Table 2 and Figure 1 show the scores of the six groups. The Kruskal–Wallis test indicated a difference between groups, $H(5) = 11.73$, p = .039, ε^2 = .13, while the one-way ANOVA did not reach significance, $F(5, 82) = 2.19$, p = .064. Group membership accounted for 11.8% of the score variance (η^2 = .118; ICC(1) = .08). The remaining 88.2% lay within the groups.

Table 2

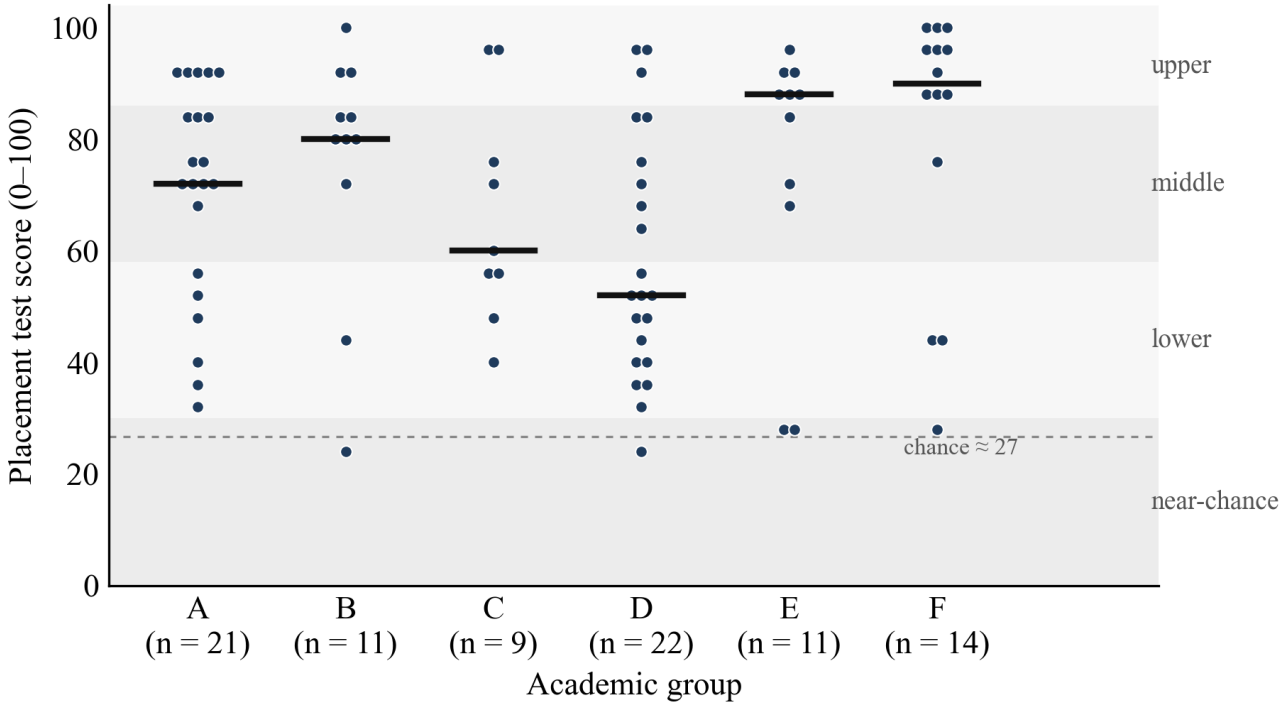
Scores and Score Bands by Academic Group (n = 88)

Group	*n*	*M*	*SD*	*Mdn*	Min–Max	NC	L	Mid	U
A	21	70.7	19.5	72	32–92	0	6	10	5
B	11	75.6	22.4	80	24–100	1	1	6	3
C	9	66.7	19.9	60	40–96	0	4	3	2
D	22	58.7	21.8	52	24–96	1	12	6	3
E	11	74.9	24.7	88	28–96	2	0	3	6
F	14	81.1	24.2	90	28–100	1	2	1	10

Note. Groups relabelled A–F. Band columns give numbers of students: NC = near-chance (≤ 28), L = lower (32–56), Mid = middle (60–84), U = upper (≥ 88).

