IB Biology (subject-focused) Extended Essay (EE)

Secondary Data Investigations (Literature & Data-based Research)

Complete Section-by-Section Guide to score 30/30

Based on: IB Extended Essay Guide (first assessment 2027) | EE Support Material | EE Assessment Criteria

Written by Peter Marier and edited with Claude (Anthropic, 2026)

HOW TO USE THIS GUIDE

This guide is organized by essay section, in the order they must appear in your Extended Essay. Each section contains:

- A description of what belongs in each section and why
- Required elements for the top markband
- The IB criterion each section addresses
- Common mistakes to avoid
- Formatting suggestions
- Notes on key differences from the Primary Data Investigations Guide and Biology IA (where relevant)

This guide is specifically for secondary data investigations (essays based on data obtained from published literature, databases, or existing studies). If you are collecting your own data through experimentation, use the ***Primary Data Investigations Guide*** *instead.*

⚠ AVOID / COMMON MISTAKE

AN ESSAY, NOT A LAB REPORT – Read this first!

The extended essay is an essay, not a lab report. It must be presented as continuous prose (subheadings still used). Your evaluation should be woven into the discussion – not placed in a separate section after the conclusion. The conclusion must come last (before reference list). These structural expectations are fundamentally different from the Biology IA.

For secondary data EEs specifically: the essay must go beyond summarising existing studies. You must conduct your OWN original analysis of extracted data. Essays that simply restate facts or data taken directly from sources score poorly. The IB explicitly states: “essays that simply restate facts or data taken directly from the sources are of little value.”

Table of Contents

ASSESSMENT CRITERIA OVERVIEW

The EE is marked out of 30 across five criteria. Examiners use a best-fit approach and mark positively. The top markband does not require a flawless essay.

Criterion	Strands	Top Band Descriptor
A: Framework (6 marks)	Research question Research methods Structure	RQ is relevant, clear and focused in relation to the scope of the essay. Methods are suitable, explained and applied effectively. Structural conventions effectively support communication of the research.
B: Knowledge (6 marks)	Knowledge Terminology Concepts	Comprehensive, relevant research materials establish knowledge of subject matter. Terminology is accurate and consistent. Concepts are explained and used effectively to demonstrate understanding.
C: Analysis & Argument (6 marks)	Analysis Line of argument	Analysis is effective and consistently produces relevant findings. A clear, sustained line of argument links the RQ, research findings and conclusions.
D: Discussion & Evaluation (8 marks)	Discussion Evaluation	A balanced discussion of the significance of the findings is fully supported by appropriate evidence. Strengths and limitations are explained.
E: Reflection (4 marks)	Evaluative Growth	Consistently evaluative with specific examples. Consistent evidence of the student's growth and transfer of learning.

1. TITLE PAGE

While the title page is not assessed, it sets the professional tone. It is a required structural convention.

Required elements

- **Research question** – stated in full, exactly as it appears throughout the essay
- **Student code** – no personal name, supervisor name, or school name anywhere in the essay (anonymity is mandatory)
- **Subject: Biology** – clearly stated as the DP subject the essay relates to
- **Word count** – the total word count (max 4,000). If you used explanatory footnotes, state: "The stated word count includes explanatory footnotes."

✅ **FOR TOP MARKS:** A professional, anonymous title page with all four required elements supports readability and demonstrates awareness of structural conventions (Criterion A – Structure strand).

⚠ **AVOID / COMMON MISTAKE**

Do NOT include a separate essay title or topic line. The 2027 guide requires only the four elements above – the research question serves as the title.

2. CONTENTS PAGE

Include a clear contents page listing all major sections with accurate page numbers. Not included in the word count. Active readers use the contents page to predict the direction of the essay – a good contents page acts as a signpost for the examiner.

FORMATTING TIP

Word can generate an automatic table of contents that automatically updates headings with page numbers with page links

1. Create a table by selecting 'References' > 'Table of Contents'
2. Format headings: "heading 1", subheading "heading 2", sub-sub heading "heading 3" etc.

** Table should automatically update the headings and page # but if they didn't change, select the table of contents and press fn + F9 "update entire table"*

3. INTRODUCTION (Criteria A, B, C)

The introduction tells the reader what to expect: the focus of the essay, the scope of the research, the sources to be used, and an insight into the line of argument. It is advisable to revisit and edit the introduction once the body of the essay is complete, to ensure it accurately introduces the essay as written.

HOW THE SECONDARY DATA EE INTRODUCTION DIFFERS FROM THE IA AND PRIMARY DATA EE

The core content is the same across all three: the IV–DV (or variable) relationship explained with biological theory, peer-reviewed sources cited, key terminology defined, and detailed biological processes described.

Compared to the Biology IA:

- **Broader literature scope** – the IA focuses narrowly on the IV–DV link; the EE must also survey the wider field (what has been established, what is debated, where your investigation fits) to demonstrate knowledge "in the wider context of the relevant discipline" (Criterion B)
- **Wider context required** – connect the topic to broader biological, ecological, medical, or agricultural significance
- **Line of argument previewed** – the introduction must give "an insight into the line of argument" (Criterion C), which is not required in the IA

Compared to the Primary Data EE:

- **Study organism may or may not apply** – a secondary data EE can focus on a single species (e.g. synthesising published data on *Apis mellifera*), a group of species (e.g. comparing metabolic rates across ectothermic reptiles), or a broader biological system (e.g. coral reef bleaching events). If a specific organism is the focus, introduce it as you would in a primary data EE (binomial nomenclature, italicised, suitability justified). If the investigation spans multiple species or systems, introduce the broader biological context instead
- **Emphasis on the research gap** – the introduction must clearly articulate why existing individual studies are insufficient and why a synthesis, comparison, or re-analysis of secondary data is needed to address the RQ
- **Methodological preview differs** – instead of previewing an experimental approach, explain the analytical approach (e.g. "this essay will compare data from X studies using Y statistical method to determine…")

tl;dr *Secondary data introduction = IA background + broader literature review + preview of line of argument + clear justification for WHY a secondary data approach is appropriate + preview of the analytical method.*

3.1 Context, Background, and Literature Review

This is the foundation of the essay. For a secondary-data biology EE, the literature review is even more critical than for a primary-data EE because the published literature IS your data source. You must demonstrate comprehensive knowledge of both the biological processes and the existing body of research.

Required elements

- **Specific biological context** – explain the biological system and process(es) directly relevant to your research question with enough depth for the reader to interpret your later analysis
- **Relationship between variables established** – use biological theory and cited literature to explain WHY the variables/factors you are investigating are expected to be related
- **Broader significance** – connect the topic to a wider biological, ecological, medical, or agricultural context – demonstrate knowledge "in the wider context of the relevant discipline"
- **Comprehensive literature review** – survey the current state of research on your topic. What has been established? What is uncertain or debated? What gap does your synthesis address? For a secondary data EE, this section must be substantially more thorough than for a primary data EE because it forms the foundation for your entire analysis
- **Justification for secondary data approach** – explain why a synthesis/comparison of existing data is the appropriate method for your RQ (e.g. the question requires data across a larger scale than a student could collect, multiple geographic regions, longitudinal data spanning years/decades, or comparison across species)

- **Study organism or system introduced** – if the essay focuses on a specific species, name it (binomial nomenclature, italicised) and briefly explain why it is a suitable subject for secondary data analysis (e.g. well-studied, extensive published data available, ecological or medical importance). If the essay spans multiple species or a broader system, introduce the system or taxonomic group and justify the scope
- **Peer-reviewed sources only** – cite academic journal articles and textbooks throughout. Do NOT use blogs, Wikipedia, or revision websites
- **Key terminology defined** – define technical or subject-specific terms when first used, otherwise it comes across as jargon
- **Figures where helpful** – include a diagram or image (with Figure caption and in-text reference) if it helps explain a process or the system being investigated

✅ **FOR TOP MARKS:** Criterion B (Knowledge) top band: "Comprehensive, relevant research materials are used to establish knowledge of the subject matter." Criterion B (Concepts) top band: "Relevant concepts are explained and used effectively to demonstrate understanding."

For secondary data EEs, the top band for Criterion B is only achievable if the student demonstrates mastery of the relevant biology AND critical awareness of the existing research landscape.

