

Baron-Cohen et al. (eyes test) AS

Read online at pastopic.com/cambridge/psychology-9990/notes/baron-cohen-et-al

Last checked 5 October 2026

© 2026 Pastopic. Free for personal study. Please share the link, not the file.

What you need to know

The syllabus asks the same things of every core study. For Baron-Cohen et al. (2001), which tested a revised version of the "Reading the Mind in the Eyes" test on adults with and without Asperger syndrome or high-functioning autism (AS/HFA), you should be able to:

1. **Explain the psychology being investigated: theory of mind and social sensitivity.**
2. **Give the background:** what the original 1997 eyes test was, and why it needed improving.
3. **State the aims** of the study.
4. **Describe the procedure and its methodology:** the research methods (a natural experiment comparing groups, plus correlations), the four groups (size, who they were, how they were recruited), the experimental design, how the revised test was built and taken, the controls (the glossary, the gender-recognition task, IQ matching), the other measures (the Autism Spectrum Quotient and IQ), and the variables.
5. **Discuss the ethical issues**, such as possible distress for the AS/HFA group, consent and confidentiality.
6. **Give the results:** the eyes test scores of each group, the sex differences, the AQ results, the correlation between the two tests, and how the results could be shown and read (graphs, means, standard deviations).
7. **State the conclusions** that were or could be drawn.
8. **Evaluate every part of the study:** its methods, sample, design, data, validity, reliability, generalisability and ethics.
9. **Relate it to the AS issues and debates:** mainly the application of psychology to everyday life, individual and situational explanations, and nature versus nurture. Only adults took part and no animals were used, so the debates about children and animals apply only as contrasts (for example, problems if children had been tested).
10. **Explain why it belongs to the cognitive approach**, using its two assumptions, and compare it with the other cognitive core studies, Andrade and Pozzulo et al.

Understand it

The psychology: theory of mind and social sensitivity

Theory of mind (ToM) is the ability to work out mental states, such as beliefs, wishes, intentions and feelings, in **yourself and in other people**. With it we can:

- realise that other people can think and feel differently from us;
- **explain** why someone did something ("she slammed the door because she was angry");
- **predict** what someone will do next;
- put ourselves in someone else's shoes, which is closely linked to **empathy**, understanding how another person feels.

Simple ToM tests made for young children, such as the Sally-Anne false-belief task, are too easy for most adults, so they can't show small differences in adult mind-reading. The eyes test was built as an **advanced, adult** test of ToM: you look at a photo of a person's eye region and decide what they are thinking or feeling.

Social sensitivity is being able to pick up and understand the feelings of other people **in a social situation**, using cues such as facial expressions, eyes and body language, and to respond to them appropriately. It has two parts, and both are needed for a full definition: the **empathic** part (recognising that others have feelings) and the **social situation** in which you notice them. It does not mean being "sensitive" as a personality trait or being easily upset.

Asperger syndrome and high-functioning autism are conditions on the autism spectrum. People with AS/HFA have average or above-average intelligence, but often find social interaction hard, and one explanation is a weaker theory of mind. In "the psychology being investigated", a description of AS/HFA earns at most one mark: write mainly about ToM and social sensitivity.

Background: the original eyes test and its problems

The **original eyes test (1997)** was a questionnaire completed online, and worked like this:

- **25** photos of the eye region, taken from pictures of famous people, both actors and actresses;
- each photo had **two** words to choose from, and the two were **opposites** (one positive, one negative);
- the mental states included **basic** emotions (such as happy or sad) as well as complex ones;
- the participant chose the word that best described what the person was thinking or feeling.

The researchers listed **eight problems** with the validity of the original test. The ones exam questions ask about are:

Problem with the original	Why it hurt validity
Only two response options (a forced choice)	A guesser is right about half the time, so scores partly show luck, not mind-reading
The two words were opposites	You only had to tell positive from negative, which is easy

Problem with the original	Why it hurt validity
It included basic mental states	Basic emotions are too easy, so almost everyone scores high (a ceiling effect)
Only 25 items	The range of possible scores is narrow, so it can't separate people well
Some words were hard to understand	A wrong answer might show poor vocabulary, not poor ToM

All of these make the test **poor at discriminating**: too many people, with and without AS/HFA, end up with similar scores.

The aims

- To **test a revised version of the eyes test** to see whether it fixes the problems of the original and still shows a deficit in adults with AS/HFA.
- To see whether **adults with AS/HFA are impaired** on the revised test, that is, whether they show a weaker theory of mind than other adults.
- To investigate **sex differences**: whether females score higher than males on the eyes test (in the general population and students).
- To see whether eyes test scores are **negatively correlated** with scores on the **Autism Spectrum Quotient** (AQ), a measure of autistic traits.

