Designing a reliable

Theory Test

for the

International Biology Olympiad

Hans Morélis
revised version
october 2011

The IBO Theory Test

Introduction

The IBO Guide contains several directives about the IBO tasks. Most important are:

1) Assessment experts should be involved in the design and wording of the test questions, the structure of the tasks as a whole and the marking and ranking procedure (IBO Guide, page 24).

2) All questions should focus on reasoning, problem solving and understanding. Questions dealing with just knowledge should be expelled. (IBO Guide, page 26).

3) The design of the questions should enable objectively marking and scoring (IBO Guide, page 24), e.g.

   - Multiple choice (MC)
   - Matching
   - Sequencing
   - Filling out (only numbers, numeric, or alphabetical codes)
   - Judging a set of statements, whether each is right (true) or wrong (false)

   We call these questions *closed* or *pre-coded* as there is only one correct answer which should be represented as a number, letter, or code.

4) In order to improve the reliability of IBO tests it is advised to reduce the proportion of MC questions in IBO tests and increase the amount of the other pre-coded type of questions (IBO Guide, page 26).

5) The total number of IBO theoretical questions should not exceed 100 (IBO Guide, page 26).

6) In order to facilitate the translation process of the tests the host country will do the best to avoid unnecessary words in the tasks (IBO Guide, page 24). Test questions should be as concise as possible (IBO Guide, page 27).

These directives are valuable and based upon years of experience. Following these directives will improve the quality and reliability of the tasks and speed up Jury discussion. But unfortunately the directives are often ignored. Maybe IBO hosts are not aware how important it is to follow the directives, so we offer some background information.

1 Assessment experts

Producing good quality IBO questions seems simple for experts in biology. But the truth is that it requires special knowledge and skills in assessment, which most biology experts not have. Of course local biologists of IBO host countries being responsible for the IBO test will have a pretty good idea about the scientific background of the IBO questions, but this doesn't mean that these experts also are able to produce good and reliable tasks as this really is something totally different.
Unfortunately most biology experts are not aware of this and that's why they normally are not in favor of consulting assessment experts. They just think it is not necessary. But the result is that rather some IBO's contained poorly designed questions, resulting in skipping many questions or long lasting discussions of the International Jury.

IBO 2011 was no exception. In part A of the Theory test 22 out of 68 questions were not designed in line with general directives concerning test production. See **X** in the table.

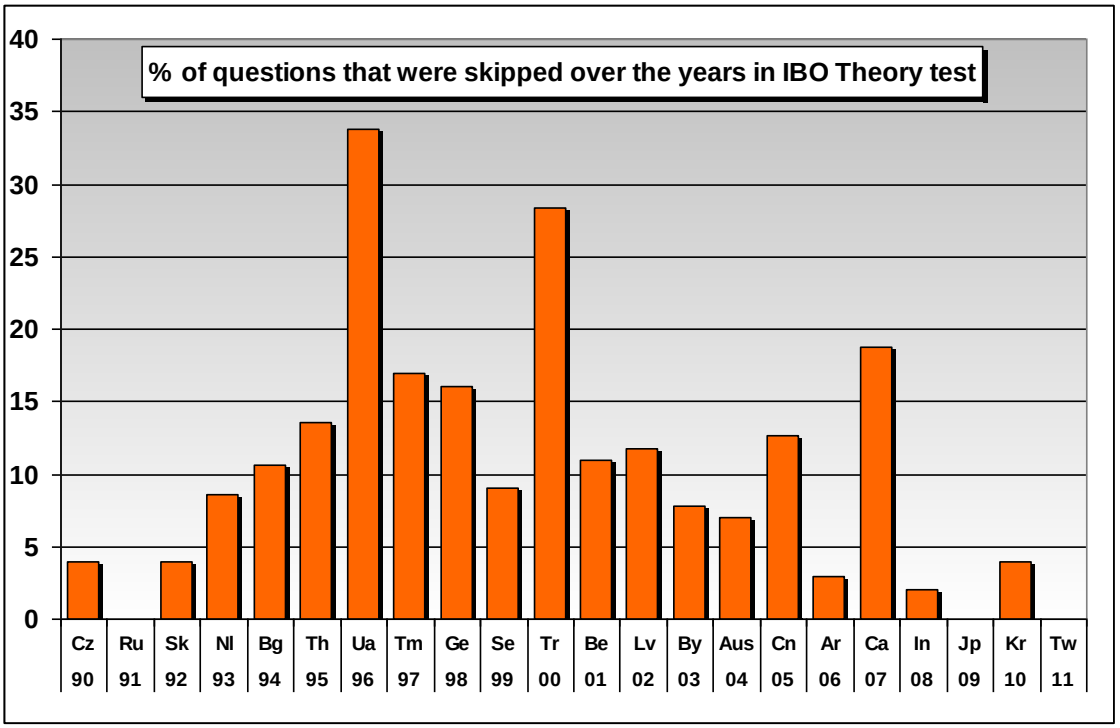
IBO 2011, part A										
Nr	1	2	3	4	5	6	7	8	9	0
1-10		**X**			**X**					
11-20			**X**				**X**		**X**	
21-30	**X**	**X**					**X**	**X**	**X**	
31-40	**X**	**X**	**X**		**X**					
41-50		**X**	**X**			**X**				**X**
51-60	**X**					**X**	**X**	**X**		

X = design not in line with assessment directives

It resulted in discussions, which for a good deal were focusing upon restyling questions and improving wording and design. This is a strong argument to involve assessment experts in the production of the tasks. Present them the raw version of the tasks and ask advice about the design of the test questions and about possible improvements of these questions and of the marking and scoring procedures. An analysis of IBO tasks over the years proved that really with not much effort the design of questions and so the quality of the test could be better. This will help reducing time spent by the international Jury in discussing questions.

It is good to realize that real assessment experts normally are not in the Biology Department. It is better to recruit these experts in the Department of Education. That's the place the know-how should be. Also Social Sciences especially Psychology is the place to be, as designing reliable questionnaires and tests belong to their key activities. That these people probably are no biologists is not a problem.
For real assessment experts it is quite easy to judge and improve pre-coded questions. They also are able to transform MC questions into the more reliable questions based upon filling out, matching, sequencing, judging statements true or false. We will offer in the next pages many examples of these transformations.

One aspect about all this is promising. The introduction in 2009 of an international subcommittee (ISC), screening the tasks on forehand, works fine. Thanks to this committee the portion of questions, skipped in IBO, has declined drastically. See diagram. On the other hand it is clear that the ISC has only a few days for their job so their influence on upgrading questions is limited. A first screening by assessment experts still is vital.