Figure 1

Placement Test Scores of Individual Students by Academic Group



Note. Dots are students (tied scores side by side); bars mark group medians; shading shows the four bands; the dashed line marks the chance score (about 27). Source: the author's placement test data, September 2026 (n = 88). Plotted with matplotlib 3.9.4 using code prepared with the assistance of Claude (Anthropic, Claude Opus 5.5) on 27.09.2026, from the author's data.

The spread was similar in every group. The Brown–Forsythe test found no difference in variance between groups, W = 0.10, p = .99. Every group spanned at least three of the four bands, and three groups (B, D and F) spanned all four. The distance between the weakest and the strongest student in a group ranged from 56 to 76 points, that is, from 14 to 19 items. The median student scored 12 points (three items) away from the median of their own group; 43.2% of students were more than 12 points away, and 22.7% more than 20 points away.

The existing groups were not more homogeneous than random allocation. Their pooled within-group standard deviation (21.9 points) was close to the mean of 5,000 random allocations of the same students to groups of the same sizes (22.7 points; 95% interval 21.6–23.2; p = .066).

– **Alternative arrangements (RQ3)**

Table 3 compares the real groups with three alternative arrangements. Level-based sets across the institute produced the most homogeneous units, with score ranges of only 4 to 24 points. Tiers inside the existing groups reduced the spread by about a third (two tiers) and by about half (three tiers). With three tiers, the median range inside a tier was 16 points (four items). The largest range within a tier was 60 points. It occurred where a near-chance student was the lowest scorer in a small group and shared a tier with much stronger peers.

Table 3

Spread of Scores Within Teaching Units Under Different Arrangements (n = 88)

Arrangement	Units	Pooled within-unit SD	Change
Existing academic groups	6	21.9	—
Random allocation, 5,000 runs: mean [95% interval]	6	22.7 [21.6–23.2]	+3%
Two tiers inside each group	12	15.1	−31%
Three tiers inside each group	18	11.6	−47%
Level-based sets across groups	6	5.1	−77%

Note. Level-based sets keep the six observed group sizes and are filled in rank order. Tiers split each group by score into parts of equal or near-equal size. Change is relative to the existing groups.

– **Item areas across tiers (RQ4)**

Table 4 shows facility by item area for the three tiers. The lower tier (*n* = 31) answered the functional items well (78.1% correct) but only about a third of the grammar and lexis items. In this tier, 13 items had a facility below .50, and 11 items had a facility below .25, the rate expected from blind guessing on a four-option item. The lowest facilities in the lower tier were on items 10 (*supply*) and 21 (*arriving*), at .06 each, and on items 9 (*Provided*) and 19 (*set*), at .10 each.

Table 4

Percentage of Correct Responses by Item Area and Score Tier

Item area	Lower tier (*n* = 31)	Middle tier (*n* = 33)	Upper tier (*n* = 29)	Cohort (*N* = 93)
Functional (items 1–5)	78.1	90.3	96.6	88.2
Grammar (7 items)	37.8	70.1	96.6	67.6
Lexis (13 items)	31.5	71.6	90.5	64.1
All 25 items	42.5	75.0	93.5	69.9
Items with $p < .50$	13	3	0	3
Items with $p < .25$	11	0	0	0

Note. Lower tier = scores ≤ 56 (near-chance and lower bands); middle tier = 60–84; upper tier ≥ 88. The last two rows count items by facility (*p*) within each tier.

The middle tier (*n* = 33) answered about 70% of the grammar and lexis items correctly. Only three items fell below .50 in this tier: item 6 (present perfect with *for*, .48), item 9 (*Provided* as a conditional conjunction, .48) and item 10 (the collocation *electricity supply*, .42). The upper tier (*n* = 29) answered at least 79% of responses correctly on every item; its lowest facility was again on item 10.