The revised eyes test

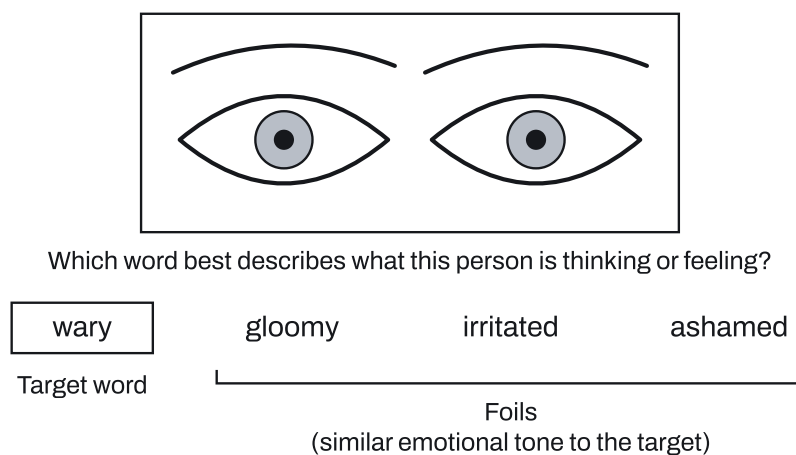


Illustration only: not a real item from the test

Feature	Original (1997)	Revised (2001)
Number of photos (items)	25	36 pairs of eyes
Words per photo	2, opposites	4: one target word and three foils

Feature	Original (1997)	Revised (2001)
Emotional tone of the words	Opposite	The foils have a similar valence (emotional tone) to the target
Mental states	Basic and complex	Complex mental states
Help with words	None	A glossary of the words, with definitions and examples

Choosing the target words and foils (before the main study): each item was piloted on **eight judges**. An item was kept only if **at least five** of the eight agreed that the target word fitted best, and **no more than two** of them chose any single foil. Items that failed got new words and were piloted again.

The glossary: participants were told to read through the glossary before starting, to mark any words whose meaning they were unsure of, and to look those words up whenever they needed to during the test. This made sure a low score came from not reading the eyes, not from not knowing the words.

Why the changes improve validity: four options cut the chance of a lucky guess from 1 in 2 to 1 in 4; 36 items widen the range of possible scores; similar-valence foils and complex states make the items harder, so scores are not all squeezed at the top; the glossary removes vocabulary as a confounding variable. Together these let the test **discriminate** between people better.

A participant who guessed every item would expect to get the number of items divided by the number of options right: $25 \div 2 = 12.5$ on the original and $36 \div 4 = 9$ on the revised test.

The sample: four groups

Group	Who	Size	Recruitment	Details
1	Adults with AS/HFA	15 , all male	Volunteer (self-selecting) sample: replied to adverts in an autism society's magazine	Mean age 29.7 years; mean IQ 115 (normal range); similar social class and education to Group 2
2	Adults from the general population	122	Adult community education classes in Exeter and public library users in Cambridge (opportunity sampling is accepted)	Mean age about 28 (ages known for only 88 of them); a broad range of jobs and education levels; similar class and education to Group 1
3	Students	103 , equal numbers of males and females	Cambridge University undergraduates	Mostly studying science; mean age 20.8 ; A-grade A levels, so assumed to have high IQ

Group	Who	Size	Recruitment	Details
4	Adults from the general population , IQ-matched	14 , all male	Randomly selected	Mean IQ 116 , matched to Group 1; mean age 28 , the same age spread as Group 1

No one in Groups 2, 3 or 4 was diagnosed with AS/HFA. The groups add up to $15 + 122 + 103 + 14 = 254$ participants.

Group 1 and Group 4 compared: both were all male, adults of a similar age, small (15 and 14) and matched on IQ (115 and 116). They differed in **how they were selected:** Group 1 volunteered, Group 4 was chosen at random.

The procedure and the other measures

1. **Every group** took the **revised eyes test:** for each of the 36 photos, the participant chose which of the four words best described what the person in the photo was thinking or feeling. The glossary was available throughout.
2. **Group 1 only** also did a **control task:** for each photo they judged the **gender** (sex) of the person. This checked that they could use information from eyes to make a simple judgement, so a low eyes test score could not be blamed on being unable to see or use the photos at all.
3. Groups 1, 3 and 4 completed the **Autism Spectrum Quotient (AQ)**, a self-report questionnaire that measures **autistic traits**; a **higher** AQ score means **more** autistic traits. Group 2 did **not** do the AQ.
4. **IQ** was measured for Groups 1 and 4 with a standardised IQ test (the mark scheme accepts the WASI-R). Group 3's IQ was assumed to be high; Group 2's IQ was not measured.

Variables, design and methods

- **Research method:** an **experiment** with a **naturally occurring IV**, often called a **natural** (or quasi-) experiment. The researchers could not assign people to have AS/HFA, so they compared groups that already existed. Don't call it a **laboratory** experiment; mark schemes don't credit that. The study also used **correlations** (eyes test scores with AQ scores).
- **Independent variables:** whether a participant had a diagnosis of **AS/HFA** or not (Group 1 versus the other groups); and **sex** (male or female, compared in Groups 2 and 3).
- **Dependent variables:** the **eyes test score** (number of items correct, out of 36) and the **AQ score**.
- **Experimental design:** **independent measures**, since each person was in only one group.
- **Controls:** the same 36 photos, the same four words per photo and the same instructions for everyone (a **standardised** procedure); the glossary; Group 4 matched to Group 1 on IQ and age; the gender-recognition task for Group 1.
- **Data:** **quantitative** (scores on the eyes test, AQ and IQ).

