Chapter 4: Critical Analysis — Mark Scheme

This mark scheme covers every Question, Activity and Consolidation-Exercise part in the worksheet. Because Chapter 4 consists largely of open-ended critical-evaluation tasks, credit should be given generously for any well-reasoned point that matches the spirit of the points below, even if not worded identically. Candidates are not expected to reproduce every point listed for full marks; the lists show the range of creditable ideas an examiner would draw on.

4.1 What is critical analysis?

Question A

No real evidence is given at all – the HR Director simply asserts, without data, that no individual has been held back because of background. This is an unsupported, sweeping generalisation ("no individual…") that is essentially unfalsifiable: it could never be checked or disproved from the statement alone. It is not a well-constructed argument because there is no evidence stage (e.g. statistics on promotion rates by background) to support the reasoning, and the statement conflates individual-level fairness with the possibility of a systemic/structural pattern in who gets promoted. Credit also: it is essentially an appeal to the speaker's own authority (as HR Director) rather than to verifiable evidence.

Question B

a) Evidence used: her own claim that teen/young-adult budgeting programmes "have all been failures", together with her personal/unpublished research suggesting poor financial habits are formed in early childhood by observing parents. Conclusion: funding and effort should be redirected away from programmes aimed at teenagers/young adults and towards primary-school children and their parents.

b) Possible flaws: "have all been failures" is a strong, emotive, unsupported claim – no data on the programmes' outcomes is given; she cites "my own research" without any detail, making it unverifiable and vague; she has a vested interest as the newly appointed chair of the charity, so may be selectively presenting evidence to justify a change in direction (possible bias); even if childhood habit-formation is correlated with adult financial difficulty, this does not prove that the existing teen programmes caused no benefit, nor that a new early-years programme would succeed (correlation vs. causation, and no evidence that the proposed alternative works either).

Question C

a) Correct. Among the 16 patients who actually have the condition (12 test positive + 4 test negative), 4 test negative, i.e. $\frac{4}{16} = 25\%$. This matches exactly what the first sentence claims (the test misses the condition in about one in four people who have it – the false-negative rate).

b) Incorrect. This confuses two different conditional probabilities. The 25% figure is $P(\text{negative test} \mid \text{condition})$, but the second sentence needs $P(\text{condition} \mid \text{negative test})$. From the table, of the $4 + 970 = 974$ patients who test negative, only 4 actually have the condition, so $P(\text{condition} \mid \text{negative test}) = \dfrac{4}{974} \approx 0.4\%$, not 25%. This is a classic statistical/legal fallacy (transposing the conditional, sometimes called the "prosecutor's fallacy") applied to medical testing.

4.2 Clarity

Question D

a) Emotive language: "blatantly obvious", "No sane person could think it sensible". Vague language: "performance aid" (which aids, how much advantage?), "the disqualification of so many sprinters" (no number or detail given).

b) The author assumes the reader already knows the details of the altitude-tent ban and the specific sprinter disqualification case being referred to, and understands the anti-doping/performance-aid rules well enough to judge the comparison – background knowledge that may not be safe to assume.

c) Self-contradiction: the author argues it is inconsistent/unfair to allow one performance aid (glucose gel) while banning another (implicitly, in the first two sentences), yet in the very next sentence defends the ban on a different performance aid (altitude tents) as "fully justified" – applying the opposite principle to a similar case.

d) The author assumes, without stating it, that glucose gel is used for a genuine medical necessity (treating diabetes) rather than purely for performance enhancement, which is why it should be an exception – this distinction between medical need and enhancement is never made explicit. Similarly, it is implied (but never stated) that the disqualified sprinters were using banned performance-enhancing methods (e.g. altitude tents) rather than being disqualified for an unrelated reason.

Activity 1

Only a photograph of one fossa is available, so only claims about *at least one* fossa can be safely deduced.

a) Cannot be deduced – one photo cannot establish a fact about *all* fossas.

b) Cannot be safely deduced as stated – "some" usually implies more than one, and only one animal has been observed.

c) Can be deduced – the animal in the photo is at least one fossa with reddish-brown fur (assuming the colour is correctly judged from the photo).

d) Can be deduced, and is the safest/weakest true statement – even if only part of the animal is visible, at least one side is reddish-brown.