Grammar and lexis subscores were closely related (Spearman ρ = .73). Across the cohort, students answered 67.6% of grammar items and 64.1% of lexis items correctly. The difference was not significant (Wilcoxon $W = 1545$, $p = .086$, rank-biserial $r = .21$). Only 14.0% of students showed a gap of 30 percentage points or more between their grammar and lexis subscores.

DISCUSSION

The main finding is that, in this cohort, the academic group is not a unit of language level. Nearly nine-tenths of the score variance lay inside the groups, and the groups were no more homogeneous than random allocation. The Kruskal–Wallis result shows that some groups were stronger on average than others; group F, for example, had a median of 90. Yet even the strongest group included a near-chance student and two students in the lower band. The complaint that Bolitho and West (2017) recorded, that teachers struggle with the "mixed ability intake in Year 1" (p. 87), can therefore be stated in numbers for this institute. The typical student scored three items away from their group's median, and for more than a fifth of the students the gap exceeded five items.

Because the spread was equally large in every group, it is a structural feature of how groups are formed, not a problem of a particular "difficult" group. The response should therefore be departmental rather than left to individual teachers, the more so because the use of differentiated instruction depends on context, class size and teachers' self-efficacy (Pozas et al., 2020; Suprayogi et al., 2017).

The comparison of arrangements (RQ3) shows the trade-off clearly. Level-based sets would produce almost homogeneous classes. Three arguments nevertheless favour tiers within the existing groups. First, the research evidence favours within-class over between-class grouping: between-class grouping showed almost no effect on achievement (Steenbergen-Hu et al., 2016), and homogeneous grouping on its own can harm low-ability students (Deunk et al., 2018).

Second, level-based sets would mix students of biology, chemistry, land management and food technology in one class. This would weaken the link to the students' disciplines on which ESP depends (Anthony, 2018; Hyland, 2022). Third, streaming requires common timetable slots for groups whose schedules are set by their programmes. In the simulation, three tiers inside each group would halve the spread while keeping the disciplinary link. Tiers are therefore the option better supported by the available evidence.

These comparisons should be interpreted with caution. The study is cross-sectional, and the arrangements in Table 3 are simulations: they show how the spread of scores that a teacher faces would change. They do not show how students would learn under each arrangement. A smaller spread makes better-targeted teaching possible, but it is not evidence of higher achievement. The advantage of three tiers is therefore a reasoned pedagogical recommendation, based on the spread observed in this cohort and on research on within-class grouping, rather than a demonstrated effect. Whether tiered teaching improves ESP learning, and for which students, needs to be tested in a longitudinal or experimental study, for example by comparing the progress of tiered and non-tiered groups over an academic year.

Tiers do not solve every case. The widest tiers contained a near-chance student placed with much stronger peers in a small group. Such students differ from their group in kind, not only in degree, and need a separate support route: for example, a self-study grammar and vocabulary module in the learning management system, individual consultations and a progress check after several weeks.

The item analysis (RQ4) suggests what each tier needs. The lower tier handled the functional items well. These students can manage routine conversational exchanges, so ESP work can start from spoken interaction on simple professional topics, provided that language support is built in. On grammar and lexis, however, this tier chose the correct

option less often than chance on 11 items. This pattern suggests that these students were drawn to specific wrong options, which points to consistent misconceptions rather than a simple lack of knowledge. Because the total score also defines the tiers, this interpretation needs confirmation from a distractor-level analysis. If it holds, explicit, form-focused teaching of the target structures is likely to help this tier more than further exposure alone.