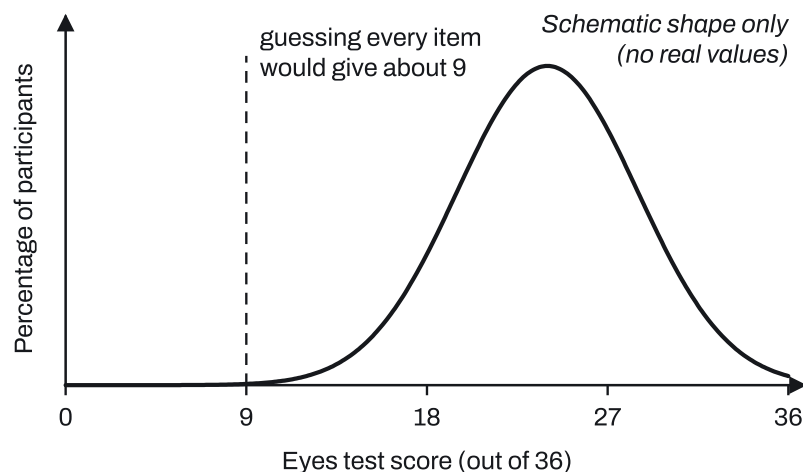
Ethical issues

- **Minimising harm:** this is the guideline the study is weakest on. Adults with AS/HFA, who were expected to find the task hard, might have felt stressed or upset by struggling with it or by knowing they were expected to score low. Hard words, even with the glossary, could add to that.
- **Informed consent:** participants agreed to take part (Group 1 actively volunteered).
- **Confidentiality:** participants' names were not published.
- The right to withdraw, deception, privacy and debriefing are **not mentioned** in the published paper, so don't claim the study kept or broke them.

The results

The eyes test

- The **AS/HFA group (Group 1)** scored **lowest** of all the groups, significantly lower than the other three. Their mean was **21.9** out of 36; the general population (Group 2) averaged between **26 and 27**.
- **Sex differences:** in Groups 2 and 3, **females** on average scored **higher than males** on the eyes test. The gap was small (mark schemes also accept that in Group 2 the two sexes scored very similarly). Groups 1 and 4 were all male, so no sex comparison is possible for them.
- The **students (Group 3)** tended to score higher than the general population (Group 2).
- The scores spread out in a **normal distribution**: when the percentage of participants getting each score is plotted, most people sit near the middle and fewer at the low and high ends.



The AQ

- The **AS/HFA group** scored **highest** on the AQ (the most autistic traits).
- In the students (Group 3), **males** scored **higher than females** on the AQ.
- Groups 3 and 4 scored roughly the same as each other.
- Across participants, AQ scores were **negatively correlated** with eyes test scores: the more autistic traits a person reported, the lower their eyes test score tended to be.

Spread (standard deviation)

A **standard deviation** (SD) measures how spread out the scores are around the mean: the bigger the SD, the more the scores vary. Two examples from the study:

What was measured	Group 1 (AS/HFA)	Comparison group
Age: mean (SD)	29.7 (14.5)	Group 3 students: 20.8 (0.8)
SD of AQ scores	6.0	Group 4 (IQ-matched): 2.9

So Group 1 was **older** on average than the students, and their ages were **far more spread out** (the students were almost all about 20). Group 1's AQ scores were also **more variable** than Group 4's.

The conclusions

- Adults with AS/HFA have an **underdeveloped (impaired) theory of mind**: they find it harder to work out mental states from the eyes, even though their intelligence is in the normal range.
- There is a **sex difference** in this kind of mind-reading: on average, **females are better** than males at judging mental states from eyes, and males report more autistic traits.
- The revised eyes test is a **more valid** advanced test of ToM than the original, because it separates people better.

Why it is cognitive

The cognitive approach has two assumptions (syllabus):

- Everyone handles information in the **same sequence**, much as a **computer** does: information comes in (**input**), the mind works on it (**process**), and a response comes out (**output**).
- People **differ from one another** in how well these mental processes work (for example **attention, language, thinking** and **memory**), and those differences can account for differences in their **behaviour and emotion**.

Baron-Cohen et al. fits both. Each participant took in a photo of eyes (**input**), worked out the mental state shown (**process**) and chose one of the four words (**output**). And the AS/HFA group scored lower than the other groups, which shows that a difference between people (in social behaviour) can be explained by a difference in a **cognitive process**, here theory of mind: they took in the same information but found it harder to process.

Key facts and formulas

There is no formula list in the 9990 exam. "Learn it" means you must remember it.