The activity illustrates how easy it is to over-generalise from a single piece of evidence to a claim about a whole population.

Question E

a) Conclusion: employers must actively encourage people to delay retirement beyond pensionable age (skilled people should not take early retirement).

b) Main reason given: losing skilled people's experience and expertise through early retirement damages the UK. This reason only partly supports the conclusion – it explains why early retirement of skilled people might be undesirable, but no evidence is given to quantify the damage, and it does not by itself establish that employers "must" act, only that it might be beneficial if they did.

c) Self-contradiction/inconsistency: the passage argues on the grounds of protecting valuable skills (an argument that only applies to "skilled people"), but then switches to an argument about avoiding poverty in old age, which applies to everyone, skilled or not. The two justifications pull in different directions and are not reconciled – the skills-based argument is undermined by the fact the poverty argument would apply equally to less skilled workers.

Question F

a) Conclusion: it is wrong to throw away leftover food. Reason given: there are people going hungry in your own town.

b) The conclusion does not follow from the reason. There is no mechanism given by which one household's food waste causes, or even relates to, hunger elsewhere in the town (no redistribution system is described). The statement assumes, without justification, that not wasting the food

would somehow reach and help the hungry – a non-sequitur; a third factor (food-distribution infrastructure, poverty, access) would actually determine whether reduced waste helped anyone.

Question G

A full answer should comment on all four aspects named in the section:

- *Emotive language*: "beautiful tree canopy", "once extensive avenues of mature trees", "make the construction firm rich", "no reasonable person could…".
- *Vague phrases*: "over 5%" (5% of what base figure, over what time period, is this a lot?), "a large number of healthy oaks" (no actual number given), "what little remains" (unquantified).
- *Unjustified assumptions*: that removing 5% of tree canopy means "all our urban greenery" is under serious threat; that the car park development will bring no benefit to the town; that the construction firm getting paid for its work is inherently objectionable; that opposing the car park automatically helps elderly residents.
- *Self-contradiction / non-sequitur*: the letter's stated concern is environmental (tree loss) but it pivots to an unrelated argument about elderly residents and local shops, which has nothing to do with tree felling – introducing a second, disconnected justification suggests the writer is stretching for extra reasons rather than building one coherent case, weakening rather than strengthening the argument.

4.3 Selectivity of data

Question H

There are 6 possible outcomes for each roll, and successive rolls are independent, so the probability of correctly predicting all 8 rolls by chance is

$$\left(\frac{1}{6}\right)^8 = \frac{1}{6^8} = \frac{1}{1\,679\,616} \approx \frac{1}{1.7 \text{ million}},$$

confirming the quoted odds (since $6^8 = 1\,679\,616 \approx 1.68$ million).

Question I

The statistical evidence commits the "prosecutor's fallacy": a low probability of winning *by chance* does not, on its own, mean cheating is likely. With odds of 1 in 9 million, if a large number of people entered (or if raffles like this happen often), it is virtually certain that *someone* will win purely by chance – indeed, if roughly 9 million tickets were sold, a winner is guaranteed. The low probability describes how unlikely it is for any *particular* named person to win, not how unlikely it is that fraud occurred; no other, independent evidence of cheating (e.g. how the fraud was supposedly carried out) is presented, so the conviction rests on a misunderstanding of probability rather than genuine proof of wrongdoing.