The middle tier needs targeted work on a few points: the present perfect with *for*, *provided* as a conditional conjunction and the collocation *electricity supply*. The last item is the hardest in the test for the whole cohort. Collocation errors are known to persist even at advanced levels (Laufer & Waldman, 2011), and the influence of the first language on collocation use is considerable (Nesselhauf, 2003). Short focused tasks built on the students' disciplinary texts can address these points without slowing down the group.

The upper tier reached the ceiling of the test. The general English course has little new to offer these students. They are the natural audience for ESP extension work, such as reading and summarising research abstracts in their field, and they can also act as peer tutors. Teachers rarely offer such challenge to high achievers proactively, even though it tends to benefit them (Ziernwald et al., 2022).

The close relationship between grammar and lexis subscores means that most students' profiles were flat. Only one student in seven showed a marked gap between the two areas. In Tomlinson's (2014) terms, readiness level is the main basis for differentiation in this cohort, not skill profile. Tiering by overall score is therefore a sound starting point; differentiation by skill profile would matter only for a minority of students.

The findings also clarify what a short general test can and cannot do. With KR-20 = .88, the test ranks students reliably enough to form tiers. The SEM of about two items, however, means that students within one SEM of a tier boundary could belong to either tier. Tier membership should therefore stay flexible and be revised with classroom evidence. This agrees with Green and Weir's (2004) caution about the diagnostic use of quick placement tests and with the concern about misplacement in placement-validation research (Fulcher, 1997; Kokhan, 2013; Wall et al., 1994). The four weakly discriminating items suggest that a local version of the test could be shortened or revised. A short ESP-specific component would add information that the general test cannot give.

Using these results to form tiers in an ESP course requires a distinction between two constructs. The test measures general English proficiency: control of core grammar, everyday lexis and routine exchanges. Readiness to learn English for professional purposes is a broader construct. It also depends on disciplinary knowledge, subject vocabulary and familiarity with academic genres, which school English rarely provides (Hyland, 2022) and which the test did not measure. General proficiency is therefore treated here as a necessary but not sufficient condition of ESP readiness.

A student who scores below chance on core structures will struggle with professional texts, while a high score does not show that a student can already handle the genres of their discipline. The scores are thus valid for one decision only: how much language support a student needs to work with ESP materials. The tiers should therefore differ in scaffolding, not in professional content. Three features of the data support this limited use: reliable scores (KR-20 = .88), flat grammar–lexis profiles and clear item-level differences between the tiers. Because tier placement is low-stakes and revisable, these

conditions are sufficient for grouping. A judgement about ESP readiness itself requires an ESP-specific measure.

The study has several limitations. It covers one institute and 93 students, so the figures describe this cohort and should not be generalised to other faculties or universities without new data. The test measures general English, not the ESP the course teaches, and it is not linked to CEFR levels. Students took it online without supervision; outside help would most likely raise the scores of weaker students, so the true spread may be even wider than observed. The ceiling effect hides differences within the upper tier. No demographic or prior-education data were available to explain the spread, for example school type or results of the national English test. The item classification is the author's, and two items lie on the border between grammar and lexis. Finally, the design is cross-sectional: it shows the spread at entry, not how it changes.

A retest with the same 25 items is planned for the end of the 2026/2027 academic year. It will show whether tiered teaching narrows the spread and which tiers gain most. A distractor-level analysis and a short ESP vocabulary measure would add to the picture. Comparable analyses for other units of the university would show whether the pattern found here is general.

CONCLUSIONS

The study leads to four conclusions. First (RQ1), the cohort was strong on average but widely spread: scores ranged from 24 to 100, and each of the four bands contained between 5% and 36% of students. The placement test was reliable for this purpose (KR-20 = .88), with an SEM of about two items.

Second (RQ2), almost nine-tenths of the variance lay inside the groups. The groups were no more homogeneous than random allocation, and the spread was equally large in each of them. The academic group should therefore not be treated as a unit of language level.

Third (RQ3), level-based sets across programmes would remove most of the spread but would break the disciplinary basis of ESP and run against evidence on between-class grouping. In the simulation, three tiers within each group would halve the spread while keeping the group intact. This makes tiers a reasoned recommendation, not a proven solution. Near-chance students need a separate support route.