Fact	Status
Baron-Cohen et al. groups: 1 = 15 male adults with AS/HFA (volunteers); 2 = 122 general population (Exeter classes, Cambridge library users); 3 = 103 Cambridge students; 4 = 14 randomly selected adults, IQ-matched to Group 1	Learn it
Baron-Cohen et al. Group 1: mean age 29.7, mean IQ 115; Group 4: mean age 28, mean IQ 116; Group 3: mean age 20.8	Learn it
Original eyes test: 25 photos of eyes, 2 opposite words, basic and complex states. Revised: 36 photos, 4 words (1 target, 3 foils of similar valence), complex states, a glossary	Learn it
Baron-Cohen et al. word selection: 8 judges; at least 5 agree on the target; no more than 2 choose any one foil	Learn it
Baron-Cohen et al. results: AS/HFA lowest on the eyes test (mean 21.9; Group 2 about 26–27) and highest on the AQ; females above males on the eyes test; males above females on the AQ (students); AQ negatively correlated with the eyes test	Learn it
Expected score from guessing = $\frac{\text{number of items}}{\text{number of options per item}}$	Learn it
Percentage = $\frac{\text{score}}{\text{maximum score}} \times 100$	Learn it
Mean = $\frac{\text{sum of the scores}}{\text{number of scores}}$; range = highest score – lowest score	Learn it
A larger standard deviation means the scores are more spread out around the mean	Learn it

How to do it

We have tagged 38 questions on this study, from 2021 to 2026 (Papers 1 and 2). 18 of them have parts of more than one type, so the counts add up to 58.

Question type	Questions
Describe or identify the sample	13
Describe the eyes test and procedure	10
Explain one strength or weakness in context	10
Outline a result or conclusion	4
Outline an aim or hypothesis	3
Define the psychology behind the study	3
Evaluate the study: two strengths and two weaknesses	3
Suggest a real-world application	3

Question type	Questions
Interpret descriptive statistics or a graph	3
Problems with using children as participants	2
Similarities with another cognitive study	2
Link a cognitive-approach assumption to the study	2

1. Describe or identify the sample (13 questions)

One mark per correct feature, from a fixed list, so learn the table in "The sample" exactly.

1. **Check which group** is asked about. The most common mistake is giving the total sample, or the size of another group.
2. **Give separate, specific points**, one per mark: size, sex, diagnosis, recruitment, age, IQ, where from.
3. **Don't repeat what the question gives you**: if it says "15 participants, all male", give two **other** features (adults, AS/HFA diagnosis, volunteers via a magazine advert, mean age 29.7, mean IQ 115, similar class and education to Group 2).

Variations you will meet:

- **"State the number"**: Group 1 is **15**; Group 4 is **14**; Group 2 is **122**; Group 3 is **103**. Only the first answer is marked.
- **The sampling technique for Group 1: volunteer** (self-selecting). "Newspaper advert" and "selective" earn nothing.
- **Characteristics** versus **features**: characteristics are things about the people (sex, age, diagnosis, IQ); features can also include the size and sampling technique. If asked for characteristics, don't give the sampling technique.
- **Similarities between Groups 1 and 4**: all male; adults of similar age; matched IQ (115 and 116); small (15 and 14). **Difference in selection**: Group 1 volunteered; Group 4 was randomly selected.
- **Group 3**: undergraduates at Cambridge; mostly science; 103; equal numbers of males and females; mean age 20.8; A-grade A levels so assumed high IQ; not diagnosed with AS/HFA. Don't give "males and females" if the question already says so.

2. Describe the eyes test and procedure (10 questions)

Short, exact recall: one mark per point.

1. **Decide which version** is asked about: the **original** (25 photos of famous actors and actresses, 2 opposite words, basic and complex states) or the **revised** (36 photos, 4 words, similar-valence foils, complex states, glossary).
2. **Match the fix to the problem**: 25 items → **36**; two options → **four**; opposites → **foils with similar valence**; hard words → **a glossary**. "Using male and female eyes" or "simpler words" are wrong.

3. **"Two ways the range of scores was increased"**: more items (**36**) and more options (**four**). Other changes, such as the glossary, don't answer this.

Variations you will meet:

- **How the target words and foils were chosen** (before the main study): eight judges; at least five had to agree on the target; no more than two could choose any one foil; failed items got new words and were piloted again. "Four words instead of two" is about the main test, not the selection, and earns nothing here.
- **The control task: gender recognition** for Group 1, deciding the sex of the person in each photo. "Saying whether a person was male or female" needs "in each photo (of eyes)" for the mark.
- **The glossary instructions**: read it through first; mark words you are unsure of; read those meanings; refer to it throughout the test. "They were told to use it" repeats the question.
- **The other tests**: the **Autism Spectrum Quotient (AQ)** and an **IQ** test. The AQ is the one that was **negatively correlated** with the eyes test (not IQ). Gender recognition is a task, not a test.
- **One strength of the AQ or IQ test** (identify only): quantitative scores that are easy to compare; published and standardised, so it can be repeated and checked for reliability.

3. Explain one strength or weakness in context (10 questions)

Name the issue (1 mark), then explain it with a detail from **this** study (1 mark). A generic point with no detail from Baron-Cohen et al. gets at most 1.