2 Knowledge

The IBO is a competition for bright students, so it is obvious that testing only knowledge is not appropriate. Still in many years rather some questions were focusing upon only recalling factual knowledge directly taken from some general Biology textbook like Campbell. The following tables show for which questions this applied in IBO 2011. [1]

IBO 2011, part A										
Nr	1	2	3	4	5	6	7	8	9	0
1-10	**K**	**K**			**K**		**K**	**K**		**K**
11-20	**K**			**K**		**K**		**K**		
21-30	**K**		**K**	**K**						**K**
31-40										
41-50										**K**
51-60										

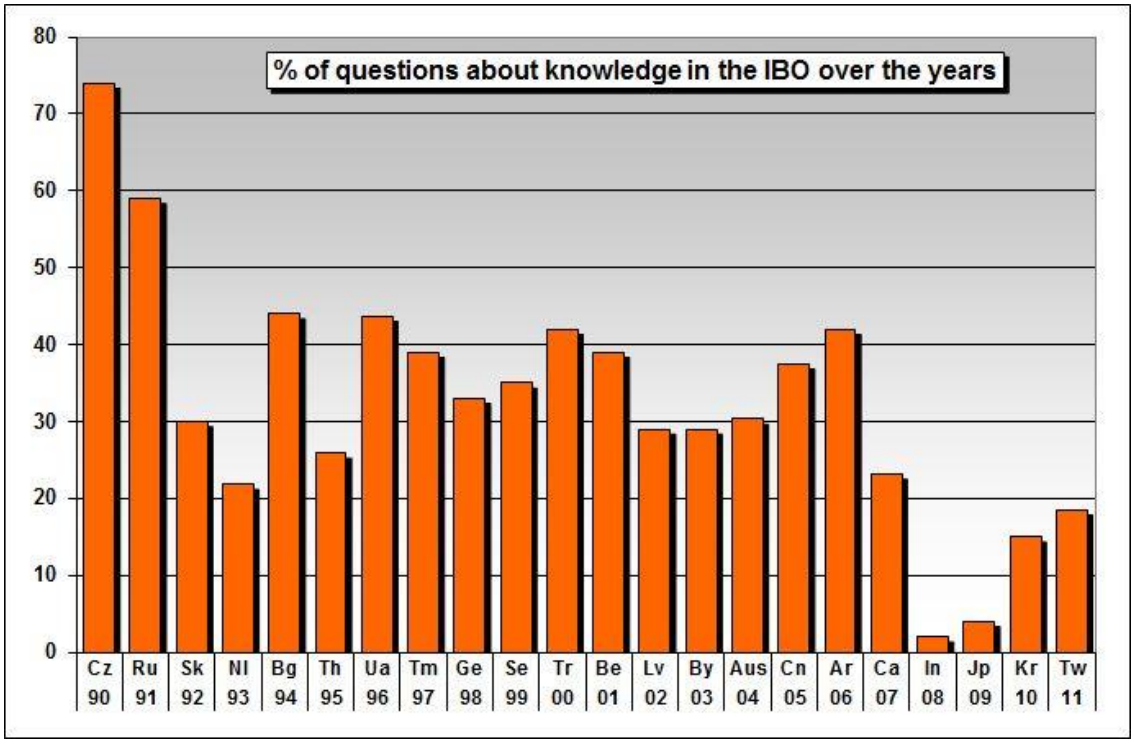
K = only factual knowledge, 16 out of 58 questions

[1] The tables and the diagram on the next page are based upon my personal interpretation of criteria defined by the National Institute for Assessment of the Netherlands. See also the well-known taxonomy of Bloom, which describes 6 levels of cognitive thinking.

IBO 2011, part B										
Nr	1	2	3	4	5	6	7	8	9	0
1-10										**K**
11-20									**K**	
21-30							**K**			
31-40										
41-50			**K**							

K = only factual knowledge, 4 out of 50 questions

IBO 2011 was no exception. See diagram.



IBO 2008 and 2009 were really good. But regrettably it is obvious that the directive that questions focusing upon only knowledge should be expelled was ignored in many IBO 's.

Remark:
Some colleagues argue that it is desirable to check the knowledge about important facts. These colleagues are representing a minority and it is not in line with the principle that our competition should enable the students to exhibit their ability in creative thinking and reasoning. Assessing important facts of course is fine, but it should always happen in a reasoning context and never just as factual knowledge.

3 Precoded questions

In order to facilitate objective marking and speed up the marking process only precoded (closed) questions are used in the IBO. Examples are multiple choice questions (MCQ) and questions based upon matching, sequencing, filling out a number or code or judging statements true/false.

Rather some people think that it is hard to test reasoning and understanding with pre-coded questions, but this is a misconception. Pre-coded questions are an effective, efficient and reliable assessment tool, but difficult to construct and not ideal for testing creative thinking.

4 Reduce MCQ

From assessment point of view IBO 1998 was very special. The German organizers fulfilled an excellent study on the composition of the theoretical task in relation to the results of students and reliability of the test. It led to the conclusion that questions based upon matching, sequencing, filling out the blank or judging statements are better (more reliable) as assessment tool than MC questions. See Report of the 9th IBO in Kiel, Chapter 3 Evaluation Report. Based upon this study it is advisable to decrease the proportion of MCQ in IBO tests and increase the amount of other types of pre-coded questions. This recommendation now is in the IBO Guide.

Regrettably since 2008 the IBO has been confronted with another argument to avoid MC questions. Modern communication facilities make it very simple to pass on the answers of MC questions. This indeed happened in 2008, 2009 and 2010. It is clear this frustrating situation should end and the solution is easy.
avoid simple answer codes, so avoid MC questions

This seems difficult to realize, but we will see that with help of assessment experts and some hints it is rather easy to transform MC questions into true/False questions.