Question J & K

a) The newspaper article (J) turns a controlled experimental finding about how managers *rated* an identical application when given a male vs. a female name into a sweeping causal claim that this bias is "why" only 22% of engineering managers are female – i.e. it presents the hiring-bias study as the (sole) explanation for female under-representation in management, which the original study does not claim.

b) The writer stripped out the methodological detail that gave the finding its credibility (sample size $n = 114$, randomised double-blind design), used a flippant/emotive tone ("Guess what!"), selected only the most dramatic findings (shortlisting and starting salary) while omitting the more nuanced result that manager gender made no difference, and asserted a causal, generalised

conclusion ("Now we know why") that goes well beyond what one experiment about ratings of a single fictional application can establish about real-world promotion statistics.

c) Most readers would find Article J easier to read: it uses short sentences, everyday vocabulary, and a narrative/conversational style, whereas Article K uses technical/statistical terminology ("randomised double-blind study", "$n = 114$", "significantly") and longer, more formal sentences. However, K is the more rigorous and reliable source, illustrating a trade-off between accessibility and accuracy.

4.4 Sampling and trialling

Activity 2

a) The sample deliberately balances age groups (25% from each of 5–8, 9–12, 13–15 and 16–17), so no single age band dominates the results, and it does cover the full 0–17 age range relevant to youth services.

b) It is likely unrepresentative because it only samples young people who already attend youth clubs – a self-selected group. It excludes under-18s who do not go to youth clubs (e.g. due to other commitments, disability, cost, lack of interest, or living in areas without a club), who may have very different views on youth services. It also may not reflect the true population split by age (forcing 25% in each band may not match the actual proportions of under-18s in each age group in the area), and it likely misses very young children (5–8 may not attend the same clubs/have limited independent opinions) and takes no account of gender, ethnicity, or socio-economic mix.

Question L

This is anecdotal evidence based on a single case ($n = 1$) and proves nothing about a general causal link between porridge and longevity. Many other centenarians did not eat porridge, and many porridge-eaters do not live to 101; other lifestyle, genetic and environmental factors are ignored. This is an example of selecting one convenient success story (confirmation bias) while ignoring the far larger number of non-confirming cases.

Question M

Mei Lin has likely fallen into a small/biased-sample error: her judgement is based on a short period and a limited number of interactions with a handful of new neighbours (small, non-random sample), which may not represent the whole island's population. She may also have unconsciously expected unfriendliness because of a (possibly unfair) stereotype or her prior experience, so any friendliness felt more surprising than it objectively was (an expectation/anchoring bias affecting her interpretation of a small sample of encounters).

Question N

a) The trial had two comparison (control-type) groups: the group given the standard anti-sickness medication (an active control, benchmarking the herbal supplement against a treatment already known to work) and the group given pills with no active ingredient (a placebo/inactive control). The purpose of the placebo pills was to allow researchers to check whether any apparent benefit of the herbal supplement was greater than the effect of simply believing you are being treated (the placebo effect), by comparing outcomes in the placebo group with the herbal-supplement group.

b) Informing participants of the three possible treatments (without telling them which one they personally received) reflects the ethical requirement of informed consent – passengers had a right to know the general nature of what they might be given (e.g. in case of allergies or known side effects, and simply to consent to taking part in an experiment at all), even though revealing which specific treatment each person received would have undermined the blinding needed for a valid trial.

4.5 Misleading with data

Activity 3

a) The bar chart has no axis values shown and a y-axis that does not start at 0, so a rise from 28% to 33.8% (an increase of under 6 percentage points, about a 21% relative increase) is stretched to look like a dramatic, possibly multiple-fold jump – a classic truncated-axis distortion used to make the headline ("Drastic effects") seem justified.

b) There is no chart at all, and the framing ("Viewers undecided") misrepresents the figures: 46% "very likely" + 29% "likely" = 75% think the policy is fair, against only 25% "not likely" – a clear majority in favour, not an undecided public. This is misleading through spin/selective wording rather than through a graphical trick.

c) The line graph plots four points at unevenly-spaced actual time intervals (Mar 09 to Sep 09 is 6 months; Sep 09 to Jan 11 is 16 months; Jan 11 to Aug 12 is 19 months) as if evenly spaced along the x-axis, and uses unevenly-spaced y-axis labels (10, 13, 18, 24). Both distortions make the rise in complaints look like a smooth, steady acceleration ("soaring"), when the true rate of increase per unit time cannot be judged from the graph as drawn.