Point	Strength or weakness	Example from the study
Independent measures: pre-existing groups	Strength	A person either has AS/HFA or doesn't, so they could only ever be in one group
Independent measures: no order effects or demand characteristics	Strength	Each person took the eyes test once, so practice and guessing the aim from another condition weren't possible
Independent measures: participant variables	Weakness	The AS/HFA group might differ from controls in other ways, such as less social experience, which could lower scores instead of ToM itself
Independent measures: sample size	Weakness	It needs more people than repeated measures, which is hard when the group (AS/HFA adults) is hard to find, so Group 1 was only 15
Standardisation (reliability)	Strength	The same 36 photos, four words per photo and glossary for everyone, so another researcher can repeat it and check the results
Controls (internal validity)	Strength	Glossary (so vocabulary doesn't affect scores); Group 4 matched on IQ; gender task (so poor eyesight or attention doesn't explain low scores)

Point	Strength or weakness	Example from the study
The revised test	Strength	Quantitative scores out of 36 can be compared between AS/HFA and control groups; it was published, so others can test new samples
Generalisability	Weakness	Group 1 was 15 male volunteers; Group 3 were Cambridge students, mostly scientists; Group 2 came from evening classes and libraries
Ecological validity	Weakness	Static photos of eyes only; in real life eyes move and we also see the whole face, body and hear the voice
Does it measure ToM?	Weakness	Choosing a word may only be matching pictures to words, and with four options a person can still guess some right

Variations you will meet:

- **Ecological validity** in three steps: **define** it (whether findings from a study generalise to real-life settings); **one problem** with the revised test (eyes are static; only the eyes are shown; no body language); **one fix** linked to that problem (video of moving eyes; photos of the whole face or body). "It was in a lab" and "it's not natural" earn nothing.
- **"Explain why the findings are not generalisable"** (4 marks): define generalisability (1), then up to three points of evidence from different groups: all-male Group 1, so not females with AS/HFA; volunteers, who may differ from non-volunteers; Group 2 from only education classes and library users; Group 3 from one prestigious university and mostly science students.
- **"Explain why each problem affected validity"** (the forced choice or basic states): identify the effect (guessing gives correct answers by chance; basic states are too easy, a ceiling effect), then explain (with two choices, half the answers can be right by luck; nearly everyone scores high, so the test can't separate people). For **how solving it improves validity**, both marks must be about the **same** problem: four options, so guessing is less likely and scores show real ability; or complex states only, so the test detects a wider range of ability.
- **"Explain one reason the procedure was standardised"** (3 marks): reason (so it can be replicated) → why that matters (so its reliability can be tested) → in this study (the 36 photos and four words are known, so the test can be repeated exactly). Or: fewer extraneous variables → so cause and effect can be seen → it is probably AS/HFA (the IV) changing the eyes test score (the DV). The question asks **why**, so a list of **how** it was standardised earns nothing.
- **Valid and not valid** (two friends disagree, 3 marks each): name the type of validity (1), then two linked points of evidence. "Has validity": internal validity from controls (same photos and words for all, the glossary, IQ matching) so the difference is likely due to AS/HFA; the revised test was designed to fix the original's problems. "Not valid": low ecological validity (static eyes, no other social cues); low population validity (15 males in Group 1). "People could guess" earns at most 1.

4. Outline a result or conclusion (4 questions)

1. **A result needs a meaningful comparison** for 2 marks: two groups (or sexes) and the direction, **and the name of the test**. "Females scored higher than males on the eyes test" (2). "Females scored higher" (0).
2. **Never** write "better", "worse" or "less able"; write "scored higher" or "scored lower".
3. **Use data when asked**: Group 1's mean of 21.9 on the eyes test (accepted 21–22), or Group 2 between 26 and 27.

Variations you will meet:

- **A result about sex**: females above males on the **eyes test** (Groups 2 and 3); males above females on the **AQ** (Group 3). Never compare sexes in Group 1 or Group 4: they were all male.
- **A result about the AQ**: the AS/HFA group scored higher than Group 3 or Group 4; males in Group 3 above females in Group 3; Groups 3 and 4 roughly equal. Group 2 did **not** do the AQ. The negative correlation with the eyes test is a result too, but if a question has already given it, it scores at most 1.
- **Group 1 compared with Group 2**: it can only be about the **eyes test** (Group 2 did no AQ or IQ test): "the AS/HFA group scored lower on the eyes test than the general population, with a mean of 21.9".
- **A conclusion** (1 mark): people with AS/HFA have an underdeveloped theory of mind; or there is a sex difference, with females better than males at reading mental states from eyes. A conclusion is a general claim, not a number.

5. Outline an aim or hypothesis (3 questions)

- **An aim** (2 marks): say what is tested and on whom: "to see whether adults with AS/HFA show a deficit on the revised eyes test compared with adults without AS/HFA"; "to investigate whether females score higher than males on the revised eyes test". "To improve the eyes test" or "to test for ToM" alone is brief (1). A finding earns nothing.
- **If one aim is given** (such as the AQ correlation), give a **different** one.
- **A hypothesis** (2 marks): a prediction naming the **group** and the **test**: "females in the general population (Group 2) will score higher than males on the eyes test" or "male students will score higher than female students on the AQ". Without the group it is brief (1); without the test it earns 0. Don't write a result in the past tense.