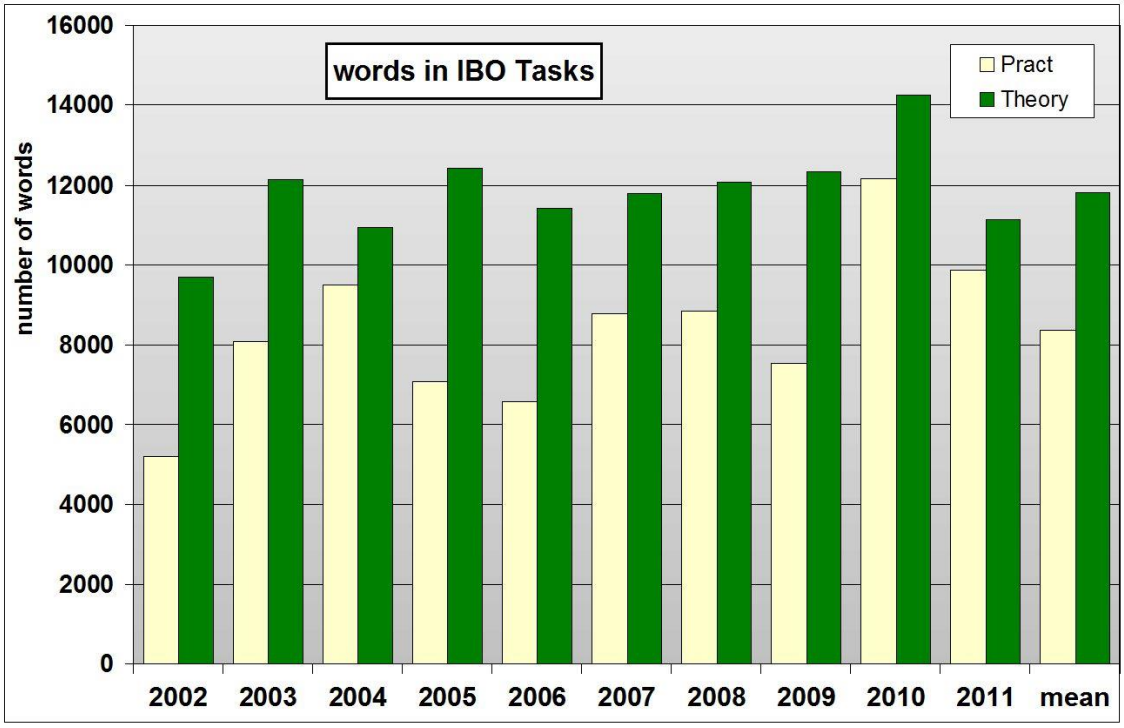
5 100 questions

Until 1998 it was quite normal to have over 120 questions, leading to endless translation sessions. Thanks to the IBO 1998 evaluation by the German host it became clear that 100 questions is fine on the condition that the proportion of MCQ is reduced as much as possible and pure knowledge questions are expelled. In doing so and with help of assessment experts it still is possible to produce a nice balanced and reliable test, covering all important biology topics.

6 Concise design

Students should be tested on their skill in reasoning about biological notions and solving biological problems. The IBO is not a language competition. This means that problems should be presented concise without abundant words and problems should not be hided behind long and complex wording. Concise wording also favors the translation process. IBO 2010 was a little bit problematic. See diagram. The test were fine, but the number of words was a lot, so translations took a long time.

Often pictures, schemes, or tables can substitute a lot of words. Real assessment experts are able to apply this and present problems clearly and briefly.
By the way, in International Chemistry Olympiad the number of characters of the test is restricted to 25.000.



Summarized

Essential for IBO Theory Test quality are the following 3 measures:

1) All questions featuring in the IBO should be designed in cooperation with assessment experts
2) Expel all questions focusing only upon factual knowledge.
3) Use primarily questions based upon matching, sequencing, filling out or judging a set of statements. Forget about MCQ.

It seems difficult to refrain from MC questions.
In the following pages we will show, with lots of examples, that it is no problem to transform MC questions in other types of pre-coded questions. All the examples are taken from IBO Theory Test 2011.

But first some general assessment directives.

Assessment directives

In order to support the process of designing good questions for IBO Theory Test some directives are offered. These are based upon principles general agreed by assessment experts and proven to be valid by evaluation of the reliability of tests. Most relevant for IBO are the following.

- Mind lay-out: not a mess, avoid page break in one question
- Indicate how many points are assigned to each question
- Mind the difficulty of the test questions.
  Very difficult and very simple questions are not good discriminating between students. Go for a 50 to 70% correct score of students. This can be checked by pretesting with an appropriate pilot group
- Mind wording: formulate as short as possible.
  The problem should be clear, concise and simple stated,
  not hided behind complex sentences, extensive wording or special jargon
- One aspect (central idea) per question
- Avoid: (double) negations and *all of the above* or *none of the above*
- Careful with extreme or vague words like: never, always, only, all, none, could, might, generally, some(times), few, often, mostly
- No overlap between questions
- No questions requiring the correct answer of the preceding question
- Go for reasoning. Expel questions focusing only upon factual knowledge
  Hint: It is easy to create reasoning questions through offering graphs, diagrams, figures, tables, or a description of a case study or situation, followed by statements to be judged true or false.

References regarding assessment directives.

- *A Guide to Teaching & Learning Practices,* IV. Student assessment, Chapter 12 Testing and Assessment issues, Florida State University, http://learningforlife.fsu.edu/ctl/explore/onlineresources/docs/Chptr12.pdf:
- *Best Practices for Designing and grading exams,* University of Michigan, CRLT Occasional papers nr 24, http://www.crlt.umich.edu/publinks/CRLT_no24.pdf
- Lucy C. Jacobs, Ph.D. (2004), *How to Write Better Tests*: A Handbook for Improving Test Construction Skills, Instructional Support Service, Indiana University Bloomington, http://www.indiana.edu/~best/write_better_tests.shtml
- Dawn M. Zimmaro, Ph.D. (2010) Writing good Multiple-Choice exams (revised version), Center for Teaching and Learning, University of Texas at Austin http://ctl.utexas.edu/assets/Evaluation--Assessment/Writing-Good-Multiple-Choice-Exams-04-28-10.pdf
- More assessment info on: http://depts.washington.edu/cidrweb/resources/exams.html

We illustrate some of the assessment directives with examples extracted from IBO-2011. Below a question which extensive wording, which could be designed more concise.

Question A 35, Original version

A35. There are three types of chemical substances that organisms emit to mediate interspecific interactions: kairomone, allomone, and synomone. Kairomone benefits individuals of another species which receives it but is disadvantageous to the emitter. Allomone benefits the emitter, and does not benefit or harm the receiver. Synomone benefits both the emitter and receiver. A plant species emits a volatile essential oil that attracts a phytophagous beetle to feed and lay eggs on its leaves. At the same time, it also attracts a parasitoid wasp, and helps this parasitic natural enemy of the beetles to locate the beetle larvae within which they can lay their own eggs. Which of the following descriptions of the role that this essential oil plays is correct?

(A) It acts as a synomone between the plant and the beetle, and an allomone between the plant and the parasitoid wasp.
(B) It acts as a kairomone between the plant and the parasitoid wasp, and a synomone between the beetle and the parasitoid wasp.
(C) It acts as a kairomone between the plant and the beetle, and a synomone between the plant and the parasitoid wasp.
(D) It acts as a kairomone between the plant and the beetle, and an allomone between the beetle and the parasitoid wasp.
(E) It acts as a kairomone between the plant and the parasitoid wasp, as well as between the beetle and the parasitoid wasp.

Improved version

A35. There are three types of chemical substances that organisms emit to mediate interspecific interactions: kairomone, allomone, and synomone. The roles of these chemicals can be classified as the table given below.

	Emitter	Receiver
Kairomone	Disadvantage	Benefit
Allomone	Benefit	No benefit/harm
Synomone	Benefit	Benefit

A plant species emits a volatile essential oil that attracts a phytophagous beetle to feed and lay eggs on its leaves. At the same time, it also attracts a parasitoid wasp, and helps this parasitic natural enemy of the beetles to locate the beetle larvae within which they can lay their own eggs. Which of the following descriptions regarding the role of this oil is correct?