3.2 Research Question

The research question is the central focus of the entire essay. For secondary data EEs, the RQ must be answerable through analysis of existing data and must be specific enough to allow a focused, systematic investigation.

Required elements

- **Variables/factors clearly stated** – identify the specific biological variables or factors being compared, with units where applicable
- **Study organism or system named** – if a single species, use common name (*Species name*) in the RQ. If multiple species or a broader system, name the taxonomic group or system (e.g. "C3 plant species", "coral reefs in the Great Barrier Reef")
- **Scope defined** – the RQ must make clear the taxonomic group, time period, geographic range, or other boundaries of the data being analysed
- **Answerable through secondary data** – the RQ must be one that can be systematically addressed by collecting, comparing, or re-analysing published data
- **Consistent wording** – copy-paste the RQ; needs to be written identically everywhere in the essay

📋 **EXAMPLE RQ FORMATS FOR SECONDARY DATA**

- **Comparative:** "To what extent does [environmental factor] affect [measured variable] of [taxonomic group] across [geographic scope], based on published ecological surveys from [year range]?"
- **Trend/relationship:** "What is the relationship between [environmental variable] and [biological response variable] in [ecosystem/region], as reported in published monitoring data from [year range]?"
- **Database analysis:** "How does the frequency of [mutation type] in the [gene name] gene vary across [comparison groups], based on data from the [database name(s)]?"
- **Meta-analysis:** "What is the overall effect of [treatment/condition] on [dependent variable] across [organism group], based on a synthesis of published experimental data?"

✅ **FOR TOP MARKS:** Criterion A (Research question) top band: "relevant to the topic of investigation, clear and focused in relation to the scope of the essay." Including specific IV values and organism name achieves this focus.

⚠ **AVOID / COMMON MISTAKE**

Do NOT write the RQ as a yes/no question – use sentence starters like 'to what extent', 'what is the relationship between', or 'how does X compare to Y'.

Do NOT write a RQ so broad that it becomes a literature review rather than an investigation (e.g. "What causes cancer?"). The RQ must be narrow enough to allow systematic data extraction and original analysis.

3.3 Hypothesis

A hypothesis provides a specific, falsifiable prediction and helps situate the RQ within biological theory. For secondary data EEs, the hypothesis predicts what the analysis of existing data will reveal.

Required elements

- **Specific falsifiable prediction** – state the expected outcome of your data analysis (e.g. "It is hypothesised that [dependent variable] will show a significant [positive/negative] correlation with [independent variable] across [organism group]")
- **Biological justification and reasoning** – explain WHY you predict this outcome using theory from the background/past studies
- **Predictive graph (recommended)** – simple graph showing the hypothesised relationship; label axes with units

3.4 Scope and Line of Argument

Brief outline of the experimental approach that will be taken and the direction your argument will follow, so the reader finishes the introduction knowing exactly what the essay will cover and how it will be structured. It's essentially the "roadmap" paragraph

Required elements

- **Outline the approach** – briefly describe what the essay will investigate and how
- **Preview the argument** – give the reader an insight into the line of argument

4. METHODOLOGY (Criterion A)

This section must demonstrate that your approach to collecting and selecting secondary data was systematic, rigorous, and appropriate for your RQ. This is the section that replaces the experimental methodology in a primary data EE. It is critically important because secondary data EEs are often scrutinised more heavily by moderators – you must clearly demonstrate that your data selection was not cherry-picked, biased, or superficial.

HOW THE SECONDARY DATA EE METHODOLOGY DIFFERS FROM THE IA AND PRIMARY DATA EE

Compared to the Biology IA:

In the IA, the methodology is structured with separate labelled subsections (IV, DV, CVs, materials list, safety table, step-by-step procedure). In the EE, this information must be presented as continuous essay prose rather than standalone tables or numbered lists. Standalone equipment/materials lists can be placed in an appendix – but examiners are not required to read appendices.

Compared to the Primary Data EE:

In a primary data EE, the methodology describes the experimental procedure (equipment, concentrations, timing, etc.). In a secondary data EE, the methodology describes HOW you systematically identified, selected, evaluated, and extracted data from published sources. The content is fundamentally different but the rigour must be equivalent:

- **Instead of experimental setup** → describe your search strategy and databases used
- **Instead of controlled variables** → describe your inclusion/exclusion criteria (how you controlled for comparability and standardized formats/data values)
- **Instead of a pilot experiment** → describe a preliminary source review that informed your final approach
- **Instead of sample size (trials)** → describe the number of studies/datasets and justify sufficiency
- **Instead of risk assessment/safety** → address data use permissions, attribution, and ethical considerations

tl;dr: *A primary data IA and EE both describe a hands-on experiment – only the format differs (tables vs prose). The secondary EE replaces the entire experimental narrative with a data-sourcing narrative. The same level of detail and rigour is expected: the methodology must be detailed enough that another researcher could replicate your data collection process and arrive at the same dataset.*

4.1 Research Variables / Factors Investigated

Even though you are not manipulating variables in a laboratory, you must clearly define the factors being investigated.

Required elements

- **Independent variable (or explanatory factor)** – the factor that varies across the studies/datasets you are comparing
- **Dependent variable (or response variable)** – the biological outcome being measured or compared
- **Confounding variables identified** – identify factors that may differ between source studies and could affect comparability (e.g. different measurement methods, different species/strains, geographic variation, time of year, sample sizes). For each, explain the potential biological impact on the DV if left uncontrolled. The strategy for minimising these is described in Section 4.4
- **SI units throughout** – SI (International System of Units) must be used throughout. Note whether source studies report the DV and IV in consistent units or whether conversion will be required. The standardisation strategy for handling unit differences is described in Section 4.4.

✓ **FOR TOP MARKS:** Criterion A (Research methods) top band: methods are "suitable for the research question" and "applied effectively." For secondary data, this means your data selection strategy must be clearly appropriate for answering the RQ.

⚠ **AVOID / COMMON MISTAKE**

Do NOT assume that because you are using published data, you do not need to define variables. Moderators will look for this. Do NOT present variables as standalone tables in the body of the essay. Describe and justify them within the narrative prose.

4.2 Search Strategy and Source Selection

A continuous narrative describing how you identified, screened, and selected your data sources. This is the equivalent of the experimental procedure for primary data — it must be detailed enough for replication. Begin by describing any initial scoping searches that informed your approach, then present the final systematic search.

The standard approach to secondary data collection follows four stages. These are the same four stages used in published systematic reviews and meta-analyses across biology, medicine, ecology, and all sciences:

1. **Identification** – search multiple databases using defined search terms and record the total number of results found. When you search more than one database (e.g. PubMed AND Google Scholar AND Web of Science), the same paper will often appear in multiple databases. These duplicates are removed so that no paper is counted or screened twice. This is not an error – it is an expected and standard consequence of searching broadly.
2. **Screening** – read only the title and abstract of each remaining record (not the full text) to quickly determine whether it is potentially relevant to your RQ. This is a time-efficient first filter. For example, a search might return 200 results, but from titles and abstracts alone you can quickly eliminate papers that are clearly about a different organism, a different variable, or an unrelated context. Only records that pass this quick filter proceed to full-text reading.
3. **Eligibility** – read the full text of each remaining article and assess whether it meets all of your inclusion/exclusion criteria (described in Section 4.3). This is where you check for extractable raw data, compatible methodology, sufficient sample size, appropriate date range, and any other criteria. Record the number excluded and the specific reason for each exclusion.
4. **Included** – the final set of studies from which you will extract data for your analysis. Record the number of studies and the total number of individual data points they provide.