6. Define the psychology behind the study (3 questions)

1. **Theory of mind** (up to 4 marks, one per point): attributing mental states to ourselves and others; mental states include desires, beliefs and emotions; linked to empathy; done by imagining what another person thinks or feels; used to **explain** others' behaviour; used to **predict** it; understanding that others can think differently from us. One mark can go to an example from the study: the eyes test, where people judge mental states from the eyes alone.
2. **Social sensitivity** (2 marks): both parts, the **empathic** part (identifying or understanding other people's feelings) **and** the **social situation**, such as noticing facial expressions or body language and responding appropriately. One part alone is brief (1); "being socially sensitive" or "being easily hurt" earns nothing.

3. **"Describe the psychology being investigated"** (4 marks): points about ToM and social sensitivity (and what makes a test valid and reliable), plus one example mark from the study. Describing AS/HFA earns at most 1.

7. Evaluate the study: two strengths and two weaknesses (3 questions)

1. Plan four points of **equal depth**, one of which is the **named issue** (quantitative data, ethics, generalisations, validity and so on). Without it the answer is capped at Level 3 (6 marks).
2. For each point: name it, say whether it is a strength or weakness of **this** study, give **at least two** specific details, and explain **why** it matters for the conclusions.
3. **Useful strengths**: standardisation and reliability; controls and internal validity; quantitative data. **Useful weaknesses**: ecological validity; generalisability; ethics (minimising harm); participant variables between groups.
4. **Quantitative data** can be either side: a strength (eyes test scores out of 36 allow objective comparison of all four groups and of the sexes, and a correlation with AQ) or a weakness (a score doesn't show **why** someone chose a word or how they felt doing the test).
5. **Ethics**: say only what the paper reports. Consent and confidentiality were kept; minimising harm is the weakness (possible distress for the AS/HFA group); don't claim anything about withdrawal, deception or debriefing.
6. Don't describe the study and don't write a learned "GRAVE" essay that misses the named issue; the examiner reports say both score little.

8. Suggest a real-world application (3 questions)

1. **What** the application is, clearly based on the study (1 mark).
2. **How** it would be done, explicitly (1 mark). "Give children the eyes test at school" isn't explicit enough.
3. Keep it **ethical** and **forward-looking**: something to do in future, not a description of the study, a future research idea or an explanation of past behaviour.

Ideas: the eyes test as one **aid in assessing** adults for AS/HFA (a low score may point to a weaker ToM, worth a full assessment by a clinician; it isn't meant as a diagnosis on its own); **training** emotion recognition, with a therapist going through photos and explaining what in the eyes shows each state; **schools** screening for students who may need support with social skills, then offering small-group sessions.

9. Interpret descriptive statistics or a graph (3 questions)

1. **Means**: say what they **show** about the groups ("Group 1 were older, on average, than Group 3"). "The mean of Group 1 was bigger" just reads the table and earns 0.
2. **Standard deviations**: a larger SD means the scores are **more spread out** around the mean, so say which group's ages or scores varied more ("Group 1's ages were much more spread out than the students"). "Group 1's SD was bigger" or "deviated more" earns 0.
3. **The graph** of eyes test results: horizontal axis **eyes test score**, vertical axis **percentage of participants**; the shape is a **normal distribution**.

10. Problems with using children as participants (2 questions)

For each problem: **identify** it (1), **elaborate** (1), then **link** it to the eyes test (1).

- **Understanding the words:** children have smaller vocabularies, so even with a glossary they may not know the words for complex mental states; a wrong answer would then show vocabulary, not ToM.
- **Understanding the task:** they might treat it as a game and choose at random, so the score wouldn't measure ToM.
- **Less social experience:** they have had fewer years of reading other people, so they would be expected to struggle with complex states anyway.
- **Ethics (protection from harm):** 36 items of hard judgements could feel like an exam and cause stress.
- **Boredom:** 36 items is long for a child, so attention may drop and they may guess.

"The results wouldn't generalise to adults" earns nothing.

11. Similarities with another cognitive study (2 questions)

The 8-mark part of Paper 1 Question 9 or 10: two similarities (or differences), **4 marks each**, marked in levels:

Level	What it needs
4	The point is well explained, with an example from both studies
3	Well explained with one study's example, or both studies used briefly
2	Brief, with an attempt at one example; or well explained with no example
1	Brief, with no examples

1. **Choose a point you can develop:** both experiments with an IV, both quantitative data, both low mundane realism, both independent measures (with Andrade).
2. **State the point, give each study's detail, then explain** what follows ("so both could look for cause and effect").
3. **Avoid weak points:** "both used humans", "both are cognitive" and "both laboratory" stay in Level 1; never call Baron-Cohen et al. a laboratory experiment; anything about the sample earns nothing if the sample is banned.