	Plant - Beetle	Plant – Wasp	Beetle - Wasp
A	Synomone	Allomone	
B		Kairomone	Synomone
C	Kairomone	Synomone	
D	Kairomone		Allomone
E		Kairomone	Kairomone

The following question A 43 had two defects:

- more than one aspect tested in one question and
- asking what is NOT correct.

A 43 Dr. Yeh treated the seeds of the above-mentioned homozygous blue-flower plants with chemical mutagen to produce a mutant population. Three recessive mutants, wf1, wf2, and wf3, produced white flowers were selected. He crossed the mutants and obtained the following results: wf1 x wf3 produced F2 offspring with only white flowers, and wf2 x wf3 produced F2 offspring with blue and white flowers in a ratio of 9:7. According to these data, which of the statements below is **NOT correct**?

A wf1 and wf3 are unable to complement each other.
B wf2 and wf3 are able to complement each other.
C wf1 and wf3 are in the same locus.
D wf2 and wf3 are not in the same locus.
E The F1 offspring from crossing wf1 and wf2 will all produce white flowers.

Three aspects are tested: complement, locus and white flowers.
It is better (more reliable) to ask these aspects in separate questions.
Example:
Which are able to complement each other. Put a cross.

	Yes	No
wf1 and wf2		
wf1 and wf3		
Wf2 and wf3		

Etc.

Normally “NOT” questions have a lower reliability than positive formulated questions. In A43 the combination [Not correct – unable] in alternative A and [Not correct – not in the same locus] in alternative D is extra tricky, because both include a double negation.
In fact is it much better to transform *NOT correct* questions into a set of true/false statements to be judged. In the following pages we will show some examples.

Transforming MC questions into other pre-coded questions.

Unfortunately in IBO we experienced that cheating is easy with MC questions.
So avoiding MC questions in IBO tests would be nice. This also completely is in line with our IBO Guide (see page 26) and recommendations from IBO 1998 organizers.
In using MC questions from IBO 2011 we will show in the next pages that transforming MC questions into other type of pre-coded questions is rather easy. The most simple approach is the following.

MC questions have alternatives A, B, C, etc. It is not difficult to get rid of these codes.
Use alternatives without numbers, letters or codes.
Students should indicate the correctness of each alternative one by one on the answer sheet. That’s all.

General example:

Put a cross in the appropriate boxes

	True (correct)	False (wrong)
1st Alternative		
2nd Alternative		
3rd Alternative		
4th Alternative		

Using this principle all MC questions pretty easily can be transformed into a set of True/False questions. See examples in the next pages.

True/False

IBO 2011, question A3, Original design.

A3. In some cells, synthesis of isoleucine from threonine is catalyzed by the sequential action of five enzymes **a, b, c, d** and **e**, which produce 4 intermediates A, B, C and D, and the end product isoleucine, respectively. What is most likely to happen when isoleucine is overproduced and there is an ample supply of threonine in cells?

A Isoleucine associates with threonine to inhibit the activity of enzyme a.
B Isoleucine associates with intermediate D to inhibit the activity of enzyme e.
C Isoleucine binds to enzyme a and inhibits its activity.
D Isoleucine binds to enzyme e and inhibits its activity.
E Threonine is converted into isoleucine continuously through the 5 enzymes.

Transformed version.

A3. In some cells, synthesis of isoleucine from threonine is catalyzed by the sequential action of five enzymes **a, b, c, d** and **e**, which produce 4 intermediates A, B, C and D, and the end product isoleucine, respectively. What is most likely to happen when isoleucine is overproduced and there is an ample supply of threonine in cells?
Indicate in the table what is true and what is false.

	True	False
Isoleucine associates with threonine to inhibit the activity of enzyme a.		
Isoleucine associates with intermediate D to inhibit the activity of enzyme e.		
Isoleucine binds to enzyme a and inhibits its activity.		
Isoleucine binds to enzyme e and inhibits its activity.		
Threonine is converted into isoleucine continuously through the 5 enzymes.		

Now it's no MC question anymore ⇨ Less chance on cheating and better reliability.

Using this strategy we also get rid of unwanted negation questions. Four of these were featuring in IBO-2011. See the following one.

IBO 2011, question A4, Original design.

A4. In some prokaryotic organisms, SO_4^{2-} is used as the final electron receptor at the end of electron transport chain during cellular respiration. Which of the following statements regarding cellular respiration in these prokaryotic organisms is **NOT correct**?

A It is anaerobic respiration.
B The reception of electron by SO_4^{2-} is accompanied by the production of H_2O.
C Operation of the electron transport chain builds up a proton motive force.
D ATP is produced.
E Production of ATP is correlated with the mobility of H^+.

Using **NOT correct** reduces the reliability of this question. This can easily avoided in applying the following transformation.

Better design

A4. In some prokaryotic organisms, SO_4^{2-} is used as the final electron receptor at the end of electron transport chain during cellular respiration. Which of the following statements regarding cellular respiration in these prokaryotic organisms is true/false? Put a tick.

	True	False
It is anaerobic respiration.		
The reception of electron by SO_4^{2-} is accompanied by the production of H_2O.		
Operation of the electron transport chain builds up a proton motive force.		
ATP is produced.		
Production of ATP is correlated with the mobility of H^+.		

We now get rid of *NOT correct* and it's no longer a MC question.
⇨ Less chance on cheating and better reliability.

See also IBO 2011, question A22

A22. *Agrobacterium tumefaciens*-mediated transformation, a widely used method to transfer foreign genes into the plant genome, has contributed to the considerable successes that plant biotechnology has already achieved. For instance, a gene encoding the coat protein (CP) of papaya ringspot virus (PRSV) was used to generate the virus-resistant transgenic SunUp papaya in Hawaii. The construct used for transformation includes the *CP* gene and a selectable marker gene (*nptII*) conferring kanamycin resistance. Both *CP* and *nptII* genes are driven by a constitutive cauliflower mosaic virus (CaMV) 35S promoter. According to the above information, which of the following statements is **NOT correct**?

A The SunUp papaya is resistant to kanamycin.
B The SunUp papaya contains some DNA sequences from CaMV.
C The SunUp papaya contains some chromosomal DNA of *Agrobacterium tumefaciens.*
D The SunUp papaya contains a portion of the Ti plasmid termed T-DNA.
E The SunUp papaya contains the *nptII* gene.

All alternatives start with *The SunUp papaya.* This can be simplified.

Improved version.

A22. *Agrobacterium tumefaciens*-mediated transformation, a widely used method to transfer foreign genes into the plant genome, has contributed to the considerable successes that plant biotechnology has already achieved. For instance, a gene encoding the coat protein (CP) of papaya ringspot virus (PRSV) was used to generate the virus-resistant transgenic SunUp papaya in Hawaii. The construct used for transformation includes the *CP* gene and a selectable marker gene (*nptII*) conferring kanamycin resistance. Both *CP* and *nptII* genes are driven by a constitutive cauliflower mosaic virus (CaMV) 35S promoter. According to the above information, which of the following statements about the SunUp Papaya is true/False?
Put a tick.