Required elements

- **Written as continuous prose with subheadings** – not a numbered list
- **Initial scoping described** – briefly describe any preliminary searches that informed your final approach: what databases were initially explored, whether sufficient studies existed, whether the RQ needed narrowing or broadening, and which databases proved most productive. This demonstrates that the student thought

critically about the methodology before committing to data collection (equivalent to a pilot experiment for primary data)

- **Databases identified** – name every database searched (e.g. PubMed, Google Scholar, Web of Science, NCBI GenBank, UniProt, ClinVar, GBIF, IUCN Red List, FAO, WHO, NOAA, OBIS, etc.) with URLs and access dates
- **Search terms listed** – provide the exact search terms, Boolean operators, and filters used (e.g. "([DV term A] OR [DV term B]) AND ([IV term A] OR [IV term B]) AND ([organism group])")
- **Date range specified** – state the publication date range of studies included and justify it
- **Number of sources at each stage** – report how many results each search returned, how many were screened, and how many were included in the final dataset. A PRISMA-style flow diagram is *strongly recommended* (see model diagram after Section 4.3)
- **Detailed enough for replication** – a reader should be able to reproduce your search and arrive at a comparable dataset

✓ **FOR TOP MARKS:** *Criterion A (Research methods) top band: "Research methods that are suitable for the research question are explained and applied effectively."* A top-scoring search strategy reads as a transparent, systematic narrative. The examiner should understand exactly how you went from an initial database search to a final, curated dataset – with numbers at every stage. Describing how initial scoping searches informed the final methodology demonstrates intellectual honesty and methodological awareness, exactly as a pilot experiment does for primary data.

4.3 Inclusion and Exclusion Criteria

Defining clear criteria for which studies/datasets are included or excluded is essential for methodological transparency and is the secondary-data equivalent of controlling variables.

Required elements

- **Inclusion criteria** – list and justify each criterion (e.g. peer-reviewed only, specific organism/taxon, specific measurement method, minimum sample size, English language, specific time period, specific geographic range, etc.)
- **Exclusion criteria** – list and justify each (e.g. studies without raw data excluded, studies using incompatible methodology excluded, etc.). Decide upfront whether studies missing key values are included or excluded, e.g. a study reports the mean DV but not the SD, or does not state the exact sample size, will it be excluded or included with a noted limitation? State and justify this decision here.
- **Biological justification for each criterion** – explain WHY each criterion matters for the validity of your analysis
- **Confounding variables addressed** – for each confounding variable identified in Section 4.1, explain which specific inclusion or exclusion criterion controls for it. This is the secondary-data equivalent of describing how each controlled variable was held constant in a primary data EE

✓ **FOR TOP MARKS:** Clear inclusion/exclusion criteria demonstrate that your data selection was systematic and unbiased. This is the secondary-data equivalent of controlled variables and directly addresses Criterion A (Research methods) and Criterion D (Evaluation).

⚠ **AVOID / COMMON MISTAKE**

Do NOT simply state "reliable sources were selected." You must explain WHAT made them reliable and HOW you determined this. Vague criteria are treated the same as uncontrolled variables in a primary data EE.

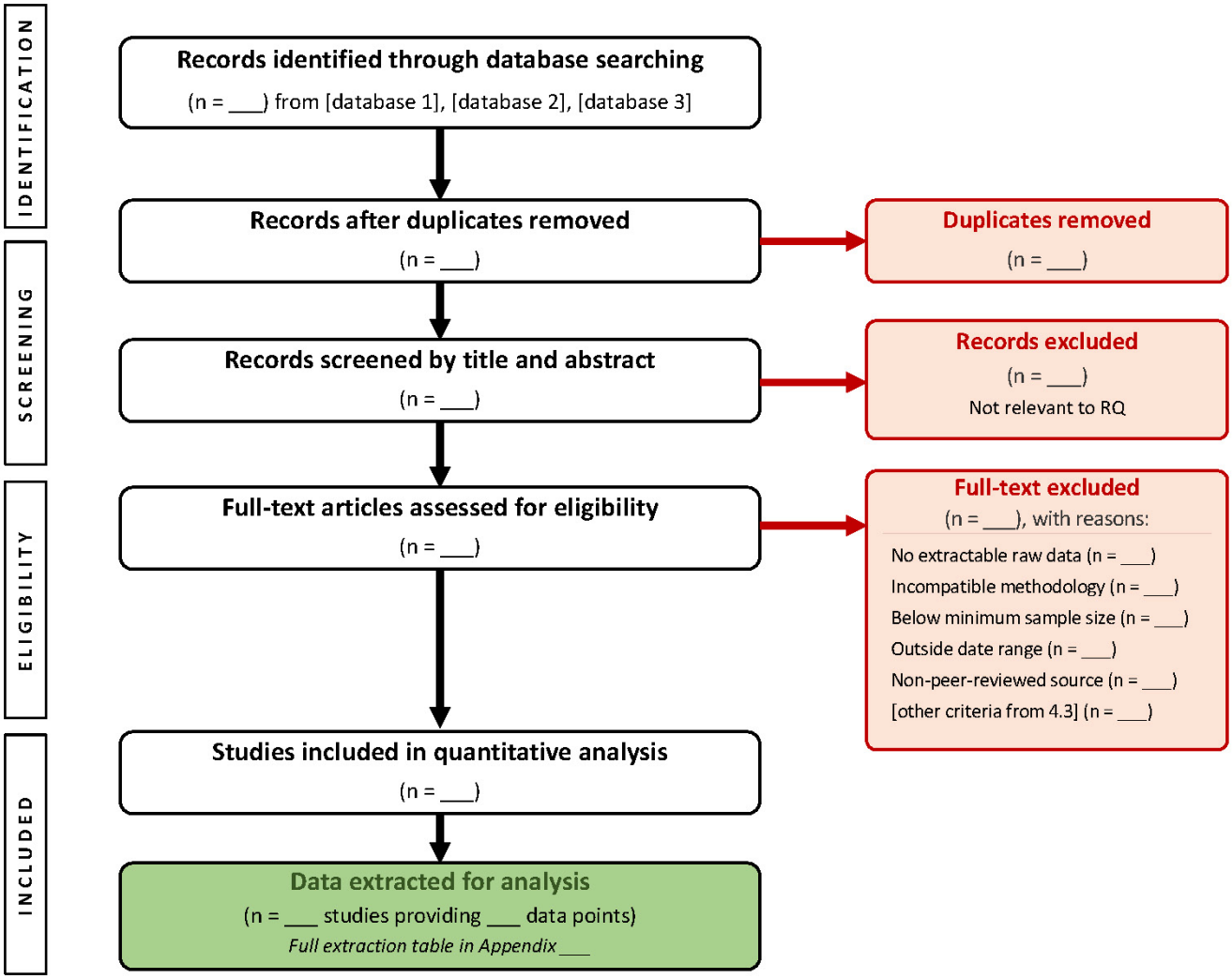
RECOMMENDED: PRISMA-STYLE FLOW DIAGRAM

PRISMA = ***Preferred Reporting Items for Systematic Reviews and Meta-Analyses*** (Page et al., 2021). It is the internationally recognised standard for transparently reporting how studies were identified, screened, and selected in any research that synthesises data from multiple published sources. The PRISMA flow diagram is the most recognisable component and is required in virtually all published systematic reviews and meta-analyses across biology, medicine, ecology, and psychology.

Including a PRISMA-style flow diagram in your EE immediately signals methodological rigour to the examiner and demonstrates awareness of standard academic research conventions. Include it as a numbered Figure in the methodology with a detailed caption and in-text reference (e.g. "see Figure 1").

Use the model below as a template:

- Replace every (n = ___) with your actual numbers at each stage
- Replace [database 1], [database 2], etc. with the actual databases you searched
- Replace the exclusion reasons with your actual inclusion/exclusion criteria from Section 4.3
- The numbers must add up: records identified – duplicates – title/abstract exclusions – full-text exclusions = studies included



4.4 Data Standardisation

Different source studies will almost certainly report data in different formats, units, or scales. Before extracting data, you must decide HOW you will standardise it so that values from different studies are directly comparable. This is a methodological decision – it must be planned and described here, not introduced for the first time in the results.