Useful points:

- **Both experiments:** Andrade manipulated doodling or not; Baron-Cohen et al.'s IV was **naturally occurring** (AS/HFA diagnosis or not); Pozzulo et al. compared children with adults and cartoon with human faces. All compared conditions on a measured DV to look for cause and effect.
- **Both quantitative:** Baron-Cohen et al. scored the eyes test out of 36; Andrade counted names and places recalled; Pozzulo et al. counted choices in line-ups. Numbers make comparisons objective.
- **Both low in mundane realism:** judging still photos of eyes; Andrade's mock phone message with shapes to shade; Pozzulo's line-ups of cartoon faces.

- **Both independent measures** (with Andrade): each Baron-Cohen participant was in one group; each Andrade participant either doodled or didn't. So neither had order effects, but participant variables could differ between groups.

12. Link a cognitive-approach assumption to the study (2 questions)

1. **The assumption** (2 marks for a full one): input → process → output, the same in everyone, like a computer; or individual differences in cognitive processing, naming **at least two** processes (attention, language, thinking, memory). Naming only one, or "we think like a computer", is brief (1).
2. **The support** (2 marks): a **result** from the study (1) and an **explicit link** to the same assumption (1). The AS/HFA group scored significantly lower on the eyes test than the other groups (result); so differences between people can be explained by differences in a cognitive process, theory of mind (link). Or: each participant saw a pair of eyes (input), worked out the mental state (process) and picked a word (output).
3. **Don't copy out your assumption again** as the link: the March 2025 report says this earned nothing.

Worked examples

Example 1

From the study by Baron-Cohen et al. (eyes test):

- (a) Describe what a participant had to do on one item of the revised eyes test. [3]
- (b) A participant answers every item of the revised test by guessing at random. Calculate the score they would expect, and compare it with a guesser on the original test. [2]
- (a) The participant looked at a photo showing only a person's eye region (1). Under it were four words, one target and three foils with a similar emotional tone (1), and they chose the word that best described what the person was thinking or feeling, looking up any unfamiliar word in the glossary (1).
- (b) Revised: $36 \div 4 = 9$ correct (1). Original: $25 \div 2 = 12.5$, so a guesser could reach half marks on the original but only a quarter on the revised test, which makes the revised score a better measure of real ability (1).

Example 2

From the study by Baron-Cohen et al. (eyes test):

- (a) Identify the independent variable and the dependent variable when Group 1 was compared with the other groups. [2]
- (b) Explain why Group 4 was matched with Group 1 on IQ. [2]
- (c) Explain why only Group 1 was asked to judge the gender of the person in each photo. [2]

(a) IV: whether the participant had a diagnosis of AS/HFA or not (**1**). DV: their score on the revised eyes test, out of 36 (**1**).

(b) Intelligence could be a participant variable that affects how well someone does on a test of words and faces (**1**). Because Group 4's mean IQ (116) was matched to Group 1's (115), a lower score in Group 1 can't be put down to lower IQ, so it is more likely to be due to AS/HFA (**1**).

(c) It was a control task to check that the AS/HFA participants could use the photos of eyes to make a simple judgement (**1**). If they could judge gender but still scored low on mental states, the low score must come from difficulty with theory of mind, not from problems seeing or attending to the photos (**1**).

Example 3

(a) Suggest one ethical real-world application based on the study by Baron-Cohen et al. (eyes test). [**2**]

(b) Explain one difference between the study by Baron-Cohen et al. (eyes test) and the study by Pozzulo et al. (line-ups). [**4**]

(a) A support service for adults with AS/HFA could run emotion-reading training sessions (**1: what**). With the client's agreement, a therapist would go through photos of eyes like those in the test, discuss which word fits and point out the cues in the eyes for each state, then check progress with a new set of photos (**1: how**).

(b) The studies differed in how the IV came about. Pozzulo et al. ran a laboratory experiment in which the researchers decided what each participant saw, such as cartoon or human faces and line-ups with or without the target. In Baron-Cohen et al. the main IV, having AS/HFA or not, could not be assigned at all: the groups already existed, so it was a natural experiment. This means Pozzulo et al. had more control over the IV and could be more confident about cause and effect, while in Baron-Cohen et al. the groups may also have differed in other ways, such as social experience. (**Level 4: well explained, with an example from both studies.**)

Example 4

Evaluate the study by Baron-Cohen et al. (eyes test) in terms of two strengths and two weaknesses. At least one of your evaluation points must be about validity. [**10**]

Plan: strengths are internal validity (the named issue) and standardisation (reliability); weaknesses are ecological validity and generalisability. Each paragraph has the point, two or more details and why it matters.

Strength 1: internal validity (the named issue). The researchers controlled variables that could otherwise explain the low scores of the AS/HFA group. A glossary meant that not knowing a word couldn't cause a wrong answer; Group 4 was matched to Group 1 on IQ (116 and 115), so intelligence couldn't explain the difference; and Group 1's gender-recognition task showed they could use the photos for simple judgements. So the difference in eyes test scores is likely to be caused by a weaker theory of mind, not by these other variables.

Strength 2: standardisation and reliability. Everyone saw the same 36 photos, with the same four words for each, chosen beforehand so that at least five of eight judges agreed on each target word. Because the items and instructions are fixed and published, another researcher could give exactly the same test to a new sample,

such as women with AS/HFA, and see whether the deficit appears again, so the reliability of the findings can be checked.

