IB Biology Internal Assessment (IA)

Secondary Data Investigations (Literature & Data-based Research)

Complete Section-by-Section Guide to score 24/24

Based on: IB Assessment Criteria (first assessment 2025) | TSM (Unpacking the IA) | May and Nov 2025 Subject Reports

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

HOW TO USE THIS GUIDE

This guide is organized by IA section, in the order they must appear in your report. Sections contain:

- A description of what belongs in each section
- A detailed checklist of required elements
- Suggested formatting for tables/graphs
- Common mistakes to avoid

AVOID / COMMON MISTAKE

THE CAP RULE – Read this first!

The IB does NOT average strands within a criterion. The LOWEST strand caps the entire criterion score.

For example, if Strand 1 = 5–6, but Strand 2 = 3–4, your maximum for that criterion is 4 marks – regardless of the strength of the other strands.

The single most important language shift: moving from DESCRIBED (3–4) to EXPLAINED/JUSTIFIED (5–6). Students who describe their choices land in the 3–4 band. Students who explain WHY land in the 5–6 band.

SECONDARY DATA IA – KEY CONTEXT

The IB explicitly recognises secondary data as a valid IA approach.

The official Biology IA criteria explicitly list "the selection of the databases or model and the sampling of data" as a methodological consideration assessed under Criterion A. Investigations that use published data, online databases, simulations, or other secondary sources can score the full 24/24.

However, secondary data IAs are often scrutinised more heavily by moderators because the methodology must demonstrate the same rigour as a primary-data investigation – just with the rigour applied to data sourcing, screening, extraction, and standardisation rather than to laboratory technique.

Two critical points before you begin:

1. **You must conduct ORIGINAL analysis** – not simply summarise what other researchers have already concluded. The IB explicitly states essays/reports that "simply restate facts or data taken directly from the sources are of little value." You must extract data, reorganise it, recalculate it, perform your own statistical tests, and create your own graphs.
2. **The IA word limit (3,000) and the four-criterion structure of the IA still apply.** This is NOT an Extended Essay. Use labelled sections with tables (as in any IA), not continuous prose with subheadings.

CRITERION A – RESEARCH DESIGN (6 marks)

This criterion assesses how effectively you communicate the methodology – both its purpose and its practice – used to address your research question. For secondary-data IAs, the methodology covers your search strategy, source selection, data extraction, and standardisation. All three strands must reach the 5–6 band.

1. Research Question

The RQ is the title of your IA. For secondary-data investigations it must be specific, focused, and answerable through systematic analysis of existing data.

Required elements (for 5–6 band)

- **Explanatory factor (IV) stated with full range/categories and units** – the factor that varies across the studies, dataset records, or time points you are analysing
- **Response variable (DV) stated with unit and time/scale of measurement**
- **Study organism or system named** – if a single species, use common name (*Species name*); if multiple species/a broader system, name the taxonomic group, ecosystem, or database
- **Scope of dataset defined** – year range, geographic range, or other boundaries that define which data is included
- **Consistent wording** – copy-paste the RQ; it must be written *identically* everywhere in the IA

✅ **FOR 5–6 BAND:** RQ explicitly states the explanatory factor (with range/categories), the response variable (with unit and time/scale), the organism or system, and the scope (year range, geographic range, or database). The question makes clear WHY analysing this dataset answers it.

⚠ **AVOID / COMMON MISTAKE**

Do NOT write the RQ as a yes/no question – use sentence starters like ‘how does’ or ‘to what extent’. Do NOT omit the range of IV. Do NOT use vague terms like 'amount' – use proper SI units. Do NOT state a derived variable directly as the DV in the RQ, state the raw measurement variable directly

EXAMPLE RQ FORMATS FOR SECONDARY DATA

- **Comparative** – "To what extent does [environmental factor] affect [DV] (unit) of [taxonomic group] across [geographic scope], based on published field studies from [year range]?"
- **Trend/relationship** – "What is the relationship between [environmental variable] (unit) and [DV] (unit) in [ecosystem/region], as reported in monitoring data from [year range]?"
- **Database analysis** – "How does the frequency of [mutation/feature] in the [gene/protein] vary across [comparison groups], based on data from [database name]?"
- **Multi-study synthesis** – "What is the overall effect of [treatment/condition] on [DV] (unit) in [organism], based on a synthesis of published experimental data from [year range]?"

2. Background

The background provides all theory and context a reader needs to understand and follow the investigation. For secondary-data IAs this section is especially important because the published literature IS your evidence base – the reader must understand the biology before they can evaluate your synthesis.

Required elements (for 5–6 band)

- **Specific context** – describe the exact biological system in which the RQ is embedded (not just the broad topic)
- **Detailed biological explanation** – explain the main biological process(es) directly involved in your investigation with enough depth to interpret future results
- **IV–DV relationship explained** – use theory to establish WHY the explanatory factor is expected to affect the response variable in a biologically meaningful way
- **Study organism/system justified** – explain why this organism or system was chosen for secondary-data analysis (well-studied, extensive published data, ecological/medical importance, suitable for synthesis)
- **Prior peer-reviewed studies cited** – describe what is already known and identify the specific gap your synthesis/re-analysis addresses
- **Justification for the secondary-data approach** – briefly explain WHY a synthesis/comparison of existing data is appropriate for this RQ
- **Academic sources only** – peer-reviewed journal articles and textbooks, and trusted institutional databases; NOT blogs, Wikipedia, revision sites. IA-generated content
- **Figures where helpful** – include a diagram, map, or schematic (with Figure caption and in-text reference) if it helps explain the system

✅ **FOR 5–6 BAND:** Background focuses on the EXACT variables being tested — not a general topic overview. The justification for using secondary data is explicit. Sources are peer-reviewed and are cited correctly. The reader finishes the section understanding WHY a synthesis of existing data will answer this specific question.