Required elements

- **Unit conversions identified** – if sources report the DV in different units state which standard unit you will convert to and show the conversion method
- **Normalisation method described** – if raw values are not directly comparable due to differences in baseline conditions (e.g. different control group values, different body sizes across species), explain how data will be normalised (e.g. percentage change from control, per-unit-mass values, etc.)
- **Handling of variation in reported statistics** – explain what you will do when sources report different summary statistics (e.g. some report mean ± SD, others report median and IQR, others report only range). State whether you will exclude incompatible data or apply a conversion method (with citation)
- **Weighting strategy** – if combining data across studies with different sample sizes, explain whether and how values will be weighted (e.g. weighted mean by sample size). Larger studies (with more data and therefore more statistical reliability) should contribute proportionally more to pooled estimates than smaller studies. The simplest appropriate method is weighting by sample size (see worked example below). State the formula used and justify your choice

WEIGHING BY SAMPLE SIZE – WORKED EXAMPLE

Formula: Weighted mean = Σ(each study's mean × its sample size) ÷ Σ(all sample sizes)

Worked example:

Suppose three studies report the mean value of a DV:
Study A: mean = 24.2, n = 200
Study B: mean = 31.5, n = 15
Study C: mean = 22.8, n = 85

Unweighted mean (treats all studies equally):
(24.2 + 31.5 + 22.8) ÷ 3 = 26.2

Weighted mean (accounts for sample size):
(24.2 × 200 + 31.5 × 15 + 22.8 × 85) ÷ (200 + 15 + 85)
= (4840 + 472.5 + 1938) ÷ 300
= 7250.5 ÷ 300 = 24.2

Notice how Study B (n = 15) pulled the unweighted mean up to 26.2, despite being far less reliable than Study A (n = 200). The weighted mean of 24.2 better reflects the overall body of evidence because Study A's larger, more reliable dataset has proportionally more influence.

FOR TOP MARKS: A clearly described standardisation strategy demonstrates methodological rigour and shows the examiner that your consolidated dataset is valid for comparison. This directly supports Criterion A (Research methods) — methods are "suitable" and "applied effectively" and also strengthens Criterion D (Evaluation) by pre-emptively addressing a common source of error.

⚠ AVOID / COMMON MISTAKE

Do NOT wait until the Results section to explain how you standardised data – describe the plan here in the methodology, then present it in the results.

Do NOT silently combine data reported in different units or formats without explaining the conversion.

Do NOT ignore studies that report data differently from the majority – either convert or exclude with justification,

4.5 Ethical Considerations

While secondary data EEs do not involve laboratory safety hazards, or environmental considerations, ethical considerations still apply.

Required elements

- **Data use permissions** – confirm that all databases used are publicly accessible or that appropriate permissions were obtained for restricted datasets
- **Proper attribution** – all data must be cited to its original source; presenting others' data as your own is academic dishonesty
- **Human/animal data ethics** – if using data originally collected from human participants or animals, briefly note that the original studies followed ethical guidelines
- **Potential for harm** – consider whether your conclusions could be misused or misinterpreted (e.g. genetic data, disease data)

✅ **FOR TOP MARKS:** Criterion A (Research methods) top band: "Research methods that are suitable for the research question are explained and applied effectively." A top-scoring methodology reads as a systematic, transparent, replicable narrative. The examiner should understand exactly how you built your dataset.

⚠ AVOID / COMMON MISTAKE

Do NOT present the methodology as a numbered list of steps – this is an essay, not a lab report.

Do NOT assume that because you didn't conduct an experiment, you don't need a rigorous methodology. This is the most common reason secondary data EEs score poorly on Criterion A.

5. TYPES OF SECONDARY DATA INVESTIGATIONS

Secondary data biology EEs can take several forms. Understanding which type your essay falls into helps you structure your methodology and analysis appropriately. In all cases, the student must conduct ORIGINAL analysis — not simply report what others have found.

TYPE 1: SYNTHESIS OF DATA FROM MULTIPLE PUBLISHED STUDIES

Description: Extract quantitative data from multiple peer-reviewed studies that investigated a similar question and combine/compare them systematically.

Example RQ: *"What is the overall effect of [chemical/treatment] on [biological response] in [organism], based on a synthesis of published field studies from [year range]?"*

Data sources: Peer-reviewed journal articles accessed via PubMed, Google Scholar, Web of Science

Recommendations: source selection must be "sufficiently wide and reliable". Clear inclusion/exclusion criteria; data extraction table; standardisation of units and measures across studies; original statistical analysis

TYPE 2: DATABASE ANALYSIS

Description: Extract raw or processed data directly from large publicly accessible scientific databases and conduct original analysis on the dataset.

Example RQ: *"How does the frequency of [mutation type] in the [gene name] gene differ between [condition A], [condition B], and [condition C], based on data from the [database name]?"*

Data sources: NCBI GenBank, UniProt, ClinVar, gnomAD, COSMIC, GBIF, IUCN Red List, NOAA Coral Reef Watch, FAO, WHO, OBIS, Tree of Life Web Project, EBI, Ensembl, PDB, etc.

Recommendations: Database clearly described with access dates and version numbers; data extraction methodology explicit; search/filter parameters documented; original statistical analysis performed on the extracted dataset; limitations of the database acknowledged

TYPE 3: COMPARATIVE ANALYSIS OF PUBLISHED DATA

Description: Collect published data from studies that investigated different conditions, species, or populations and create an original comparison not present in any single source.

Example RQ: *"To what extent does the [physiological variable] of [taxonomic group] correlate with [environmental/geographic factor], based on published data?"*

Data sources: Peer-reviewed studies reporting comparable data on different species/populations/conditions

Recommendations: The comparison must be ORIGINAL – the student is creating a new dataset by bringing together data that has not previously been combined in this way; units standardised across studies; confounding variables addressed

TYPE 4: TEMPORAL TREND ANALYSIS FROM MONITORING DATA

Description: Use long-term monitoring datasets to analyse temporal trends in a biological variable.

Example RQ: *"What is the relationship between [abiotic variable] and the frequency of [biological event] in [ecosystem/region] from [start year] to [end year], based on [database/institution] monitoring data?"*

Data sources: Government monitoring databases (NOAA, EPA, national biodiversity databases), institutional long-term datasets (AIMS, LTER), WHO/FAO global health/agriculture data

Recommendations: Clear justification for the time period selected; data completeness and gaps addressed; original statistical analysis of trends (regression, correlation); confounding temporal factors acknowledged

INVESTIGATIONS CAN COMBINE TYPES

Real investigations often do not fit neatly into a single type. For example, a student investigating the relationship between [variable A] and [variable B] in [organism] across two environmental conditions (e.g. years with and without a specific event) could be doing a comparative analysis (Type 3) of a correlation, using data extracted from a long-term monitoring database (Type 2), split by a temporal condition (Type 4).

This is perfectly normal and acceptable. When your investigation spans types:

- **Identify the primary type** — ask: what is the main analytical action? If you are comparing groups, it is primarily Type 3. If you are analysing a trend over time, it is primarily Type 4. If you are pulling data from a single database, it is primarily Type 2. If you are pooling results from many separate studies, it is primarily Type 1
- **Use that type's table structure as your starting point** — then adapt by adding columns from other types as needed (e.g. add a "Year" column to a Type 3 table if the comparison involves a temporal element)

⚠ AVOID / COMMON MISTAKE

CRITICAL FOR ALL TYPES: Regardless of the type, a secondary data EE must demonstrate ORIGINAL ANALYSIS. The student must extract data, reorganise it, process it statistically, create original graphs/tables, and argue their own conclusions. An essay that merely summarises what each study found is a literature review, NOT a secondary data investigation.

6. RESULTS AND DATA PROCESSING (Criteria A, C)

Present all extracted data clearly and systematically, then process and analyse it using appropriate quantitative methods. This section is where you demonstrate that you have done MORE than simply read the literature — you have created an original, consolidated dataset and analysed it yourself.

HOW SECONDARY DATA EE RESULTS DIFFER FROM THE IA AND PRIMARY DATA EE

Compared to the Biology IA:

The data presentation standards (tables, processed data, statistics, graphs) are essentially the same as the IA — same rigour, same formatting expectations. The only structural difference is that in the IA, the interpretation happens in a separate Criterion C section (Analysis → Conclusion). In the EE, the data is presented here and the interpretation is woven into the Discussion.