d) A cumulative total can never decrease, so a chart of cumulative revenue will always show an upward-sloping (or flat) line even in months where actual (non-cumulative) sales are falling – "revenue continues to grow" is therefore misleading; the underlying monthly figures could be declining while the running total still rises.

e) A pie chart implies the slices are parts of one whole totalling 100%, but $48 + 55 + 73 = 176\%$, far more than 100. The three percentages must relate to three separate measures (e.g. approval ratings for three different candidates measured independently, where respondents could support more than one, or answer separate yes/no questions) rather than shares of a single race, so presenting them as pie slices falsely implies they are mutually exclusive shares of one contest.

f) Inverting the y-axis (running from 30 at the top down to 0 at the bottom) reverses the visual impression of the trend: a line that is actually falling will appear to rise up the page, and vice versa, misleading any reader who assumes (as is conventional) that higher up means a higher value.

Question O

a) The relative claim ("40% increase", from 5 to 7 injuries) is arithmetically correct but is based on very small absolute numbers (only 2 extra injuries), so expressing it as a large percentage change makes it sound far more alarming ("casualties soar") than the real-world change warrants; no context is given on how many cyclists actually used the towpaths in each year, and the comparison with roads (412 to 458, an $\approx 11\%$ increase but a much larger absolute rise of 46) is not highlighted, even though roads plainly carry a far greater absolute injury burden.

b) The number of injuries could rise even with towpaths just as safe (i.e. the same risk per cyclist/per journey) if the number of people cycling on towpaths has increased (e.g. due to greater popularity of cycling, more towpaths being open, better weather). Without a rate (injuries per cyclist, or per mile cycled), a rise in the raw count tells us nothing about whether the towpaths themselves have become more dangerous.

Question P

a) The headline ("Your toddler's babbling may reveal a speech delay") presents a small, preliminary piece of research as if it offers a ready, practical diagnostic tool for parents, which overstates what a single small study with one relevant finding can support.

b) Vague words such as "may have had a breakthrough", "a small number", "poorly understood", and "could enable" are unquantified and non-committal, signalling that the actual strength and

size of the finding is being glossed over.

c) This baby would have been a false positive: they would have been identified (based on low syllable variety) as being at risk of speech delay despite not going on to have one, so their family could have been put through unnecessary anxiety, and the "extensive practical, emotional and financial commitment" of early intervention described at the start of the article would have been undertaken needlessly for a child who did not need it.

d) The evidence is very weak: of 14 babies studied, only one was later diagnosed with speech delay, and that child did not even have the lowest syllable-variety score (only the third lowest) – so the supposed marker did not correctly identify the one true case as the most extreme, and with a sample this small (one positive case among 14) no reliable statistical conclusion can be drawn about the marker's predictive power. The research finding is far too weak and preliminary to justify the confident, practically-oriented tone of the newspaper headline and article.

4.6 Coincidence?

Activity 4

The three people had each visited all four earthquake-affected regions around the time of the earthquakes, which sounds remarkable, but this is easily explained if they share (or have shared) an occupation that routinely takes people to disaster-prone regions either shortly before an event (e.g. geologists, seismologists, or risk assessors visiting to study or advise on hazard) or shortly afterwards (e.g. aid workers, journalists, or disaster-relief professionals visiting after a disaster). A shared underlying factor (profession) explains the correlation between their travels and the earthquakes far more plausibly than coincidence or any special premonition.

Activity 5

a) Likely a third factor rather than a direct causal link between Devon moving in and more foxes: e.g. a change in local food waste/bin habits generally, a seasonal or wider urban trend in fox numbers, or new building/food sources elsewhere pushing foxes into the area, coinciding by chance with Devon's arrival. (If foxes genuinely did increase, then more foxes plausibly causes more raided bins – but this is a coincidence of timing with Devon's arrival, not something Devon caused.)

b) Plausibly a third factor: a candidate's conscientiousness, organisation, punctuality and enthusiasm could cause both arriving early and performing well/being well regarded in the interview, rather than early arrival itself causing the job offer (though a minor direct effect, e.g. making a good first impression, is also possible).