⚠️ **AVOID / COMMON MISTAKE**

A broad overview of the topic is insufficient – connect theory directly to YOUR specific explanatory factor and response variable. Background sources cited here must also be revisited in the Conclusion. Do NOT include a section explaining why you personally chose this topic (engagement is no longer assessed). Do NOT use "I couldn't do this in the lab" as the only justification for the secondary-data approach – frame it positively (scale, scope, data availability).

3. Hypothesis

Technically optional but strongly recommended. A hypothesis provides a specific, falsifiable prediction and helps situate the RQ within biological theory. Null/alternative hypotheses belong in the Statistical Analyses section, not here.

Required elements

- **Specific directional prediction** – state how each explanatory-factor group/value will affect the response variable
- **Biological justification** – explain WHY you predict this outcome using theory from the background/past studies
- **Predictive graph (recommended)** – a simple sketch showing the hypothesised line/curve/comparison; label axes with units

4. Type of Secondary Data Investigation

Secondary-data biology IAs can take several forms. Identifying which type your investigation falls into helps you structure your methodology, data extraction, and statistical analysis appropriately. In ALL cases, you must conduct ORIGINAL analysis – the IA cannot simply summarise what other researchers have already concluded.

State explicitly in your report which type your investigation belongs to (you may combine types – see note below).

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", typically 8-10. 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 type, a secondary-data IA must demonstrate ORIGINAL ANALYSIS. You must extract data, reorganise it, process it statistically, create original graphs/tables, and argue your own conclusions. An IA that merely summarises what each study found is a literature review – not a scientific investigation – and will not score above 1–2 on Criterion B Strand 3 (processing relevant to RQ).

5. Independent Variable (Explanatory Factor)

Describe what factor varies across the studies/records/time points in your dataset, what range/categories it takes, and WHY you chose this range and these specific increments or groups.

Required elements (for 5–6 band)

- **Variable stated clearly** – named with full SI units (continuous) or categories defined (discrete)
- **Range/groups listed** – 5 groups (4+ treatment groups plus a baseline/control) recommended for robust statistical analysis; for continuous IV, sufficient values to demonstrate a relationship. (This doesn't apply for Type 4 studies).
- **Baseline/reference identified and justified** – if comparing across categories, what is the appropriate baseline? (e.g. ambient conditions, control population, baseline year)
- **Range justified** – explain WHY this range is biologically meaningful (cite published literature establishing the relevant range)
- **Increments/categories justified** – explain WHY these intervals or groups were selected (data availability, biologically meaningful thresholds, published optima)
- **How values are extracted** – describe exactly how each IV value is taken from the source (e.g. study's reported treatment value, database field, time stamp); See Section 8 (search strategy)

✅ **FOR 5–6 BAND:** Both range AND increments are justified using biological reasoning supported by citations or pre-trial data. The justification explains the biological significance of the chosen boundaries

⚠ AVOID / COMMON MISTAKE

"I chose these values because that was what the studies reported" is a generic justification. EXPLAIN the biological relevance of your chosen range. If a preliminary scan of the literature shaped your final IV groups, include a brief description of what you reviewed and what decisions resulted.

6. Dependent Variable (Response Variable)

Describe exactly what is being measured (across studies/records/time points), how it is reported in your sources, and WHY this measurement is appropriate for addressing the RQ.

Required elements (for 5–6 band)

- **Variable named with SI units and time/scale of measurement**
- **Measurement method(s) described** – explain how your source studies/databases obtained the DV. If methods differ across sources, note this here – the strategy for handling differences is described in Data Standardisation.
- **DV selection justified** – explain WHY this measurement best answers the RQ (why not another proxy? why these units? why this time scale?)
- **Number of data points per group** – state how many studies/records contribute to each IV group and justify why this is sufficient for the planned statistical test (minimum of 5 data points per group required to calculate SD)

⚠ AVOID / COMMON MISTAKE

In a secondary-data IA, each study's reported value is one data point in your analysis. With fewer than 5 data points per group you cannot calculate SD, and most statistical tests have very low power. Prioritise gathering enough data points per group over investigating multiple IVs or an overambitious range.
If your IV is continuous, aim for at least 8–10 total data points so that a Pearson/Spearman correlation can produce a meaningful p-value.

7. Confounding Variables

In a secondary-data IA, you cannot directly "control" variables in the way a primary-data investigator can – instead, confounding variables are factors that differ between your source studies (or database records) and that could affect the DV. Your strategy is to MINIMISE their impact through carefully chosen inclusion/exclusion criteria, and to MONITOR/acknowledge those that remain.

💡 TABLE FORMAT REQUIRED

Present confounding variables in a table with four columns: (1) Variable name, (2) Biological impact on DV – explained with a citation, (3) Strategy for minimisation – which inclusion/exclusion criterion or standardisation step addresses it (see Section 9), (4) Residual impact – whether it remains as a monitored confounder and how its effect on results will be assessed.

Confounding Variable	Biological Impact	Strategy for Minimization	Residual Impact

For EACH confounding variable – required elements (for 5–6 band)

- **Variable named**
- **Biological impact explained** – WHY would this variable affect the DV if not minimised? (cite a source)
- **Minimisation strategy described** – which specific inclusion/exclusion criterion or standardisation step addresses it (cross-reference Section 9 and Section 10)
- **Residual impact acknowledged** – if the variable cannot be eliminated, how will its effect on the conclusion be considered later in the Analysis and Evaluation?