Compared to the Primary Data EE:

In a primary data EE, this section presents raw experimental data (individual trial values), processes it (means, SD), and applies statistical tests. In a secondary data EE, the equivalent section presents:

- **A data extraction summary table** (equivalent to "raw data") showing key values extracted from each source
- **Processed/standardised data** (recalculated means, converted units, weighted averages)
- **Statistical analyses** performed on your consolidated dataset
- **Original graphs** created from the extracted data (NOT screenshots from source papers)

***Tl;dr**: The rigour and formatting expectations (captions, units, uncertainties, graph standards) are identical across all three. The IA and primary EE present student-collected trial data; the secondary EE presents student-extracted and student-consolidated published data. In all cases, the student must show their OWN processing and analysis.*

6.1 Data Extraction and Organization

This is the equivalent of "raw data" for secondary data EEs. Present the key quantitative data extracted from each source in a structured, transparent format.

Required elements

- **Detailed table caption** – format: 'Table X – Summary of extracted data on [DV] across [number] studies investigating [topic]. Data extracted on [date].' Include definitions of any abbreviations used
- **Column headers** – include: Source (author, year), Sample size (n), IV value/category, DV value (mean ± SD where available), measurement method, and any relevant notes
- **Unit standardisation applied** – All values in this table should already be standardised to consistent units and decimal places as described in Section 4.4. If any values were converted, indicate this in the table caption.
- **If dataset is large** – include a representative summary in the main body and the full extraction table in an appendix (with a reference statement: "Full data extraction table available in Appendix A")

Choose a table format below that best matches your type of investigation (see Section 5). Adapt columns as needed

EXAMPLE DATA EXTRACTION TABLE – TYPE 1: SYNTHESIS FROM MULTIPLE PUBLISHED STUDIES

Table 1 – Summary of extracted data on [dependent variable] ([unit]) of [organism/group] exposed to [treatment/condition], from [number] published studies ([year range]). All values standardised to [unit] as described in Section 4.4.

Source	[IV] (unit)	Species	n	[DV] (unit ± SD)	Study Design	Notes
Author A (year)						
Author B (year)						
…						

EXAMPLE DATA EXTRACTION TABLE – TYPE 2: DATABASE ANALYSIS

Table 2 – Summary of [dependent variable] ([unit]) of [gene/protein/organism] grouped by [independent variable], extracted from [database name] (version ___, accessed [date]). All values standardised to [unit] as described in Section 4.4.

Record ID	[IV]	n	[DV] (unit ± SD)	Notes
ID_001				
ID_002				
…				

Note: for database extractions the Record ID is the unique identifier the database assigns to each entry

EXAMPLE DATA EXTRACTION TABLE – TYPE 3: COMPARATIVE ANALYSIS

Table 3 – Comparison of [dependent variable] ([unit]) across [number] species/populations/conditions, compiled from [number] published studies ([year range]). All values standardised to [unit] as described in Section 4.4.

Source	[IV]	Species/population	n	[DV] (unit ± SD)	Location
Author A (year)	Species X				
Author B (year)	Species Y				
…	…				

EXAMPLE DATA EXTRACTION TABLE – TYPE 4: TEMPORAL TREND ANALYSIS

Table 4 – [Dependent variable] ([unit]) in [ecosystem/region] over [time period], extracted from [database/institution] monitoring data (accessed [date]).

[IV] year/time point	n	[DV] (unit ± SD)	[Abiotic variable] (unit)	Data source	Notes
Year 1					
Year 2					
…					

6.2 Processed Data

Show all calculations performed on the extracted data, explain why and how each calculation was done, and present the results in a clearly formatted table.

Required elements

- **Description and justification of each calculation** – for each measure (weighted mean, pooled standard deviation, percentage change, rate, etc.), explain WHAT it shows and WHY it was performed
- **Method stated** – specify software and version (e.g., 'Data were analysed using Microsoft Excel for Mac (Version 16.107)') OR show a sample calculation with correct formula and working
- **Weighting and normalisation applied** – if combining data across studies, apply the weighting and/or normalisation strategy described in Section 4.4. Show sample calculations with correct formulae and working (unit conversions should already be reflected in the data extraction table in Section 6.1)
- **Processed data table** – with detailed caption; columns for factor/category, weighted or pooled mean, SD, n, and any derived measures

✓ **FOR TOP MARKS:** "Results must be presented in a standardized format… clearly labelled with appropriate headings, units and numbering." "Examples of calculations should include mathematical uncertainties on the measurements collected." This contributes to Criterion A (Structure). Criterion C (Analysis) top band: "Analysis in the essay is effective and consistently produces relevant findings." The sciences guidance: analysis "must follow standard processes, including qualitative and quantitative approaches and statistical methods where appropriate, and may include mathematical transformation."

6.3 Statistical Analyses

Select and apply statistical tests that are appropriate for your data type, sample size, and the nature of your secondary dataset. Justify every test choice. The statistical demands for secondary data EEs differ somewhat from primary data EEs because you are often working with summary statistics rather than raw trial-level data.

HOW STATISTICAL DEMANDS MAY DIFFER FOR SECONDARY DATA

The fundamental requirement is the same: you must use appropriate statistical tests and justify every choice. However, the nature of the data often differs:

- **You may only have access to summary statistics** (means, SDs, sample sizes) rather than raw data points. This limits which tests can be applied but does NOT excuse you from statistical analysis
- **Weighting by sample size expected** – larger studies should contribute more to pooled estimates than smaller ones

***tl;dr**: The level of statistical sophistication should match the data available. Do not force tests that require raw data when you only have summary statistics.*

LINKED RESOURCES

Consult the following flowcharts to help choose appropriate statistical test (links to online calculators embedded):
If using Raw Data → STATS FLOWCHART – RAW DATA
If using reported means → STATS FLOWCHART – REPORTED MEANS

Required elements

- **Outlier test** – check for statistical outliers in the raw data using Q_1, Q_3 and IQR. Do not remove outliers – flag them in the raw data table and consider presenting results both with and without the outlier(s). If working with reported means, check whether any study's value is an outlier relative to other studies.
- **Normality test** – Shapiro-Wilk test to check normality of each IV group's data (prerequisite for parametric tests)
- **Choice of main test justified** – based on normality, data type, and sample size (*see Stats Flowchart*)
- **Post-hoc test if $p < \alpha$** – Tukey's HSD (following ANOVA) or Dunn's test (following Kruskal-Wallis); present pairwise comparisons
- **Correlation test** – if IV is continuous: Pearson's (parametric) or Spearman's (non-parametric); include H_o, H_a, r, R^2, p-value, and inference
- **Null and alternative hypotheses stated** – for each statistical test
- **Online calculator acknowledged** – if used, state the calculator name/URL; this does not replace justification of the test choice

EXAMPLE TABLE FORMAT

e.g. format for statistical tests (also include table caption with additional information and selected α)

	Hypotheses	*p-value*	Inferences
Test name	H_0 - H_a -		

FOR TOP MARKS: A top-scoring statistical analysis demonstrates a logical chain: outlier check → normality check → variance check → appropriate main test → post-hoc (if significant) → correlation (if IV is continuous). Every test choice is justified based on the results of the preceding test and data available.

⚠ AVOID / COMMON MISTAKE

Do NOT force ANOVA on data where you only have one value per "group" (each source study = one data point). In this case, correlation or regression may be more appropriate.

Do NOT omit statistical analysis because you "only have secondary data." The IB expects statistical methods regardless of data source. Correlation coefficient must be accompanied by a significance test — r^2/r_s alone is not sufficient.

Use SD for $n \leq 30$. SEM is more appropriate for $n > 30$.

⚠ AVOID / COMMON MISTAKE

In a secondary data EE, each study's reported mean is one data point in your analysis. This means your effective sample size is often small (e.g. 8–15 studies). This has important implications for which statistical tests are appropriate:

- **Correlation (Pearson's / Spearman's):** you need at least 8–10 studies for a correlation to be statistically meaningful. With fewer than 8 data points, the p-value will almost never reach significance regardless of the trend
- **Two-group comparison (t-test / Mann-Whitney):** you need at least 4–5 studies per group, so 8–10 studies total minimum
- **Three+ group comparison (ANOVA / Kruskal-Wallis):** you need at least 3–4 studies per group, so 9–12+ studies total for three groups
- **Chi-squared:** needs expected frequency ≥ 5 per cell and n ≥ 20 total

With fewer than ~10 data points per group, the Shapiro-Wilk normality test has very low statistical power – it will almost never reject normality even if the data is not normally distributed. This means a "passed" result cannot be trusted. In this situation, skip the normality test and go directly to non-parametric tests (Spearman's, Mann-Whitney, Kruskal-Wallis). This is not a weakness — it is the statistically honest approach. State this justification explicitly in your essay.