Weakness 1: ecological validity. The task was to judge still photos of the eye region only. In everyday life eyes move, and we judge feelings from the whole face, body language, voice and what is happening around us. A person who struggles with a static photo might do better in a real conversation, or the other way round, so the scores may not show how well people read minds in real social situations.

Weakness 2: generalisability. The AS/HFA group was only 15 men who volunteered through an autism society's magazine, so the findings may not apply to women with AS/HFA or to people who don't volunteer. The students were Cambridge undergraduates, mostly scientists, and Group 2 came only from education classes and library users. So the size of the group differences may not hold for the wider population.

How it would be marked: four points of equal depth, each with specific details and an explanation of why it matters, including the named issue: the top level (9–10 marks). Leaving out validity would cap it at 6.

Common mistakes

- **Giving the total sample or another group's size.** "The number in Group 1" is 15, not 254 or 122. The March 2026 report lists this as a common error.
- **Results with no test named, or with value words.** "Females scored higher" earns nothing; "females scored higher than males on the eyes test" earns 2. Never write "better", "worse" or "less able".
- **Comparing sexes in Group 1 or 4.** Both were all male, so a "female AS/HFA" result is wrong.
- **Saying Group 2 did the AQ.** Only Groups 1, 3 and 4 did; Group 2 did only the eyes test.
- **Mixing up the fixes:** "male and female eyes" or "simpler words" are not how the original's problems were solved. It was 36 items, four words, similar-valence foils and a glossary.
- **Confusing how words were chosen with how the test worked.** The judges' rule (at least five of eight agree) is the selection; "four words" is the test.
- **Giving a lay meaning of social sensitivity**, such as being shy or easily hurt. Give the empathic part and the social situation.
- **Copying out the assumption** as the "link" to a finding. Give a result, then explain how it shows a cognitive process at work.
- **Calling it a laboratory experiment.** The IV (AS/HFA) was naturally occurring.
- **Reading numbers off a table** for a statistics question. Say what the means or standard deviations **show**.
- **A generic weakness** ("not ecologically valid", "it's in a lab") with no detail from the study. Name what about the eyes test is unrealistic.
- **Applications that aren't explicit**, such as "use the eyes test in schools", or that describe the study instead of a future use.

How to get the marks

- **"State" and "identify"** need a short exact answer: a number, a name, a feature. Only the first answer is marked.
- **"Outline"** needs the point with a little detail: a result with a comparison and the test named; an aim with what was tested and on whom.
- **"Describe"** needs several separate points, one mark each, with exact numbers (no tolerance on group sizes or IQ).
- **"Explain"** needs the reason (1) and a link to Baron-Cohen et al. (1); in 3-mark versions, the reason, why it matters and the link.
- **"You must use data"**: quote a figure, such as Group 1's mean eyes test score of 21.9.
- **Statistics**: interpret, don't copy: "more spread out", "older on average".
- **Applications**: what it is and how it is done, ethical and forward-looking.
- **Comparisons** (8 marks): two points, each with an example from **both** studies and an explanation of what follows.
- **"Evaluate ... two strengths and two weaknesses"** (10 marks): four equal points, the named issue included, two or more details from the study in each, no description.

Quick check

1. State the size of each of the four groups and the total number of participants.
2. A participant gets 27 of the 36 items right on the revised eyes test. Calculate their percentage score, and how many more items they got right than a person guessing at random would expect.
3. Explain why using three foils with the same emotional tone as the target made the test better than using two opposite words.
4. In a replication, five (made-up) participants score 24, 30, 27, 21 and 28 on the revised eyes test. Calculate the mean and the range.
5. Outline one conclusion from the study about sex differences.

Answers

1. Group 1: **15**; Group 2: **122**; Group 3: **103**; Group 4: **14**. Total $15 + 122 + 103 + 14 = \mathbf{254}$.
2. $\frac{27}{36} \times 100 = \mathbf{75\%}$. A guesser expects $36 \div 4 = 9$, so the participant got $27 - 9 = \mathbf{18}$ more right.
3. With two opposite words, a participant only had to decide whether the eyes looked positive or negative, which is easy and gives a 1 in 2 chance of guessing right. With four words of similar tone, they must pick out the exact mental state, so guessing is harder (1 in 4) and the score shows real theory of mind.
4. Mean = $\frac{24 + 30 + 27 + 21 + 28}{5} = \frac{130}{5} = \mathbf{26}$. Range = $30 - 21 = \mathbf{9}$.
5. Females are, on average, better than males at working out mental states from the eyes, as females in the general population and student groups scored higher on the eyes test than males.

Past questions to try

- 9990/12/O/N/23 Q1: the measure negatively correlated with the eyes test, then the Group 4 sample.
- 9990/22/M/J/22 Q4: why two problems of the original test hurt validity, and how a fix helped.
- 9990/13/M/J/23 Q7: two ethical real-world applications.
- 9990/11/M/J/24 Q7: two problems if children were used as participants.