✅ **FOR 5–6 BAND:** Each confounding variable has a clear chain: variable → biological impact on DV (explained with citation) → specific minimisation strategy (inclusion criterion, exclusion criterion, or standardisation step) → assessment of residual impact. This is the secondary-data equivalent of the "controlled vs. uncontrolled variables" distinction in a primary-data IA.

⚠️ **AVOID / COMMON MISTAKE**

Simply stating "studies used different methods" without explaining biological impact or minimisation strategy is insufficient. Each confounding variable must have a specific link to the inclusion/exclusion criteria or standardisation strategy. Common relevant confounders in secondary-data IAs include: differences in measurement technique, sample size differences between studies, geographic variation, time of year of data collection, species/strain differences, reporting bias, publication bias.

THE METHODOLOGY OF A SECONDARY DATA IA (SECTIONS 8-11)

In a primary-data IA, the "Methodology" is a single step-by-step lab procedure. In a secondary-data IA there is no lab procedure – the methodology IS the systematic process of sourcing, filtering, standardising, and extracting data.

- **Section 8 – Search Strategy and Databases** (where the data came from)
- **Section 9 – Inclusion and Exclusion Criteria** (how sources were filtered)
- **Section 10 – Data Standardisation** (how values were made comparable)

Do NOT add a separate "Methodology" section that repeats these – that wastes word count and confuses moderators. Collectively, Sections 8–10 must allow a reader to reproduce your dataset.

8. Search Strategy and Databases

A clear, transparent narrative describing how you identified and accessed your data sources. This is the secondary-data equivalent of the Materials & Apparatus section in a primary-data IA – it identifies the "tools" used to obtain data.

The standard approach to systematic data collection follows four stages used in published systematic reviews:

1. **Identification** – search databases using defined search terms and record the total number of results. Duplicates across multiple databases are removed.
2. **Screening** – read only the title and abstract of each result to determine whether it is potentially relevant to your RQ. This is a time-efficient first filter.
3. **Eligibility** – read the full text and assess whether the source meets all inclusion/exclusion criteria (Section 9). Record the number excluded and the reason for each.
4. **Included** – the final set of studies/records from which you extract data for analysis.

Required elements (for 5–6 band)

- **Databases identified** – name every database searched with URL and access date (+ version/edition if applicable)
- **Search terms listed** – provide the exact search strings, Boolean operators, and filters used (e.g. "([DV term A] OR [DV term B]) AND ([IV term]) AND ([organism])")
- **Date range specified** – state and justify the publication date range or temporal scope of data included
- **Number of sources at each stage** – report how many results each search returned, how many were screened, how many were excluded (and why), and how many were included
- **Pre-screening described** – if you adjusted your search terms or scope after preliminary scoping searches, briefly describe what you found and what decisions resulted (the equivalent of pre-trialling in a primary-data IA)
- **Detailed enough for replication** – another researcher should be able to reproduce your search and arrive at a comparable dataset
- **PRISMA-style flow diagram strongly recommended** – a single figure that visualises the entire identification → screening → eligibility → included pipeline. Because the diagram requires both the search numbers (this section) and the exclusion reasons (Section 9), the full template is presented at the end of Section 9.

✅ **FOR 5–6 BAND:** A top-scoring search strategy reads as a transparent, systematic narrative. The moderator should understand exactly how you went from an initial database search to a final, curated dataset – with numbers at every stage. Describing pre-screening decisions demonstrates intellectual honesty and methodological awareness, equivalent to a pilot experiment in a primary-data IA.

9. Inclusion and Exclusion Criteria

Defining clear criteria for which studies/records are included in or excluded from your dataset is essential for methodological transparency. This section is the secondary-data equivalent of the "Control Variables" section in a primary-data IA – it is how you actively minimise variability that would otherwise confound your conclusions.

💡 **TABLE FORMAT REQUIRED**

Present in a table with three columns: (1) Criterion, (2) Inclusion or Exclusion, (3) Justification – why this criterion matters for the validity of the analysis (cite a source where relevant).

Criterion	Inclusion or Exclusion	Justification

For EACH criterion – required elements (for 5–6 band)

- **Criterion named** – be specific (e.g. "Sample size n ≥ 20", "Published 2010–2025", "English language", "Peer-reviewed journals only")
- **Inclusion or exclusion stated** – clearly identify whether the criterion ADMITS or REJECTS data
- **Biological/methodological justification** – explain WHY this criterion matters for the validity of your analysis, and link it to a specific confounding variable from Section 7 where applicable
- **Decision rule for missing values stated** – if a source reports the mean but not SD/doesn't state n/no measurement date will it be included with a noted limitation or excluded? State this upfront and justify

Typical criteria worth considering

- **Source type** – peer-reviewed journals, institutional reports, authoritative institutional sources (e.g.WHO, FAO, EPA), validated databases; NOT blogs, or revision sites
- **Time period** – recent studies only (e.g. last 15 years) to ensure modern methodology; or wider range if temporal trends are part of the RQ
- **Geographic scope** – specific region(s) only, or global, depending on the RQ
- **Taxonomic scope** – specific species, genus, or higher taxonomic group
- **Measurement methodology** – consistent measurement technique across included studies
- **Minimum sample size** – studies below a defined n threshold excluded to ensure reliability
- **Data reporting** – studies must report mean, SD, and n (or sufficient information to derive these)
- **Language** – typically English-language only, acknowledged as a potential bias

✅ **FOR 5–6 BAND:** Clear inclusion/exclusion criteria demonstrate that your data selection was systematic and unbiased. Each criterion is explicitly linked back to a confounding variable from Section 7 OR justified on grounds of source quality/reliability. The decision rule for handling incomplete data is stated upfront so the moderator can see no cherry-picking occurred.