6.4 Graphs

Graphs must plot **processed data** (means, NOT raw trial data NOT screenshots from source papers). Choose graph type that matches your IV type. Quality over quantity – include only the graph(s) essential to answering the RQ.

Graph caption – required elements

- **Graph number and title**
- **The original source(s) of the data**
- **What is being measured and over what time period**
- **Source of error bars** (e.g., 'Vertical error bars show ±1 standard deviation') – can be calculated from raw or from published studies
- **Statistical information** – r^2, R^2, *p-value* from correlation test; or letters indicating post-hoc significance (if applicable)

Scatter plot (continuous IV)	Bar chart (categorical/discontinuous IV)
• X axis: IV with name and units • Y axis: Mean DV with name and units • LOBF: solid line; display r^2 / r_s and *p-value* from test • Vertical error bars: ±1 SD (SEM if n > 30) • Horizontal error bars: ±1 IV uncertainty * If analyzing temporal trends from monitoring data a **time-series scatter plot** may be used with the variable plotted against time with a regression line	• X axis: IV group descriptor (no units needed) • Y axis: Mean DV with name and units • Vertical error bars: ±1 SD (SEM if n > 30) • Letters above bars: indicate post-hoc significance groupings (e.g., Tukey p < 0.05)

⚠ AVOID / COMMON MISTAKE

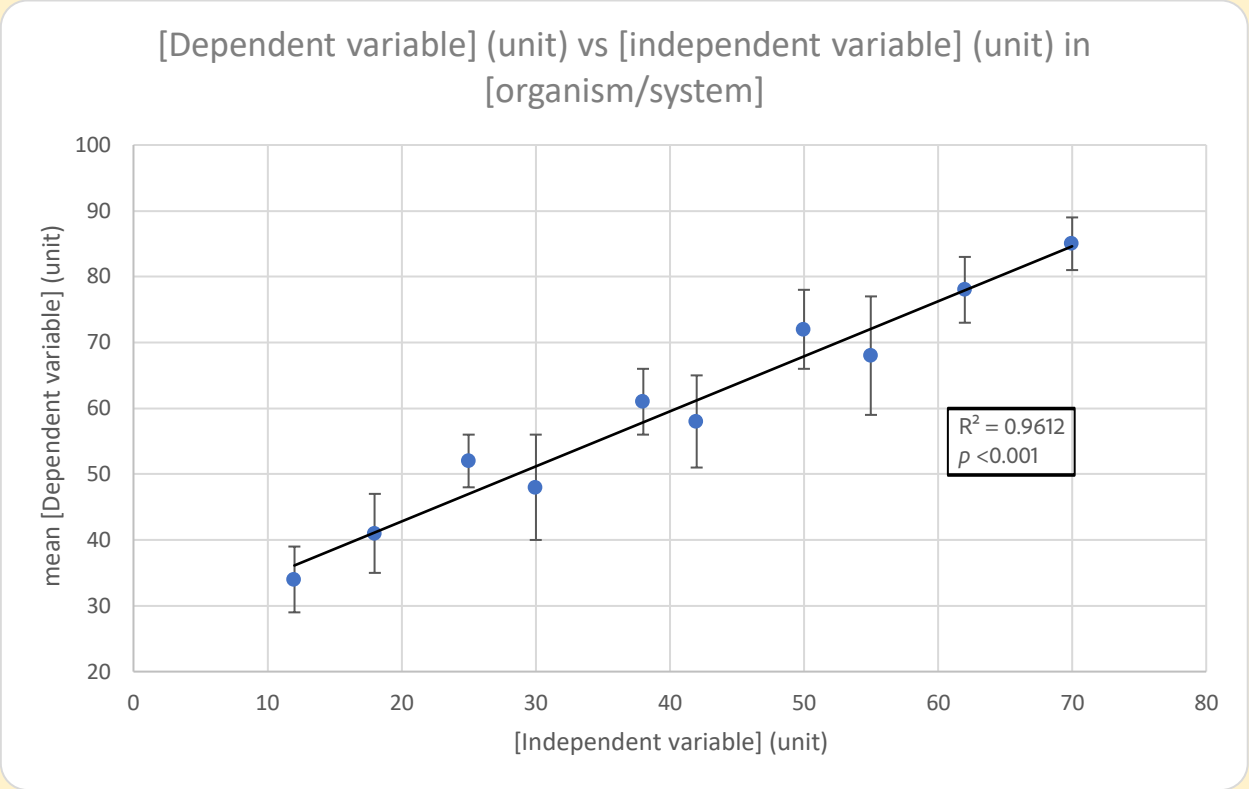
Do NOT include any graphic not referenced in the body text.
Do NOT screenshot graphs from source papers – create your OWN original graphs from your extracted data. Using another researcher's graph is not original analysis and will not score well.

Do NOT screenshot graphs from software – copy and paste as high-resolution images.
Do NOT plot raw trial data on graphs – the Y axis should show mean DV values.
Make graphs large and easy to read. Font size on axis labels should be at least 12pt.

*Below are examples of potential graphs in secondary data Ees. For more examples of graph formatting, consult the **Primary Data Investigations Guide** which provides some examples of graphs derived from raw data.*

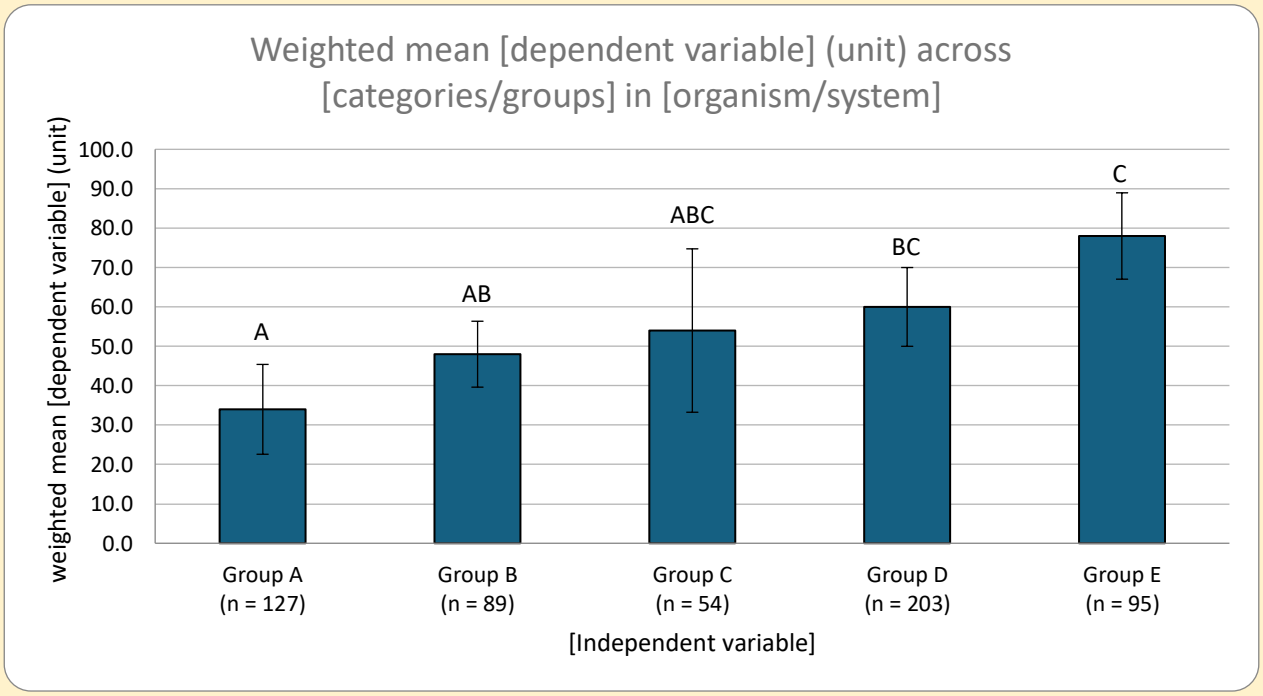
EXAMPLE GRAPH FORMAT (SCATTER PLOT)

Graph 1 – Scatter plot [dependent variable] (unit) against [independent variable] (unit) across 10 published studies on *Study species*/system. Each data point represented the reported mean from one study (see Table X for full source details). Errors bars show ±1 Standard deviation as reported by each study. Coefficient of determination displayed under trendline along with *p-value* from Pearson correlation coefficient.