c) A genuine causal link is plausible (A can cause B: strong engineering research feeds directly into patentable innovation), but a third factor is also likely, such as a country's overall investment in R&D, wealth/GDP, or education system, which drives both strong university research facilities and high patent registration.

d) Third factor, not causation: both are more likely driven by increased sun exposure/outdoor behaviour and by increased awareness/screening (more people getting checked for skin cancer, leading to more diagnoses, while also buying more sunscreen), and by long-term trends (population growth, more holidays in the sun) affecting both variables over time. There is no plausible mechanism by which sunscreen use would cause melanoma.

e) Third factor / spurious time-trend correlation: both internet usage and obesity have risen worldwide over recent decades alongside broader trends such as economic development, urbanisation and more sedentary lifestyles; these underlying trends plausibly drive both variables rather than internet use directly causing obesity (though a minor indirect mechanism – more screen time meaning less physical activity – could contribute a small genuine causal effect too).

4.7 Critical analysis of models

Activity 6

a) Factors affecting energy received from the Sun: variation in solar output/solar activity cycles, changes in the Earth's distance from the Sun and orbital eccentricity, axial tilt and precession (Milankovitch cycles), and the amount of incoming sunlight reflected back to space by clouds, ice, or aerosols (albedo) before it can be absorbed.

b) Factors affecting energy radiated from the Earth: the concentration of greenhouse gases (CO_2, methane, water vapour) in the atmosphere, which trap outgoing infrared radiation; cloud cover; and the reflectivity/absorptivity of the Earth's surface (e.g. ice and snow cover vs. ocean or forest), which affects how much energy is absorbed and re-radiated.

Question S

a) The major flaw is inferring a single direction of causation (more CO_2 causes higher temperature) purely from correlation, while ignoring the very lag mentioned just before the extract: historically, CO_2 levels have tended to *follow* temperature changes by several centuries, suggesting that in the ice-core record temperature change came first, with CO_2 acting as an amplifying feedback rather than the sole/initial cause. The commentator also ignores the possibility that a third factor (e.g. Milankovitch orbital cycles altering solar energy received) drives both temperature and, indirectly, CO_2 release from the oceans.

b) Yes – but this cannot be established simply by reading correlation off the two graphs; it rests on independent physical/chemical evidence (the known radiative/greenhouse properties of CO_2 molecules, demonstrated in laboratory physics and atmospheric science) showing a plausible causal mechanism by which added CO_2 traps additional outgoing radiation and warms the Earth. So while the specific ice-core correlation argument as presented is flawed as a stand-alone proof, the broader claim that CO_2 increases can cause warming is supported by separate scientific evidence.

Consolidation Exercise 4

Question 1

The pictogram uses currency symbols of increasing number/size ("£" vs. "£££") to represent bills tripling, with no numerical axis or actual monetary figures given. This is misleading because such pictograms are typically perceived by their visual area rather than by the count of symbols alone; if the larger set of symbols is also drawn bigger or more densely, the perceived increase can look far greater than an actual threefold rise (a 2-D area scaling of a 1-D value change). There are no underlying numbers (e.g. actual pound-figures per household, or the time period/type of bill), so the reader cannot judge whether "triple" refers to typical, average, or a cherry-picked example, nor whether inflation or usage changes have been accounted for.

Question 2

a) Evidence: studies showing that larger vehicles cause more road wear *on average*. Conclusion: it is fair for drivers of larger vehicles to pay higher road tax because "driving a larger vehicle increases damage to infrastructure".

b) It does not follow for an individual driver. The evidence is about an *average* across the population of larger vehicles; applying an average, population-level tendency to justify a claim about what *any given* individual's driving does ("driving a larger vehicle increases damage") is an unjustified generalisation (an ecological-fallacy-type leap). A specific driver of a larger vehicle might drive very little, drive gently, or mostly on already-robust roads, so may cause no more (or even less) wear than an average smaller-vehicle driver; the average could also be skewed by a subset of

very heavy vehicles (e.g. vans/lorries) within the "larger vehicle" category rather than applying evenly to, say, large family cars or SUVs.