⚠️ **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 way as uncontrolled variables in a primary-data IA. Do NOT silently exclude inconvenient data – every exclusion must be documented and justified.

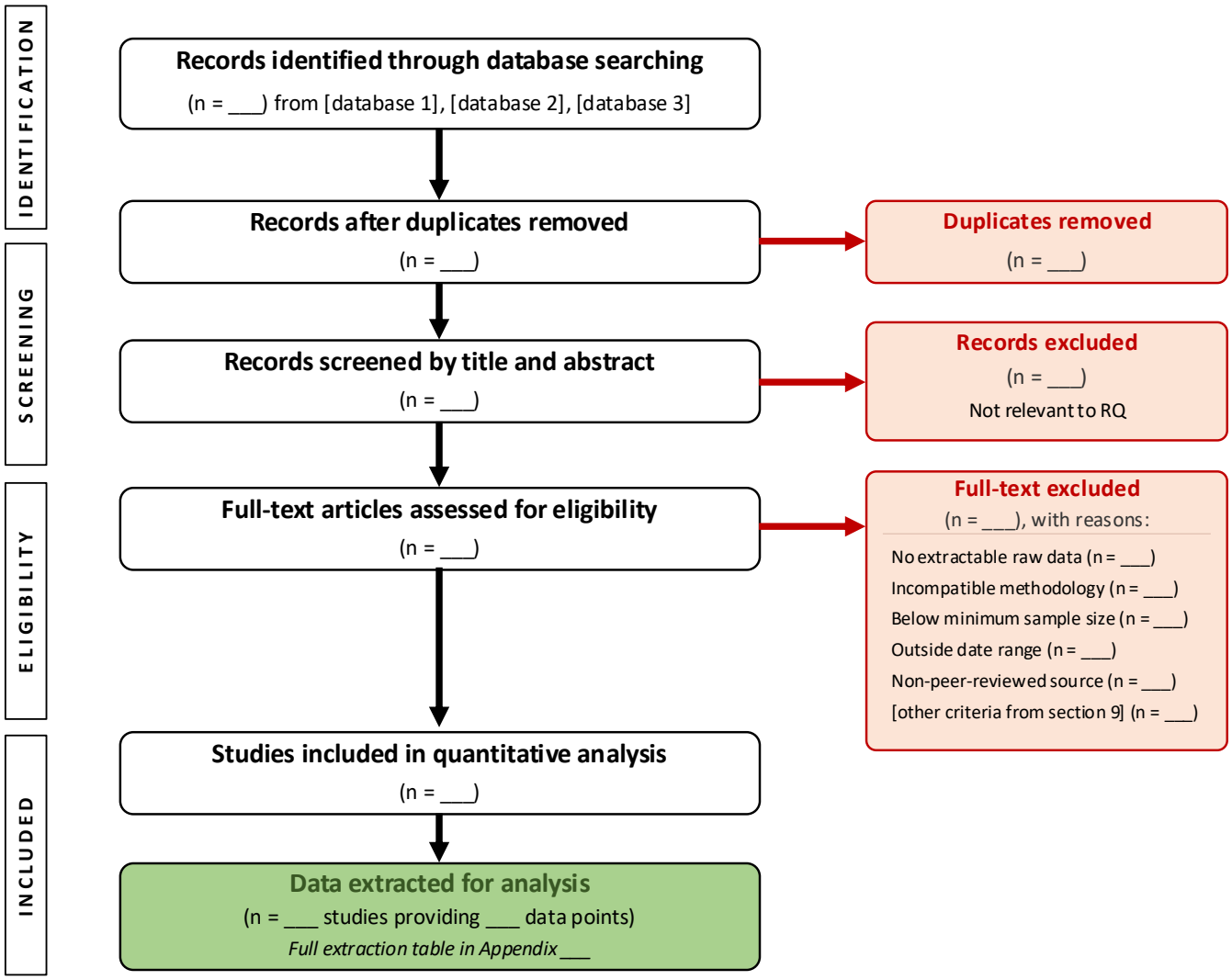
RECOMMENDED: PRISMA-STYLE FLOW DIAGRAM

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 IA 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



10. 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 values from different studies are directly comparable. This methodological decision must be planned and described here, not introduced for the first time in the Data Analysis.

Required elements (for 5–6 band)

- **Unit conversions identified** – if sources report the DV in different units, state which standard SI unit you will convert to and show the conversion (or include sample conversion in Processed Data)
- **Normalisation method described** – if raw values are not directly comparable due to differences in baseline (e.g. different control values, different body sizes), explain how data will be normalised (e.g. % change from control, per-unit-mass values, z-scores)
- **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 + IQR, others only mean and n). State whether you will exclude incompatible reporting or apply a conversion (with citation).
- **Weighting strategy** – if combining data across studies with different sample sizes, explain whether and how values will be weighted. Larger studies should contribute proportionally more to pooled estimates. The simplest defensible approach is weighting by sample size.
- **Inconsistent IV reporting** – If source studies report the IV as a range rather than a point value, state a decision rule before extraction and apply it consistently (e.g. use the midpoint of the range). Any assumption made must be noted in Section 18 as a limitation.

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 moderator that your consolidated dataset is valid for comparison. Unit conversions, normalisation, and weighting are all addressed prospectively – not improvised in the Data Analysis section.