EXAMPLE GRAPH FORMAT (BAR CHART)

Graph 2 – Bar chart of weighted mean [dependent variable] (unit) of *Study species* across 5 [categories/conditions], compiled from 14 published studies. Error bars show ±1 Standard deviation. Groups which do not share letters denote significant difference (*post-hoc* Tukey $p < 0.05$). Sample sizes shown in parentheses.



✅ **FOR TOP MARKS:** Every graph and figure must be referred to in the body text at the point where it supports your argument. "When it comes to writing the essay, include only graphics, tables and other information that support the points you are making. Be sure to refer to the graphic in the text. If you do not refer to the graphic, then it is not supporting your points and it should not be used."

7. DISCUSSION (Criteria B, C, D)

The discussion is where you (1) interpret your results, (2) explain them biologically, (3) compare them with published literature, and (4) evaluate your methodology. This is where the highest marks are earned across Criteria C and D. Every paragraph must be analytical, not descriptive. Analytical writing explores the why, how, and so what – descriptive writing merely states what happened.

HOW THE SECONDARY DATA EE DISCUSSION DIFFERS FROM THE IA AND PRIMARY DATA EE

Compared to the Biology IA:

In the IA, Criterion C separates Analysis from Conclusion, and Criterion D is a standalone evaluation table placed after the conclusion. In the EE, all of these elements are handled within the Discussion section. Evaluation must appear before the conclusion — not after it. Literature comparison happens throughout the discussion, not saved for the conclusion. Unlike the current IA criteria, the EE Criterion D explicitly assesses both strengths AND limitations.

Compared to the Primary Data EE:

Same structure: interpret → explain biologically → compare with literature → evaluate. However, for secondary data EEs:

- **You compare your SYNTHESIS with individual study findings** rather than comparing your experiment with published studies
- **Evaluation focuses on SOURCE quality and ANALYTICAL limitations** rather than experimental errors (e.g. random vs systematic)
- **Differences between source studies should be discussed** — differences in methodology, species, or conditions between source studies should be considered as potential explanations for patterns or discrepancies (the IB requires awareness of "limitations or uncertainties inherent in their approach")

***tl;dr**: The IA separates analysis and evaluation into distinct sections with tables. The primary EE merges them into a prose Discussion with evaluation woven in. The secondary EE uses the same prose structure but shifts the evaluation from experimental limitations to source quality, analytical limitations, and differences between source studies.*

ANALYTICAL WRITING TIP – PEEL TECHNIQUE

Use PEEL to structure each paragraph in your discussion:

P – State the Point you are making

E – Provide Evidence (data, citation) to support it

E – Explain how the evidence supports the point

L – Link the point to your research question and the next point in your argument

Recommended Structure:

Use "Discussion" as the overall section heading. Within it, use subheadings that reflect the content of YOUR investigation – name your specific variables, trends, and themes.

Subheading 1: The main relationship between IV and DV

Name specific variables in the subheading. This is the "big picture" — what does your synthesised data show overall?

- **Overall trend described with data** – describe the relationship or pattern revealed by your analysis with reference to specific values from your processed data table and graph
- **Variability across studies** – discuss the spread/range of values across source studies — what does this tell you about consistency of the relationship?
- **Statistical results interpreted** – interpret the main test results (p-value, r^2, post-hoc groupings) — explain what the results mean for your conclusion, not just the numbers
- **Biological explanation** – explain WHY this overall trend/pattern occurs using mechanisms, pathways, or processes from published literature

Subheading 2: Key features, anomalies, and between-study differences

Zoom into anything that deviates from, or adds nuance to, the main pattern:

- **Outlier studies** – if any source study reports values that deviate from the overall trend, explain possible reasons (different methodology, species, conditions, sample size)
- **Sub-group differences** – if different categories of studies (e.g. field vs lab, tropical vs temperate) show different patterns, discuss why
- **Between-study differences discussed** – discuss why different source studies may have produced different results (e.g. different species, conditions, sample sizes, methodologies). Consider whether these differences affect the reliability of your pooled findings
- **Evaluation woven in** – if a methodological limitation explains a discrepancy (e.g. different measurement techniques across studies), discuss it here with a specific explanation

Subheading 3: Comparison with existing reviews and established knowledge

Compare your results with the studies cited in your introduction:

- **Agreements** – where your findings align with existing reviews or accepted biological theory – explain why this consistency strengthens the conclusion
- **Discrepancies** – where your findings differ – explain possible reasons (e.g. different inclusion criteria, different time periods, different analytical methods)
- **Significance** – what your synthesis adds to or clarifies about existing knowledge

Subheading 4: Strengths, limitations, and improvements

Evaluate the methodology and sources. This is especially important for secondary data EEs because the IB explicitly states: "The student must comment on the quality, balance and quantity of the secondary sources and data used. They are also expected to show an awareness of any limitations or uncertainties inherent in their approach."

- **Strengths of the approach** – identify specific methodological strengths (e.g. systematic search strategy, large combined sample size, diversity of source studies, use of established databases)
- **Limitations** – **data quality** – discuss limitations arising from the quality of source data (e.g. inconsistent measurement methods across studies, different sample sizes, missing data, potential for reporting bias)
- **Limitations** – **analytical** – discuss limitations of your analytical approach (e.g. inability to control for confounding variables across studies, limitations of summary statistics vs raw data, small number of source studies)
- **Source bias considered** – acknowledge that published studies may over-represent significant or positive results, which could skew your synthesis. Consider whether the studies you found represent the full picture or only a subset of research conducted on this topic
- **Improvements** – for each limitation, propose a realistic, specific improvement and explain WHY it would address the issue
- **Sources evaluated** – assess the quality and relevance of the published literature used (peer-review status, journal impact factor, sample sizes, funding, etc.)
- **Unresolved issues** – identify what remains unanswered and suggest directions for future research

SUBHEADINGS TIPS

Name your specific variables in the subheadings – **do not use generic labels**. You may combine or split these subheadings depending on complexity. For example, if there is no clear threshold or optimum, subheadings 1 and 2 can be merged.

Every paragraph under every subheading must contribute to answering the RQ. If it does not advance the argument, remove it.

✓ **FOR TOP MARKS:** Criterion C (Analysis) top band: "Analysis in the essay is effective and consistently produces relevant findings." Criterion C (Line of argument) top band: "A clear, sustained line of argument links the research question, research findings and conclusions." Criterion D (Discussion) top band (7–8): "A balanced discussion of the significance of the findings is fully supported by appropriate evidence." Criterion D (Evaluation) top band (7–8): "An evaluation of the effectiveness of the essay is present, with relevant strengths and limitations EXPLAINED."

⚠ AVOID / COMMON MISTAKE

Do NOT place evaluation after the conclusion (it is part of the discussion). The conclusion must be the final written section of the essay. Do NOT use generic evaluation points ("more sources", "better databases") without specifying what, why, and how. Do NOT simply summarise each source paper in sequence — this is a literature review, not a discussion. Your discussion must be organised by THEME or FINDING, with evidence drawn from across your sources.