	True	False
is resistant to kanamycin.		
contains some DNA sequences from CaMV.		
contains some chromosomal DNA of *Agrobacterium tumefaciens.*		
contains a portion of the Ti plasmid termed T-DNA.		
contains the *nptII* gene.		

Remark.
With this True/False approach alternatives have to be in the question and on the answer sheet. This is unpleasant. It easily can be avoided. Use student names and each student claims one of the alternatives as statement, which each has to be judged true or false.
Example: Five students conclude that Sunup Papaya:

Alex	is resistant to kanamycin.
Andrew	contains some DNA sequences from CaMV.
Ann	contains some chromosomal DNA of *Agrobacterium tumefaciens.*
Antonio	contains a portion of the Ti plasmid termed T-DNA.
Archie	contains the *nptII* gene.

Tick off their conclusions with a tick on the answer sheet

	True	False
Alex		
Andrew		
Ann		
Antonio		
Archie		

Or designed horizontal

(from left to right)

	Alex	Andrew	Ann	Antonio	Archie
True					
False					

+ / —

Sometimes it is more appropriate to switch to + (yes) and – (no) instead of True / False.

See IBO-2011 question A6. Original version.

A6. Which structural or physiological feature of bacteria is commonly used as a drug target to kill bacteria effectively but with very little harm to human cells?

A Glycolysis
B Components of plasma membrane
C Components of ribosome
D Components of the electron transport chain in aerobic respiration
E Requirement of oxygen

Improved, version.

A6. Which structural or physiological feature of bacteria is commonly used as a drug target to kill bacteria effectively but with very little harm to human cells?

I Glycolysis
II Components of plasma membrane
III Components of ribosome
IV Components of the electron transport chain in aerobic respiration
V Requirement of oxygen

Tick of **+** (Yes) or **—** (No) on the answer sheet.

	+	**—**
I		
II		
III		
IV		
V		

or

	I	II	III	IV	V
+					
—					

Another example: IBO-2011 question13. Original version.

A13. Hypersensitive response is one of the plant defense responses to pathogens. Each of four pathogen strains, a to d, produce a distinct range of effectors. One of the effectors, Avr, recognized by a specific receptor protein encoded by the

resistance (R) gene in the host plant is present in strains b and c. Host plants B and D produce the R protein. Which plant(s) are likely to develop a hypersensitive response after the host plants A to D are infected by pathogens a to d (a→A, b→B, c→C, d→D), respectively?

A A only
B B only
C C only
D D only
E B and C
F B and D

There are 6 alternatives A up to F. The letters A up to D and a up to d also are used in the stem of the question. That's rather confusing. Students have to draw a conclusion about a→A, b→B, c→C, d→D. This means 4 different aspects, which each have to be judged true or false. Totally 2^4 = 16 different answers are possible. But there are 6 alternatives so only 6 out of 16 choices are offered. That's not good (less reliable). Furthermore B is featuring three times in the alternatives. It is obvious B should be correct, which means that alternative A, C and D are not likely to be correct. So the real choice in this question is between B, E and F.

It is easy to improve the question in the following way.

A13. Hypersensitive response is one of the plant defense responses to pathogens. Each of four pathogen strains, a to d, produce a distinct range of effectors. One of the effectors, Avr, recognized by a specific receptor protein encoded by the resistance (R) gene in the host plant is present in strains b and c. Host plants B and D produce the R protein. Which plant(s) are likely to develop a hypersensitive response after the host plants A to D are infected by pathogens a to d (a→A, b→B, c→C, d→D), respectively?

Put a tick.

	Yes	No
A		
B		
C		
D		

Or

	A	B	C	D
Yes				
No				

We have seen now that question A13 is an example of a multiple question with a limited number of alternatives, while the total number of choices is much more. This happened in IBO-2011 in many other questions. We show two examples.

Question 29, original version.

A29. Compared to a healthy individual what are the levels of the following hormones in an individual with primary hyperthyroidism (hypersecretion of thyroid hormone)?
Thyrotropin-releasing hormone (TRH), thyroid-stimulating hormone (TSH), Thyroid hormones T3 and T4

↑: increase ↓: decrease —: remains unchanged

	TRH	TSH	T3	T4
A	↑	↑	—	↑
B	↑	↑	↑	—
C	↓	↓	↑	↑
D	↓	↓	↓	↓
E	↓	↑	↑	↑

In this question we see only five alternatives (A up to E), but in fact the total number of possible choices is $3^4 = 81$! Besides this the use of the symbols ↑, ↓ and — is a little bit unusual and confusing. It would be nicer to have the following fill out design.

Question 29, improved: less chance on cheating and more reliability:

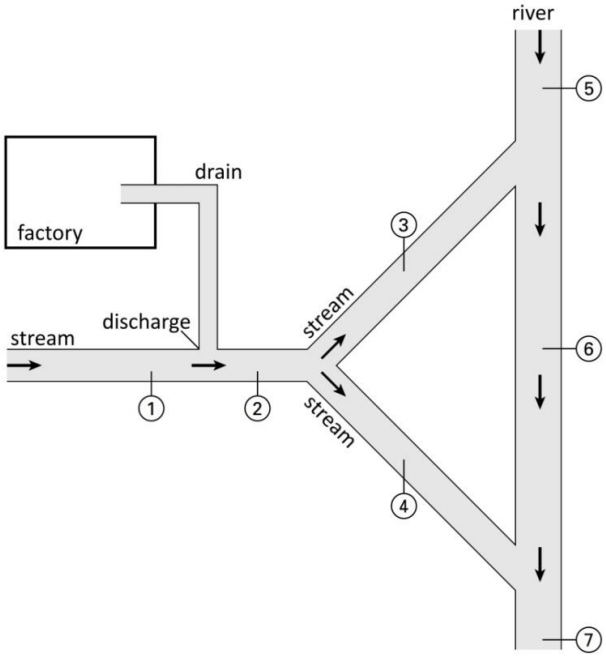
A29. Compared to a healthy individual what are the levels of the following hormones in an individual with primary hyperthyroidism (hypersecretion of thyroid hormone)?
Thyrotropin-releasing hormone (TRH), thyroid-stimulating hormone (TSH), Thyroid hormones T3 and T4
Tick off in the table

	increase	decrease	remains unchanged
TRH			
TSH			
T3			
T4			

Question A46 was

A46. A group of students would like to know how the discharge of waste water from a factory might influence water quality of a river. The picture shows 7 potential sampling locations (① to ⑦) in relation to the locations of the factory and the river. Which locations are essential to be included in the sampling in order to draw valid conclusions about the pollution of the river by the factory?