Fourth (RQ4), the tiers differ mainly in overall level, not in skill profile. The lower tier needs explicit work on core grammar and collocation, the middle tier needs targeted work on a few structures and collocations, and the upper tier needs ESP extension work beyond the reach of the test.

In practice, the Department can use the placement test to form flexible tiers inside each group. It can build a shared bank of tiered ESP tasks and offer a support module for students near the chance level. The planned end-of-year retest will show whether this approach narrows the spread.

CONFLICT OF INTEREST

The author declares that there are no conflicts of interest regarding the publication of this paper.

FUNDING

The author declares that this study received no specific financial support.

ARTIFICIAL INTELLIGENCE STATEMENT

During the preparation of this manuscript, the author used AI tools to organise the anonymised data, write the code for the statistical analysis, and search for references. No personal data were entered into the tool. After using this tool, the author reviewed and edited the content, checked all figures against the data and all references against the original sources, and takes full and sole responsibility for the content of the publication.

REFERENCES

Anthony, L. (2018). *Introducing English for specific purposes*. Routledge. https://doi.org/10.4324/9781351031189

Basturkmen, H. (2010). *Developing courses in English for specific purposes*. Palgrave Macmillan. https://doi.org/10.1057/9780230290518

Bolitho, R., & West, R. (2017). *The internationalisation of Ukrainian universities: The English language dimension*. British Council Ukraine; Stal. https://www.teachingenglish.org.uk/sites/teacheng/files/Pub-UKRAINE-REPORT-H5-EN.pdf

Bondie, R. S., Dahnke, C., & Zusho, A. (2019). How does changing "one-size-fits-all" to differentiated instruction affect teaching? *Review of Research in Education, 43*(1), 336–362. https://doi.org/10.3102/0091732X18821130

Cambridge University Press & Assessment. (n.d.). *Test your English: General English*. https://www.cambridgeenglish.org/test-your-english/general-english/

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume*. Council of Europe Publishing. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

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 practices in primary education. *Educational Research Review, 24*, 31–54. https://doi.org/10.1016/j.edurev.2018.02.002

Fabeku, O. E., & Fasanmi, S. A. (2024). Appraisal of transition challenges and adjustment of first-year students at a Nigerian university. *Educational Challenges, 29*(2), 80–96. https://doi.org/10.34142/2709-7986.2024.29.2.05

Fulcher, G. (1997). An English language placement test: Issues in reliability and validity. *Language Testing, 14*(2), 113–139. https://doi.org/10.1177/026553229701400201

Goundar, P. (2023). Examining Fijian first-year university students' English proficiency on CEFR levels. *Australian Journal of Applied Linguistics, 6*(2), 73–93. https://doi.org/10.29140/ajal.v6n2.1050

Green, A. B., & Weir, C. J. (2004). Can placement tests inform instructional decisions? *Language Testing, 21*(4), 467–494. https://doi.org/10.1191/0265532204lt293oa

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. *Educational Evaluation and Policy Analysis, 29*(1), 60–87. https://doi.org/10.3102/0162373707299706

Hughes, A., & Hughes, J. (2020). *Testing for language teachers* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781009024723

Huszti, I., & Lechner, I. (2025). Challenges of teaching multilingual students in tertiary education. *Educational Challenges, 30*(1), 106–118. https://doi.org/10.34142/2709-7986.2025.30.1.07

Hyland, K. (2022). English for specific purposes: What is it and where is it taking us? *ESP Today, 10*(2), 202–220. https://doi.org/10.18485/esptoday.2022.10.2.1

Kokhan, K. (2013). An argument against using standardized test scores for placement of international undergraduate students in English as a Second Language (ESL) courses. *Language Testing, 30*(4), 467–489. https://doi.org/10.1177/0265532213475782