⚠ **AVOID / COMMON MISTAKE**

Do NOT wait until 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 conversion. Do NOT ignore studies that report data differently from the majority – either convert or exclude with justification. Do NOT average SD from multiple groups – instead use a range of reported SDs (e.g. SD = 1.2-3.8) and report median SD as your representative value or used pooled SD formula if studies share the same dV and comparable sample sizes.

11. Ethical Considerations

Secondary-data IAs do not involve laboratory safety hazards or environmental disposal issues, but ethical considerations still apply. Note: include only relevant points. If a category does not apply, you may omit it.

When to include and what to address

- **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** – if your source studies were conducted on human participants or animals, briefly note that the original studies followed ethical guidelines (peer review provides this assurance)
- **Potential for harm or misuse** – if your conclusions could be misinterpreted (e.g. genetic data, disease frequency data, population-level conclusions about human groups), briefly acknowledge this
- **Sensitive data** – if working with patient data, anonymised individual-level data, or data on endangered species, address how the source materials handled this

CRITERION B – DATA ANALYSIS (6 marks)

This criterion assesses how you have recorded, processed, and presented your data in ways relevant to the research question. For secondary-data IAs, "recording" means transparently extracting values from your sources into a master dataset, and "processing" means standardising, summarising, and statistically analysing that dataset. All three strands must reach the 5–6 band.

12. Data Extraction Table (Raw Data)

All quantitative values taken from your sources, presented transparently. This is the "raw data" of a secondary-data IA. Every value must be traceable to its source.

Required elements (for 5–6 band – Criterion B, Strand 1)

- **Detailed table caption** – format: "Table X – Summary of extracted data on [DV] (unit) across [number] sources investigating [topic]. Data extracted between [dates]. Asterisk (*) denotes an outlier identified in Section 14 (Statistical Analyses)." Include definitions of any abbreviations.
- **Column headers** – include: Source (Author, year) / Record ID, IV value or category, n, DV value (mean ± SD where available, or raw value), method of measurement, original units (if conversion was needed), notes
- **Unit standardisation applied** – all values in this table should already be in the standardised SI units described in Section 10. Indicate conversions in the caption.
- **Consistent decimal places** – match the precision reported by the source studies, or choose a consistent precision and explain any rounding
- **All extracted data included** – if dataset is too large, include a representative sample in the main body and the full extraction table in an Appendix (with a justifying statement)
- **Anomalies flagged visually** – use an asterisk or coloured highlight to mark outliers identified in Statistical Analyses (Section 14)
- **Every value traceable to a source** – the Source / Record column must allow a moderator to locate the original value in the cited paper, database entry, or supplementary file
- **Per-source qualitative context in Notes column** – record methodological and contextual information per source (study design, location, instrument, author caveats, database flags); this replaces a standalone qualitative section and is drawn on in the Analysis to interpret anomalies.

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 10.

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 10.

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 10.

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]). All values standardised to [unit] as described in Section 10.

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

13. Processed Data

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

Required elements (for 5–6 band – Criterion B, Strands 1 & 3)

- **Description and justification of each calculation** – for each measure (weighted mean, pooled SD, percentage change, rate, CV) explain WHAT it shows and WHY it was performed
- **Method stated** – specify software and version (e.g. "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 strategy described in Section 10; show sample calculations
- **Unit conversions visible** – if conversions were applied during extraction, include at least one sample calculation
- **Instrument/source uncertainty acknowledged** – state the uncertainty in extracted values (often the SD as reported by the original studies, or the precision of the database's reported values). Note: propagation of uncertainties is optional and not required in IB Biology.
- **Processed data table** – with detailed caption; columns for IV group, weighted/pooled mean DV (with unit), pooled SD, CV (optional), total n across the group

14. Statistical Analyses

Select and apply statistical tests appropriate for your data type, sample size, and the nature of your secondary dataset. Justify every test choice. Present results clearly. The statistical demands for secondary-data IAs differ somewhat from primary-data IAs because you are often working with summary statistics (mean ± SD ± n from each source) 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
- **Each study contributes one data point** to your analysis. With 8–15 included studies, your "n" for many tests is small.
- **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 (for 5–6 band – Criterion B, Strands 2 & 3)

- **Outlier check** – use IQR to flag statistical outliers. Do NOT remove outliers without justification – indicate them with * in the raw data table. Consider presenting analyses both with and without flagged outliers.
- **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 data type, distribution, and sample size: if normally distributed and n is sufficient → t-test/ANOVA; if not normal or n is small → Mann-Whitney/Kruskal-Wallis
- **Post-hoc test if $p < \alpha$** – Tukey's HSD (following ANOVA) or Dunn's test (following Kruskal-Wallis); present pairwise comparisons in a table
- **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 and 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 → 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 IA, 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 IA.

15. Graph(s)

Graphs must plot processed data (means or weighted means, NOT raw trial data NOT screenshots from source papers). Choose the 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
- **Source of the data** – state that data was compiled from published sources (and reference the data extraction table)
- **Study organism / system** and the variables plotted
- **Source of error bars** (e.g. "Vertical error bars show ±1 pooled SD")
- **Statistical information** – r^2, r_s, p-value from correlation; or letters indicating post-hoc significance groupings

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)