A Locations 1, 2, 4, 7
B Locations 1, 3, 4, 7
C Locations 1, 2, 5, 7
D Locations 2, 3, 4, 6
E Locations 2, 5, 6, 7



In this MC question the number of possible choices is: 7!/(3!*4!) = 35
Only 5 of these choices are presented.
It would have been much better to substitute the 5 alternatives A up to E by:

Encircle the correct location numbers:

1	2	3	4	5	6	7

Judging statements

In IBO questions often a set of statements is presented which should be judged by the students on their correctness. In fact quite good, but only if all statements are judged separately in a true/false setting, not as a MC question. See the next examples.

IBO 2011, question 28, original version

A28. Which of the following events will result in an excitatory postsynaptic potential?
- a. Increasing sodium influx.
- b. Blocking potassium out-flux.
- c. Increasing calcium influx.
- d. Closing a chloride channel.

A Only a & b
B Only b & c
C Only a, c & d
D Only b, c & d
E a, b, c & d.

Using a up to d in the stem of a MC question and refer to these letters in the alternatives using capitals A up to E is not in line with assessment directives. But the main problem of this MC question is that 4 statements have to be judged, leading to 2^4 = 16 possible choices. But only 5 (out of 16) choices are presented. Of course it would have been better (more reliable) to design this question as a true/false or correct/wrong type in the following way.

IBO 2011, question 28, improved version

A28. Which of the following events will result in an excitatory postsynaptic potential?
- I Increasing sodium influx.
- II Blocking potassium out-flux.
- II Increasing calcium influx.
- IV Closing a chloride channel.

Indicate with a tick what is correct and what is wrong

	Correct	Wrong
I		
II		
III		
IV		

IBO-2011 question A30, original version

A30. Which of following receptors/molecules are required for the activation of Helper T cells trigged by antigen-presenting cells.

1. CD8
2. CD4
3. Class I MHC molecule
4. Class II MHC molecule
5. T cell receptor

A Only 1, 3 & 5
B Only 2, 4 & 5
C Only 3, 4 & 5
D Only 2 & 4
E Only 1 & 3

Five aspects have to be judged ⇨ total number of choices is $2^5 = 32$.
Only 5 (out of 32) alternatives is not good. It would have been better to transform this MC question in the following way.

IBO-2011 question A30, improved version

A30. Which of following receptors/molecules are required for the activation of Helper T cells trigged by antigen-presenting cells.

1. CD8
2. CD4
3. Class I MHC molecule
4. Class II MHC molecule
5. T cell receptor

Put a tick.

	required	not required
CD8		
CD4		
Class I MHC molecule		
Class II MHC molecule		
T cell receptor		

The following question (IBO-2011 question A17) is even worse:
6 statements have to be judged ⇨ $2^6 = 64$ choices.

Original version

A17. Dennis dissected a plant leaf and found bundle sheath cells full of starch granules. Which of the following characteristics can be observed in this plant?

I. Stomata open at night
II. Presence of PEP carboxylase in mesophylls
III. Presence of Rubisco in bundle sheath cells
IV. High photorespiration rate on hot summer days
V. Carbon fixation can occur in both mesophyll and bundle sheath cells
VI. Carbon assimilation rate is saturated in the early morning on summer days

A Only I, IV
B Only II, IV, V
C Only II, IV, VI
D Only II, III, V
E Only II, III, V, VI
F Only II, IV, V, VI

Looking to the 6 (out of 64) alternatives it is obvious that I must be wrong (featuring once) and II (featuring 5x) must be correct.
Of course the following design would have been much better.

Improved version

A17. Dennis dissected a plant leaf and found bundle sheath cells full of starch granules. Which of the following characteristics can be observed in this plant?

I. Stomata open at night
II. Presence of PEP carboxylase in mesophylls
III. Presence of Rubisco in bundle sheath cells
IV. High photorespiration rate on hot summer days
V. Carbon fixation can occur in both mesophyll and bundle sheath cells
VI. Carbon assimilation rate is saturated in the early morning on summer days

Put a tick

	Yes	No
I		
II		
III		
IV		
V		
VI		

or

	I	II	III	IV	V	VI
Yes						
No						

How to mark true/false type questions.

Be aware that in true/false questions it is not good to only ask what is true, as this will offer marking problems.
The number of decisions is important, not the number of correct statements.
3 statements or choices ⇒ 3 decisions
4 statements or choices ⇒ 4 decisions, etc
Every decision (true or false) has the same value, and if we give all of them 1 point, the scores of the IBO-students can be 0,1, 2, 3, or 4. This depends on the number of correct choices. An all (4 pnts) or nothing (0 pnts) score would be unfair.

A major weakness of True/false questions is guessing. Students have a 50% chance of correctly answering an item without any knowledge of the content. This is a disadvantage of this system. Luckily for IBO this is not really a problem as it is a competition, so what counts is the ranking.

A not very popular measure could be subtracting points for each incorrect answer. This was applied in some questions in IBO-2011.
An often used method is subtracting p/(n-1) points for each incorrect choice, where
p = total points for the question
n = number of choices.
It means that only one correct choice always results in zero points.
See also *GradingProposal.pdf* by Eva Deinum of the Netherlands.

Matching

Sometimes a question is dealing with matching aspects. In that case a MC question of course is not a good approach. It is much better, reliable and also easy to design a matching question. See examples.

IBO-2011,Question A51, original design

A51. A large proportion of angiosperms are pollinated by animals. Assign the following flower descriptions (I to V) to the most likely pollinator (a to e).

I. Flower white, open during night, intensive fragrant, nectar hidden in long, tight tubes.
II. Flower often with ultraviolet coloring pattern, open during daytime, pleasant fragrant.
III. Flower large and coarse, bright red, open during daytime, no fragrance but large amounts of nectar
IV. Flower large and coarse, far opened, open during night, intensive fragrant, large amounts of nectar
V. Flower reddish brown, no nectar, smell of rotten flesh

a. bats
b. birds
c. bees
d. flies
e. moths

Which of the following statement is correct?
A I-a, II-b, III-c, IV-e, V-d
B I-b, II-c, III-d, IV-a, V-e
C I-d, II-e, III-a, IV-b, V-c
D I-e, II-c, III-b, IV-a, V-d
E I-e, II-d, III-c, IV-b, V-a

In this question the total number of possible answers is 5! = 1*2*3*4*5 = 120 and we see only 5 of them. So this question is not well designed. The reliability can be improved in the following way:

IBO-2011,Question A51, improved design

A51. A large proportion of angiosperms are pollinated by animals. Assign the following flower descriptions (I to V) to the most likely pollinator (a to e).

I	Flower white, open during night, intensive fragrant, nectar hidden in long, tight tubes.	a	bats
II	Flower often with ultraviolet coloring pattern, open during daytime, pleasant fragrant.	b	birds
III	Flower large and coarse, bright red, open during daytime, no fragrance but large amounts of nectar	c	bees
IV	Flower large and coarse, far opened, open during night, intensive fragrant, large amounts of nectar	d	flies
V	Flower reddish brown, no nectar, smell of rotten flesh	e	moths

Put a tick in the correct boxes.

	a	b	c	d	E
I					
II					
III					
IV					
V					

Another possibility:
Which flower description fits to which pollinator. Fill out the correct letter.