8. CONCLUSION (Criteria C, D)

The conclusion is the final written section of the essay. It must directly address the research question and synthesise the findings. It should NOT repeat the discussion – it should draw the threads of the argument together into a clear, concise answer

Recommended Structure:

Write the conclusion as continuous prose. The following elements should appear in approximately this order:

1. Direct answer to the research question

- **State the conclusion explicitly** – open with a clear, unambiguous statement that directly answers the RQ
- **Restate the RQ** – remind the reader of the exact research question (copy-paste – same wording as everywhere else)
- **Supported by data** – refer to key statistical results (e.g., ANOVA p-value, correlation r^2) that justify the conclusion – do not re-describe trends, just reference them briefly

2. Evaluation of the hypothesis

- **Supported, partially supported, or refuted** – state which, explicitly
- **Explain why** – briefly explain which aspects of the data support or contradict the prediction, referencing specific findings from the discussion

3. Synthesis in scientific context

- **Synthesise, don't repeat** – provide a "supported, well-explained synthesis" of results – not a repetition of data trends or a re-description of the discussion
- **Broader significance** – briefly state what your findings mean in the wider biological context (implications for the field, connection to real-world applications)
- **Consistency with literature** – one or two sentences summarising whether your overall conclusion aligns with or challenges published studies cited in your introduction

4. Remaining uncertainties and future directions

- **Unresolved issues** – if the RQ was not fully answered, state this clearly and explain why
- **Future research** – suggest specific, realistic directions that would address unresolved issues (e.g. primary data experiments to test the patterns your synthesis revealed)

⚠ AVOID / COMMON MISTAKE

Do NOT introduce new data, new processing, or new sources in the conclusion. Do NOT simply restate trends already described in the discussion – synthesise them into a coherent answer.
Negative results are valid – "negative" results are just as valid as "positive" results – do not force conclusions the data does not support.
No definitive answer is acceptable – the EE guide states: "You might not be able to give a definitive answer to the question. Not having a definitive answer should not compromise your capacity to perform well, as determined by the assessment criteria: as long as your discussion and argumentation are strong."

9. REFERENCE LIST (Criterion B)

Required elements

- **APA format throughout** – all references formatted correctly and consistently → Citation generator
- **Alphabetical order** – by first author's surname
- **Every in-text citation has a corresponding entry** – and vice versa; no orphan citations or unused references
- **Database access dates** – for all databases from which data was extracted, include the access date and version number (if applicable)
- **Appropriate sources only** – peer-reviewed journal articles, academic textbooks, and trusted institutional websites; NOT Wikipedia, revision sites, or general web pages
- **Substantial reference list expected** — the IB states that for secondary data EEs, "an essay of this type would normally be expected to produce a substantial bibliography and not be limited to just a few sources." As a guideline, a thorough secondary data EE typically includes 20–40+ references

✓ **FOR TOP MARKS:** Criterion B (Knowledge) top band: "Comprehensive, relevant research materials are used to establish knowledge of the subject matter." Grade A descriptor: "There is effective engagement with relevant research areas, methods and sources." For secondary data EEs, this criterion carries additional weight because the reference list IS the evidence base for the entire essay.

⚠ AVOID / COMMON MISTAKE

Any IA that lacks references and a reference list will be submitted as 'no grade' due to doubts of authenticity

10. APPENDICES

This section is optional. Include only supplementary evidence that supports the transparency and reproducibility of the investigation but is not assessed.

Potential included elements

- **Full data extraction table** – if the complete table was too large for the main body, include it here and reference it (e.g. "Full data extraction table available in Appendix A")
- **Complete search records** – full database search results and screening records
- **PRISMA flow diagram** – if not included in the main body
- **Raw statistical output** – screenshots of raw output data tables produced by online calculators or software
- **Supplementary graphs or figures** – additional visualisations that support but are not essential to the argument

⚠ AVOID / COMMON MISTAKE

Appendix is NOT assessed by examiners – anything the student wants the examiner to read and credit should NOT go in this section. It is NOT to be used as a word-count overflow section. **Reference each appendix in the main body** (e.g. "see Appendix A")

11. REFLECTIVE STATEMENT (Criterion E on RPF)

The 500-word reflective statement is written at the end of the EE process and recorded on the Reflection and Progress Form (RPF), which is uploaded as a separate file (NOT part of the essay PDF). Comprises (1) Evaluative strand – evaluate what was LEARNED, WHY it mattered, and HOW it changed your thinking and (2) Growth and Transfer – how you developed as a learner and HOW this is connected to other contexts

Recommended structure:

- **Opening (~50 words)** – state the most significant thing you learned from the EE. The examiner should immediately see you are *evaluating*, not describing
- **Body (~350 words)** – develop 2–3 evaluative points. For each: (1) a specific moment or challenge from your EE process, (2) what you learned from it, (3) where you have applied or will apply this learning. Depth over breadth – develop a few points in detail rather than listing many
- **Closing (~100 words)** – reflect on how the EE experience shaped your thinking overall. What would you do differently, and why?

WRITING TIPS

For every sentence, ask: does this tell the examiner what I DID, or what I LEARNED? If it only tells them what you did, rewrite it to explain the learning, the insight, or the change in perspective. Possible topics to evaluate:

- **Developing a systematic search methodology** – what did you learn about finding and evaluating sources?
- **Navigating contradictory literature** – how did you learn to evaluate source quality?
- **Learning a new statistical test** – how did this change how you interpret data?
- **Narrowing or changing your RQ** – what did this teach you about research focus?
- **Dealing with missing or inconsistent data** – how did this shape your understanding of secondary data limitations?
- **Managing time and workload** – what strategies did you develop and where could they apply?
- **Understanding correlation vs causation** – how did this affect your conclusions?
- **Writing in academic register** – what did you learn about communicating science?

✓ **FOR TOP MARKS:** Evaluative: "Reflection is consistently evaluative and includes specific examples." Growth: "Reflection consistently shows evidence of the student's growth and transfer of learning."

Key words: consistently, evaluative, specific examples, growth, transfer.

⚠ AVOID / COMMON MISTAKE

Do NOT describe your process – evaluate your learning. Do NOT write general statements – give specific examples.
Do NOT forget transfer – explicitly connect learning to another context.
Do NOT exceed 500 words. Do NOT include the reflective statement in the essay – it goes on the RPF (separate upload).

12. FORMATTING AND WORD COUNT

Word count

- **4,000 words MAXIMUM** – following are excluded from the word count: contents page, diagrams, graphs, data tables, equations, calculations, in-text citations, reference list, headers, appendix, figure/table captions, RPF

⚠ AVOID / COMMON MISTAKE

Any content that goes beyond 4000 words is NOT READ and therefore NOT COUNTED in the grading. While data tables are not included, tables that include descriptive text are (e.g. controls, qualitative data, evaluation).

Layout

- **1.5x line spacing throughout**
- **Font size 12 minimum** for ALL text – including figure captions, graph axis labels, and table text
- **Page numbers on every page** – beginning with the first page after the contents page
- **Tables do not break across pages**
- **Headings/captions are not separated from their related content**

Figures and tables

- **Each figure has a name** (Fig.1, Fig.2…) AND a detailed caption (using APA guidelines)
- **Each figure referenced in text** – e.g., '(see Fig.1)' before or immediately after the figure
- **Figures placed near their in-text reference** – not on a separate page far from the citation
- **Images are not blurry** and stay within normal margins
- **Species names correctly formatted** – *Genus species* (italicised; Genus capitalised, species lowercase)

Writing style

- **Third-person passive throughout** – recommended to avoid all personal pronouns (I, we, my, our)
- **In-text citations** – every biological or scientific claim must be supported by an in-text citation (APA format)
- **Technical terms defined** – define complex/subject-specific terms clearly when first used; avoid jargon

Digital file

- **Save as PDF** – check the final PDF version for correct page numbers, image placement, consistent fonts, line spacing, and no widows/orphans before submission
- **File size under 10 MB** – optimise embedded images if needed. The RPF is uploaded separately

APA citation

- **In-text citations** – mainly parenthetical style (although narrative can be used when referring to a specific study/investigator). Direct quotations should be avoided or used very sparingly (note: they are included in word count)
- **Reference list** – alphabetical order. *Note: this is called a 'Reference List' NOT 'Works Cited' or 'Bibliography*

LINKED RESOURCE

Consult APA CITATION GUIDE for full details on in-text citations and reference list entries

⚠ ACADEMIC INTEGRITY AND AI USE

"Using artificial intelligence (AI) to write an essay that is then presented as your own is dishonest." Additionally, AI-generated material can be "considered as one of your resources… always acknowledged and cited appropriately."
Generally speaking, AI use should be avoided but if it used it must be declared and validated against other sources.