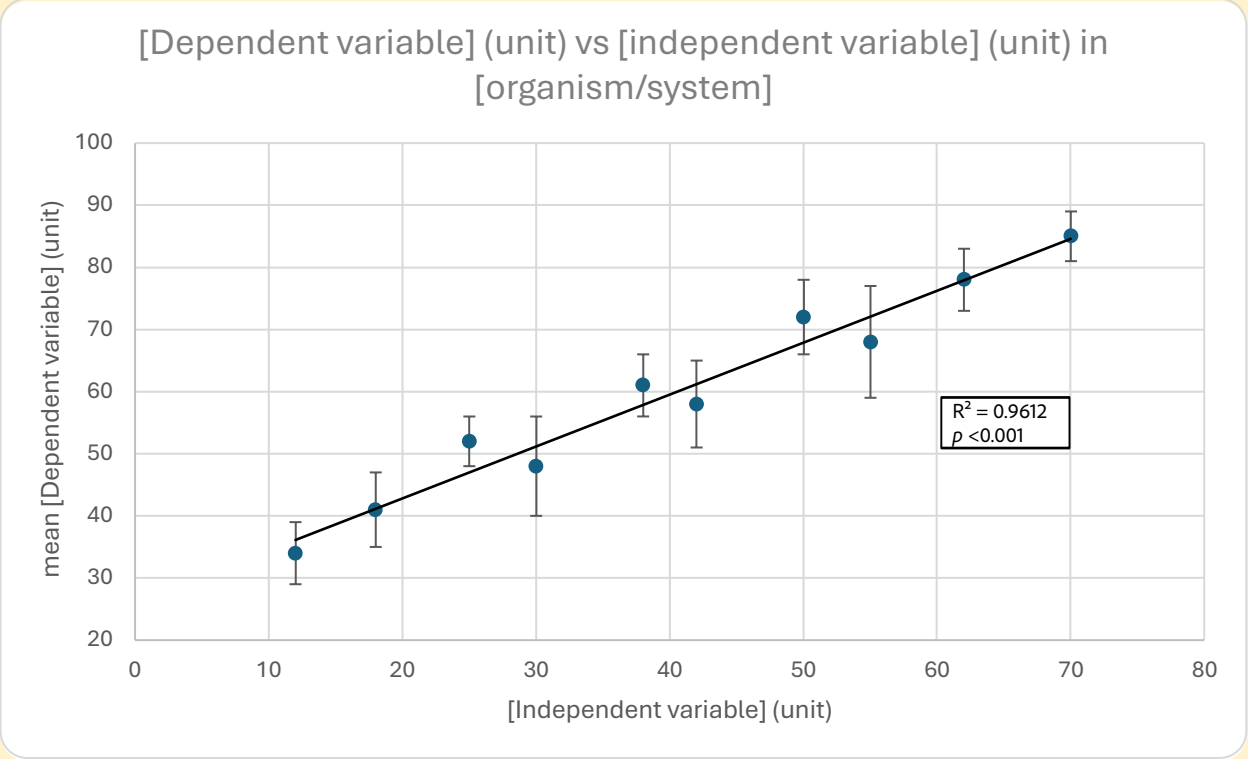
⚠ AVOID / COMMON MISTAKE

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 from Excel.
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.

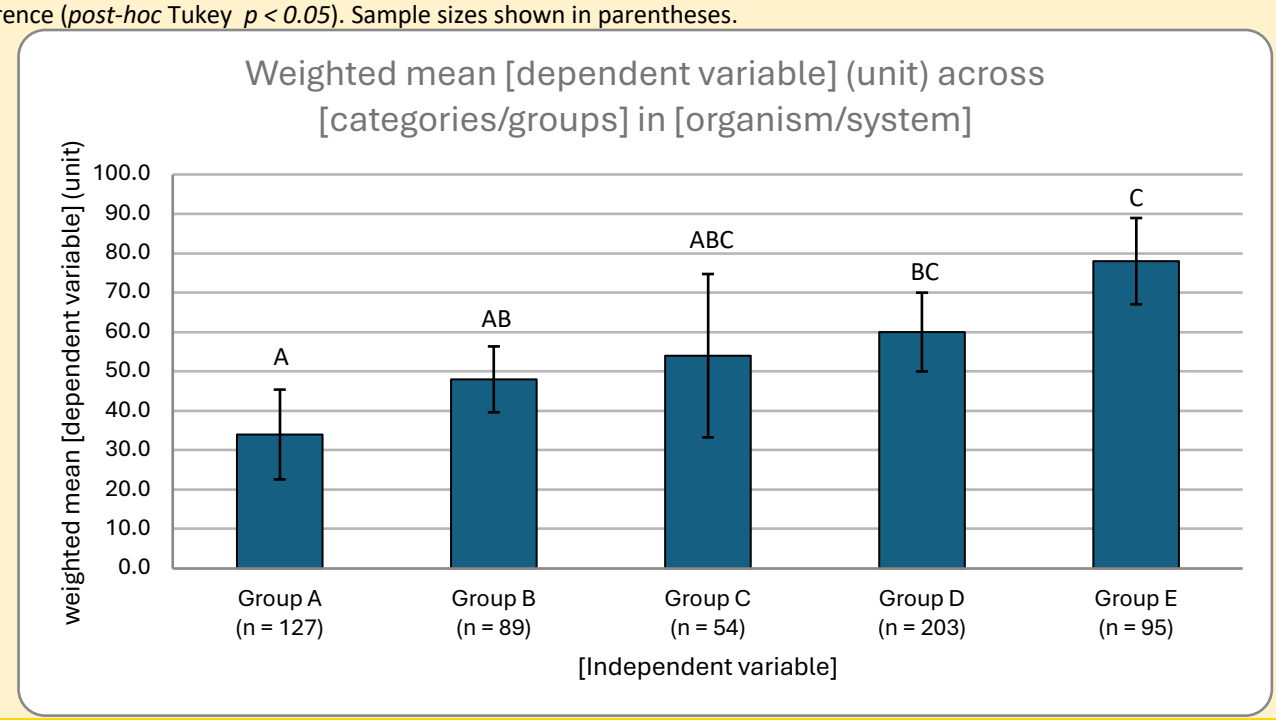
EXAMPLE GRAPH FORMAT (SCATTER PLOT)

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.