Kononova, O. (2026). Digital benchmarking tools for assessment in maritime English training. *Educational Challenges, 31*(1), 179–191. https://doi.org/10.34142/2709-7986.2026.31.1.13

Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. *Psychometrika, 2*(3), 151–160. https://doi.org/10.1007/BF02288391

Laufer, B., & Waldman, T. (2011). Verb-noun collocations in second language writing: A corpus analysis of learners' English. *Language Learning, 61*(2), 647–672. https://doi.org/10.1111/j.1467-9922.2010.00621.x

Magableh, I. S. I., & Abdullah, A. (2021). The impact of differentiated instruction on students' reading comprehension attainment in mixed-ability classrooms. *Interchange, 52*(2), 255–272. https://doi.org/10.1007/s10780-021-09427-3

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223–242. https://doi.org/10.1093/applin/24.2.223

Pozas, M., Letzel, V., & Schneider, C. (2020). Teachers and differentiated instruction: Exploring differentiation practices to address student diversity. *Journal of Research in Special Educational Needs, 20*(3), 217–230. https://doi.org/10.1111/1471-3802.12481

Smale-Jacobse, A. E., Meijer, A., Helms-Lorenz, M., & Maulana, R. (2019). Differentiated instruction in secondary education: A systematic review of research evidence. *Frontiers in Psychology, 10*, Article 2366. https://doi.org/10.3389/fpsyg.2019.02366

Steenbergen-Hu, S., Makel, M. C., & Olszewski-Kubilius, P. (2016). What one hundred years of research says about the effects of ability grouping and acceleration on K–12 students' academic achievement: Findings of two second-order meta-analyses. *Review of Educational Research, 86*(4), 849–899. https://doi.org/10.3102/0034654316675417

Suprayogi, M. N., Valcke, M., & Godwin, R. (2017). Teachers and their implementation of differentiated instruction in the classroom. *Teaching and Teacher Education, 67*, 291–301. https://doi.org/10.1016/j.tate.2017.06.020

Symonenko, S., & Osadchyi, V. (2019). Peculiarities of English language training for electrical engineering students at Ukrainian universities. In *2019 IEEE International Conference on Modern Electrical and Energy Systems (MEES)* (pp. 394–397). IEEE. https://doi.org/10.1109/MEES.2019.8896541

Tarnopolsky, O. (2018). Constructivism in ESP teaching at Ukrainian universities. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1–7). Wiley. https://doi.org/10.1002/9781118784235.eelt0934

Tomczak, M., & Tomczak, E. (2014). The need to report effect size estimates revisited: An overview of some recommended measures of effect size. *Trends in Sport Sciences, 21*(1), 19–25. https://www.wbc.poznan.pl/Content/325867/PDF/5_Trends_Vol21_2014_%20no1_20.pdf

Tomlinson, C. A. (2014). *The differentiated classroom: Responding to the needs of all learners* (2nd ed.). ASCD.

Pro zastosuvannia anhliiskoi movy v Ukraini [On the use of the English language in Ukraine], Law of Ukraine No. 3760-IX (2024) (Ukraine). https://zakon.rada.gov.ua/laws/show/3760-20 [in Ukrainian].

Wall, D., Clapham, C., & Alderson, J. C. (1994). Evaluating a placement test. *Language Testing, 11*(3), 321–344. https://doi.org/10.1177/026553229401100305

Winke, P. M., & Rawal, H. (2018). Teaching large, mixed-ability classes. In J. I. Liontas (Ed.), *The TESOL encyclopedia of English language teaching* (pp. 1–6). Wiley. https://doi.org/10.1002/9781118784235.eelt0222

Ziernwald, L., Hillmayr, D., & Holzberger, D. (2022). Promoting high-achieving students through differentiated instruction in mixed-ability classrooms: A systematic review. *Journal of Advanced Academics, 33*(4), 540–573. https://doi.org/10.1177/1932202X221112931