Question 3

a) The categories mix two different, inconsistent scales: "Dissatisfied / Satisfied / Very satisfied" is one ordered scale (about satisfaction), while "Good / Excellent" appears to be a different scale (perhaps rating quality), with no clear relationship between the two – it is unclear, for example, whether "Good" should count as more or less positive than "Satisfied". There is also no neutral or "very dissatisfied" option, and the category boundaries are not defined, so different students may interpret them inconsistently.

b) Tomas has combined everyone except the "Dissatisfied" respondents: $9 + 5 + 7 + 4 = 25$ out of a total of $4 + 9 + 5 + 7 + 4 = 29$ respondents, giving $\frac{25}{29} \approx 86\%$.

c) Reasons to doubt the 86% figure: (i) the categories are ambiguous/inconsistent, as identified in part (a), so simply adding four different labels together to mean "happy" is questionable; (ii) the sample is small (29 out of 280 students, about 10%) and, more importantly, hugely disproportionate – 18 of the 29 surveyed (62%) study Further Maths, whereas Further Maths students make up only $\frac{84}{280} = 30\%$ of the year group, so the sample over-represents a group that may have systematically different (here, more positive) views, biasing the overall figure.

d) Satisfaction ("not dissatisfied", matching Tomas's own definition) within each group: Further Maths: $\frac{16}{18} \approx 88.9\%$; not Further Maths: $\frac{9}{11} \approx 81.8\%$. Weighting by the true population proportions ($\frac{84}{280} = 30\%$ study Further Maths, $\frac{196}{280} = 70\%$ do not):

$$0.3 \times 88.9\% + 0.7 \times 81.8\% \approx 26.7\% + 57.3\% = 84.0\%.$$

So a better, representative estimate is about 84%, slightly lower than Tomas's unweighted 86%, because the oversampled Further Maths group was slightly more satisfied than the rest of the year.

Question 4

a) Surgical gloves: $\frac{176}{800} \times 100 = 22\%$. Legal services: $\frac{520}{5200} \times 100 = 10\%$.

b) Evidence: the consultancy's findings that savings of 22% and 10% were possible in the two specific spending areas examined (gloves and legal services). Reasoning: these sample savings were generalised into an overall "up to 20%" saving across the whole procurement budget. Conclusion: applying 20% to the full £30 billion budget gives the headline figure of £6 billion wasted per year.

c) The headline is not supported because the report only examined two narrow categories of spending (gloves and legal services), totalling just £6 million out of a £30 billion budget – around 0.02% of total spend – and the "up to 20%" figure is a maximum/best-case figure for those two categories, not a proven average across the enormous range of other categories (e.g. drugs, medical equipment, staffing-related procurement, energy, building maintenance) that make up the vast majority of the budget and were not studied at all.

d) Combined spending on the two areas: 800+5200 = 6000 (£k). Combined saving: 176+520 = 696 (£k). Percentage saving on the combined total: $\frac{696}{6000} \times 100 = 11.6\%$. Applying this to the full budget: $0.116 \times £30$ billion $\approx £3.48$ billion.

e) Even this recalculated £3.48 billion figure should be treated with caution: it still rests on extrapolating from just two very different categories (with quite different saving percentages, 22% and 10%, showing how much saving rates can vary by category) to the whole, much larger and more diverse budget. Without examining a properly representative sample of the many other procurement categories, any single extrapolated overall saving percentage – whether 20% or the recalculated 11.6% – is an unreliable, unjustified generalisation.

Investigation and Review

The "Critical analysis of your own work", "Review" checklist, and "Investigation: Public health policy" sections are reflective/open research tasks with no single correct answer, so no formal marking points are given here. Credit should be based on whether the candidate's own report/summary demonstrably applies the skills covered in this chapter: identifying evidence, reasoning and conclusion; noting emotive/vague language, unjustified assumptions or self-contradiction; checking for selective use of data, sampling or trialling flaws, misleading graphs, and confusing correlation with causation.