CRITERION C – CONCLUSION (6 marks)

Criterion C is written as a single narrative that moves from interpreting your processed data to explaining your findings in scientific context. It flows through three phases: (1) description of data, (2) interpretation of data, (3) biological explanation of data. This criterion assesses how successfully you answer the RQ using your data analysis in the context of accepted scientific knowledge.

16. Analysis

Describe and interpret your processed data. Do not explain causally here – that occurs in the next section (Conclusion).

Required elements (for 5–6 band – Criterion C, Strand 1)

- **Qualitative notes referenced** – describe patterns, anomalies, and any unusual contextual observations from the Notes column of the Data Extraction Table (section 12) cross-reference specific records or observations
- **Overall trends** – describe the overall relationship between IV and DV with reference to specific values from your processed data table and graph
- **Variability within groups** – discuss pooled SD values and error bar size; identify and discuss outliers statistically (IQR method) with justification for any exclusions
- **Variability between groups** – discuss error bar overlap; interpret p-values from ANOVA/Kruskal-Wallis and post-hoc tests with reference to the 0.05 significance threshold
- **Correlation results** – if IV is continuous, interpret r and R^2 values and their significance
- **Between-source variability addressed** – comment on whether the spread of values across source studies reflects genuine biological variation, methodological inconsistency, or limited sample sizes within source studies
- **Reference all figures and tables** – every table and graph must be cited in the Analysis (e.g. "See Graph 1…")

17. Conclusion

Draw your conclusion explicitly, grounded in the evidence established in Analysis. Reference your earlier analysis – do not restate it in full. Then use course biology and published literature to explain your findings causally (i.e. the biological why behind the patterns and the conclusion drawn earlier).

Required elements (for 5–6 band – Criterion C, Strand 1)

- **Answer the RQ directly** – state your conclusion explicitly and unambiguously at the start
- **Fully consistent with data** – refer back to specific data values, trends, and processed results from Analysis
- **Uncertainties interpreted** – discuss what the pooled SD, error bar overlap, outliers, and p-values collectively mean for the reliability and validity of your conclusion
- **Statistical results explained** – interpret p-values relative to the 0.05 threshold; explain what "statistically significant" or "not statistically significant" means for the conclusion
- **Hypothesis evaluated** – state explicitly whether the data supports or refutes your hypothesis and WHY
- **Correlation vs. causation** – if your investigation produces a correlation across observational data (which is typical for Types 1, 2, 3, 4), explicitly discuss whether a causal relationship can be inferred and why
- **Relevant biological terms used** – e.g. optima, maxima/plateau, thresholds, rate changes – where applicable

Required elements (for 5–6 band – Criterion C, Strand 2)

- **Biological explanation** – use course biology and cited literature to explain WHY the results turned out the way they did; this is mechanistic explanation, not re-description of trends
- **Prior studies revisited** – return to the studies cited in the Background (Criterion A); compare your synthesised findings to their individual conclusions; explain both agreements AND discrepancies using the literature

- **Unexpected results addressed** – if results were not as predicted, explain possible biological or methodological reasons using the literature; do not force conclusions the data do not support

✅ **FOR 5–6 BAND:** Background theory is actively REVISITED to explain findings — not just confirmed. Literature is used to explain unexpected results, not just expected ones. Specific data points are cited throughout the conclusion. For secondary-data IAs, the synthesis itself is part of what is being explained: WHY did pooling data across studies produce a clearer (or noisier) signal than the individual studies?

⚠️ **AVOID / COMMON MISTAKE**

A common Criterion C weakness is citing sources in the Background but NEVER returning to them in the Conclusion. Every key source should appear in both sections. Do NOT introduce new data or processing here – this is interpretation and explanation. Each statistical result or data point should appear *once* – no need for repetition. For secondary-data IAs specifically: do not treat correlation as causation just because your dataset is large – cross-study correlations in observational data are particularly vulnerable to unmeasured confounding.

CRITERION D – EVALUATION (6 marks)

This criterion assesses the evaluation of the methodology, which includes (1) weaknesses/limitations and (2) suggested improvements. Note: STRENGTHS are no longer required under the new criteria. Both strands must reach the 5–6 band.

For secondary-data IAs, weaknesses focus on SOURCE QUALITY and ANALYTICAL LIMITATIONS rather than experimental technique. The IB explicitly states examiners expect "awareness of any limitations or uncertainties inherent in their approach" – this is especially relevant for secondary-data approaches.

18. Evaluation

💡 **TABLE FORMAT REQUIRED**

Present evaluation in a two-column table: Left column – Weakness/Limitation with its relative impact on data quality and the chain of reasoning; Right column – Specific, realistic improvement that directly addresses the weakness, with explanation of expected benefit.

Weaknesses/Limitations	Suggestions for Improvement

Note: Points should clearly identify if it is a weakness or limitation – they are not synonomous:

- **Weakness**: An issue in the methodology or procedure that affected data quality – and that COULD be fixed if the investigation were repeated.
- **Limitation**: An inherent bound on what the experiment can answer, even if perfectly executed. Cannot be eliminated, only reduced.

For EACH weakness/limitation – required elements (for 5–6 band)

- **Relative impact assessed** – is this weakness/limitation minor or major? Explain qualitatively WHY it matters more or less than others by reasoning through its effect on data quality and conclusions. List the weaknesses/limitations in order from most to least impact and indicate this priority in the table caption.
- **Specific to this investigation** – not generic (avoid: "needed more sources", "human error", "better databases")
- **Impact explained** – explain HOW this weakness/limitation affected the data: what kind of error does it introduce (random or systematic)? How does this affect the conclusion?
- **Evidence from own data** – support each weakness with observations from YOUR results (e.g. unusually large pooled SD in one IV group, specific outlier studies, qualitative notes, large between-study variability)
- **Clear chain:** weakness → effect on data quality → effect on conclusion

For EACH improvement – required elements (for 5–6 band)

- **Specific improvement** – describe exactly HOW you would fix the issue in a future investigation (include specific changes to inclusion/exclusion criteria, additional databases to search, methodological standardisation steps, etc.)
- **Improvement linked to weakness** – explain WHY this specific improvement would address the identified weakness and what benefit it would provide
- **Realistic in school context** – improvements should be feasible for a high school student (e.g. expanding to additional databases, refining search terms, applying additional standardisation steps – not "conducting a meta-analysis with R")
- **Clear chain:** weakness → impact → specific improvement → expected benefit

Common categories of weakness/limitation in secondary-data IAs

- **Data quality limitations** – inconsistent measurement methods across source studies, missing SDs or sample sizes, varying levels of methodological detail provided by source authors
- **Source bias** – publication bias (significant results over-represented), English-language bias, geographic bias (well-studied regions dominate)
- **Analytical limitations** – small effective n (each study is one data point), inability to control for unmeasured confounders, limitations of working with summary statistics rather than raw data
- **Database limitations** – curation gaps, version-dependent data, inconsistent annotation across records, geographic or temporal sampling bias in monitoring databases
- **Standardisation limitations** – assumptions made in unit conversions, simplifications in normalisation, weighting strategy assumptions

✅ **FOR 5–6 BAND:** Almost no student reaches the 5–6 band for evaluation because they skip relative impact and do not EXPLAIN WHY these weaknesses/limitations and improvements are valid and relevant to the investigation. You must explain whether each weakness is minor or major and explicitly connect this to data quality and your conclusion clearly and in detail. For secondary-data IAs especially, the moderator wants to see that you understand the inherent limitations of working with other researchers' data.

⚠️ **AVOID / COMMON MISTAKE**

Don't confuse improvements with extensions – improvements must refine the original methodology, not propose an entirely new investigation (e.g. a different RQ, a different organism, or switching from secondary to primary data).

Generic weaknesses that will NOT score above 3–4:

- "Should have used more studies / larger dataset" (without specifying how many or why current n was insufficient)
- "Source bias is unavoidable" (not a methodological weakness with concrete impact)
- "Should have used more databases" (without specifying which, why, and how this would improve data)
- "Could not control all variables" (true of all secondary-data work – needs to be specific to YOUR investigation)

REFERENCE LIST

19. Reference List

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 and versions** – for ALL databases from which data was extracted, include the access date and version/edition number (if applicable). This is essential for traceability in secondary-data IAs.
- **Retrieval dates for online sources** – required for traceability
- **For database and institutional dataset sources:** access date AND dataset version or name must be included – a URL alone is insufficient, as database records are updated over time and values may change between access dates
- **Appropriate sources only** – peer-reviewed journal articles, academic textbooks, validated databases, and trusted institutional websites; NOT Wikipedia, revision sites, or general web pages
- **Substantial reference list expected** – secondary-data IAs typically include 15–30+ references because the literature IS the data source. A reference list with fewer than 10 entries is a strong warning sign to moderators.

⚠ AVOID / COMMON MISTAKE

Any IA that lacks references and a reference list will be submitted as 'no grade' due to doubts of authenticity. For secondary-data IAs, every data point in your extraction table must trace back to an entry in the reference list. The moderator should be able to verify any specific value you extracted by following the citation.

APPENDICES

20. Appendices

This section is optional but particularly useful for secondary-data IAs. Include only supplementary evidence that supports the transparency and reproducibility of the investigation but is not assessed.

Recommended appendices for secondary-data IAs

- **Full data extraction table** – if your dataset is large, include a representative sample in Section 12 and place the complete table here with a clear reference ("Full data extraction table available in Appendix A")
- **PRISMA-style flow diagram** – if not included in Section 9
- **Complete search records** – screenshots or tabulated records showing all search terms used, results returned, and screening decisions
- **Full statistical output** – screenshots of raw output from online calculators (Shapiro-Wilk results, ANOVA tables, post-hoc pairwise comparisons)
- **Unit conversion worksheets** – if many conversions were required, show the full conversion process for transparency
- **Database screenshots** – screenshots showing exactly where extracted values appeared in databases or supplementary files (useful for Type 2 investigations)

⚠ 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”). Note: while appendix is not assessed, a moderator *can* look at it to verify the student's claims

FORMATTING REQUIREMENTS

Formatting Checklist

Word count

- **3,000 words MAXIMUM** – the following are excluded from the word count: charts, diagrams, graphs, data tables, equations, calculations, in-text citations, reference list, headers, appendix, figure/table captions

⚠ AVOID / COMMON MISTAKE

Any content that goes beyond 3000 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

- **1.5–2× line spacing throughout**
- **Normal margins** (moderate at most)
- **Font size 12 minimum** for ALL text – including figure captions, graph axis labels, and table text
- **Same font throughout** – including all captions and graphs
- **Page numbers on every 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** – 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 unexplained jargon
- **RQ written identically** – every time the RQ appears in the report, the wording is identical

APA citation

- **In-text citations** – mainly parenthetical style (although narrative can be used when refering to a specific study/investigator)
- **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