Flower	I	II	III	IV	V
Pollinator					

Sequencing

Often nice questions can be created about sequences in biological events. It is useless to try to test this in a MC question, because with n events the possible number of choices again is n!.

IBO-2011, question A21, original version

A21. During leaf development in water lily, the sclereid-initials grow and elongate along the palisade mesophyll cells or the intercellular space between them. After elongation they gradually form calcium oxalate crystals in the cell wall along the cell membrane. Thereafter, they form the secondary cell wall. Four cell wall structures are: (I) primary cell wall; (II) secondary cell wall; (III) middle lamella; (IV) calcium oxalate crystals. What is the final sequence of structures in the mature sclereids of water lily, starting from the plasma membrane as the innermost layer to the outermost layer?

A I → IV → II → III
B III → IV → I → II
C I → II→ IV→ III
D III → I → IV → II
E II → IV → I → III

only 5 out of 24 possible sequences are presented
that's not good

By the way: just by inspecting the 5 alternatives it is obvious that IV must be 2nd and III must be last. This means that without any biological notion alternative B, C and D can be skipped.

IBO-2011, question A21, improved design

A21. During leaf development in water lily, the sclereid-initials grow and elongate along the palisade mesophyll cells or the intercellular space between them. After elongation they gradually form calcium oxalate crystals in the cell wall along the cell membrane. Thereafter, they form the secondary cell wall. Four cell wall structures are:
(I) primary cell wall;
(II) secondary cell wall;
(III) middle lamella;
(IV) calcium oxalate crystals.
What is the sequence of structures in the mature sclereids of water lily, starting from the plasma membrane as the innermost layer to the outermost layer?

Fill out I, II, III and IV in the correct sequence:

1st	2nd	3rd	4th

Another example.

IBO-2011, question 58, original design

A58. A scientist unearthed four plant fossils (I to IV) with some prominent structures intact. These are listed in the following table:

Structure / Fossil	Spore	Ovary	Embryo	Pollen	Xylem	Ovule
I			✓		✓	
II			✓	✓	✓	✓
III		✓		✓	✓	✓
IV	✓		✓			

According to this table, which sequence below correctly represents the order of evolution of these plants?

A I→II→III→IV
B II→III→IV→I
C III→IV→I→II
D IV→I→II→III
E II→I→IV→III
F IV→I→ III→II

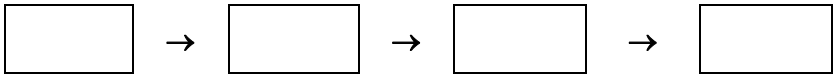
only 6 out of 24 possible sequences are presented
that's not good

IBO-2011, question 58, improved more reliable design

A58. A scientist unearthed four plant fossils (I to IV) with some prominent structures intact. These are listed in the following table:

Structure / Fossil	Spore	Ovary	Embryo	Pollen	Xylem	Ovule
I			✓		✓	
II			✓	✓	✓	✓
III		✓		✓	✓	✓
IV	✓		✓			

According to this table, fill out the correct sequence in evolution of I, II, III and IV



Numerical answers

Questions requiring a numerical answer or % of course never should be shaped as a MC question. In this case a fill-in-the blank type always is better. We illustrate this with two examples.

IBO-2011, question A27, original design

A27. A method to estimate a mammal's blood volume uses a specific radioactive isotope of iodine (^{123}I). This isotope, usually produced synthetically, has a half-life time of 13 hours. It decays to ^{123}Te, which is almost perfectly stable. To estimate the blood volume, 10 mL of iodine solution are injected into the animal's vein. The activity of the solution at the injection is 2mSv. A sample of 10 mL of the animal's blood, taken 13 hours after the injection, is 0.0025mSv. The estimate volume of the animal's blood volume is?

A 10.0 L
B 8.0 L
C 4.0 L
D 2.5 L
E 1.25 L

IBO-2011, question A27, improved more reliable design

Of course the question should end with:
Fill out:

estimate volume of the animal's blood volume	_____ L

IBO2011, question A42, original design

A42. On a remote island, Dr. Yeh discovered a new plant species, which can produce either white or blue flowers. This species is mainly cross-pollinated by insects. Genetic experiments showed that the white-flower phenotype is recessive to the blue-flower phenotype. Statistical analysis revealed that 91% of these plants on the island produce blue flowers. If one is to randomly select two blue-flower plants and cross them, then what is the approximate probability that they are capable of producing white-flowered F1 offspring?

A 0.09
B 0.21
C 0.42
D 0.49
E 0.91

IBO2011, question A42, improved more reliable design

Of course the question should end with:

Fill out:

probability in %	_____ %

APPENDIX

Information about Statistical Evaluation

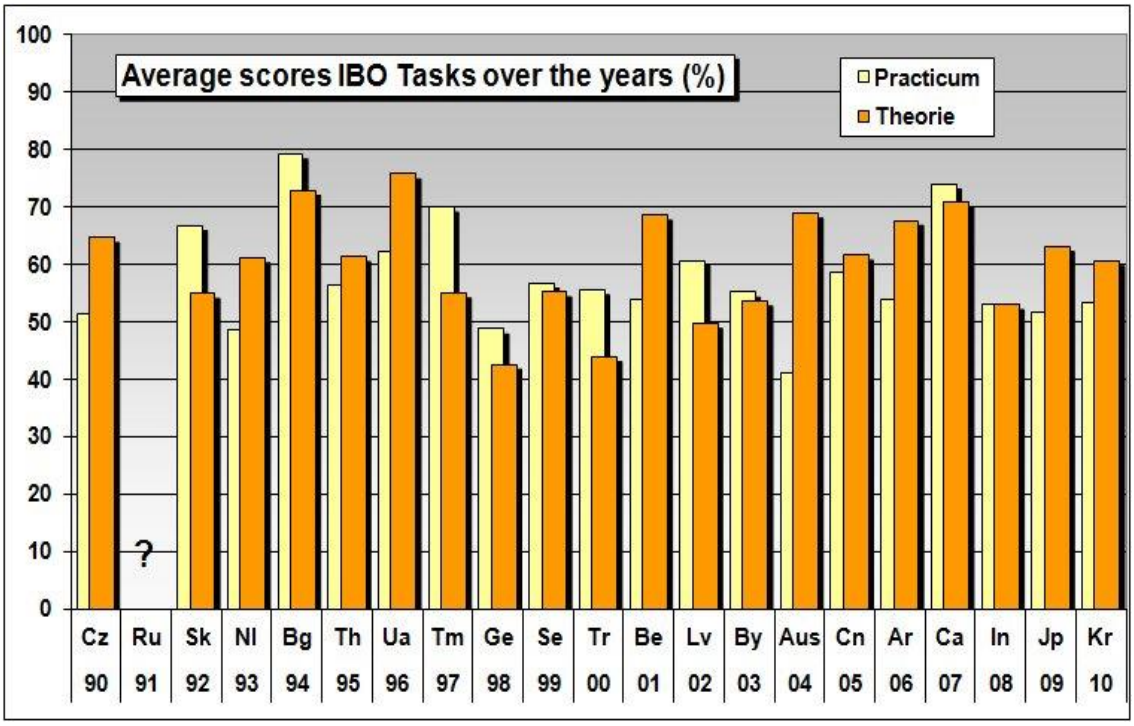
A statistical evaluation of a test offers useful information about the quality and validity of the test. It requires some expertise knowledge to interpret evaluation of data. Most important are:

Mean
The mean $\bar{x}$ is the "average" student response to a question. It is computed by adding up the number of points earned by all students on the question, and dividing that total by the number of students:

$$\bar{x} = \frac{1}{n} \sum x_i$$

Question Difficulty (p-value)

The difficulty of a question is indicated by the percentage of students answering the question correctly. We call this percentage the p-value.|
p-values are relevant for determining whether students have mastered the concept being tested.
Good questions are in the range of 30%<p<70%.
Also the p-value for the test as a whole is interesting. It gives an indication of the difficulty of the whole test. The picture shows p-values for Theory and Practice of all IBO's.

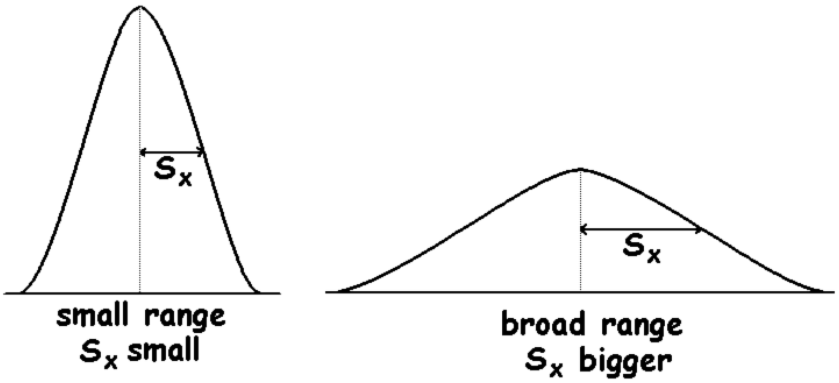


Standard Deviation (SD)

The standard deviation, written as **SD** or S_x, is related to the distribution of the scores of the students.

$$S_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}}$$

High values of **SD** means scores are spread out more so the test is more discriminating. See picture.



DI-value (Test Discrimination Index **DI**)

This offers a simple method to check the discriminating power of a test or question.

$$DI = p_u - p_l$$

p_u = overall p-value of top-group (top 25%) of the students
p_l = overall p-value of bottom-group (bottom 25%) of the students

Item Test Correlation (Rit)

Rit indicates the quality of individual items in a test. It shows whether a question is differentiating well between students as it offers information to which extent good students answer a specific question better than the bad students.

Good students mean: having a high score on the test as a whole
Bad students mean: having a low score on the test as a whole.

Rit can be calculated using the following formula:.

$$Rit = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\,S_x S_y}$$

x_i = scores of the students on a specific question
$\bar{x}$ = overall average score of students on this question
y_i = total score of each student on the test
$\bar{y}$ = overall average score of students on the test
S_x = standard deviation of the scores of all students on this question
S_y = standard deviation of the totalscores of all students on the test
n = number of students

With help of Excel it is not difficult to calculate Rit (use Pearson Product-Moment correlation).

The value of Rit always varies between –1 and +1.
Rit = +1 means that a questions is excellent. All good students did it correct, all bad student did it wrong, but this situation will never occur.
In practice, values of Rit will seldom exceed 0,50.
A question is “good” if the index is above 0,30; "fair" if it is between 0,10 and 0,30; and "poor" if it is below 0,10.
Questions with Rit = 0 are not discriminating at all.
In a selective test like our IBO these questions are useless.

Questions with Rit< 0 are suspicious, as it means that the good overall students perform worse on these questions than the overall bad students. This normally means that the question is tricky or the answer key is wrong.
If after having the test a question leads to a big dispute about the correct answer it is advisable to check Rit. In most of the cases Rit is low or even negative. We have experienced this situation several times in the IBO. It indicates that it is better to skip the question or adjust the answer key.
Unfortunately in many IBO ’s Rit was not included in the statistical evaluation, which is a pity.

In the case of just dichotomous scores for all questions (no intermediate scores, just *all* (full marks) or *nothing* (zero marks)) the Pearson Product-Moment Correlation calculation can be simplified to:

$$\mathbf{Rit} = \frac{\bar{X}_c - \bar{X}_f}{S_x}\sqrt{p * q}$$

$\bar{X}_c$ = average score on the test of the students answering the question correctly
$\bar{X}_f$ = average score on the test of the students answering this question wrongly
S_x = standard deviation of all the student scores on the test
p = proportion of students having this question correct
q = proportion of students having this question wrong

Reliability

The reliability of a test refers to which extent the test is likely to produce consistent scores. Normally the following formula of Cronbach **α** is used.

$$\alpha = \left(\frac{n}{n-1}\right)\left(1 - \frac{\sum_{i=1}^{n} S_i^2}{S_x^2}\right)$$

S_i = Standard deviation of scores per question

S_x = Standard deviation of total scores on the test

n= number of questions

Reliability	**Interpretation**
.90 and above	Excellent
.80 - .90	Very good
.70 - .80	Good
.60 - .70	Moderate
.50 - .60	Weak
.50 or below	Poor

In the case of tests with just dichotomous scores for all questions (no intermediate scores, only all (full marks) or nothing (zero marks)) it is easier to use the so called Kuder-Richardson formule (KR-20) in stead of Cronbach.

$$KR\text{-}20 = \frac{n}{n-1} * \left(1 - \frac{\sum_{i=1}^{n} p_i * q_i}{S_x^2}\right)$$

n = number of questions
p_i = proportion of students having a question answered correctly
q_i = proportion of students having answered the same question(s) wrongly
S_x = standard deviation of the total test scores

Information about t-score

In our IBO Practical and Theoretical test are balanced equally: 50% - 50%.
But both tests normally aren't equal in difficulty and this disturbs the balance.
IBO 2004 is a nice example.
The average score of the students for Theory and Practice was: 69% and 41%.
Practice was much more difficult, so (without correction) no equal balancing occurred.
Two 2004 silver winners both had a total average score of exact 64,3 %, but they differed rather much in their Pr and Th scores. See table.

Student name	Pr	Th	Overall
Georgi K.	54,3 %	74,3 %	64,3 %
You-Jin L.	38,7 %	89,9 %	64,3 %

Now the question is: who should be the best in the ranking?
Of course it should be Georgi, who performed better than You-Jin at the more difficult Pr test. But how to do this? What exactly should be balanced?

The solution is applying the t-score method. Be aware this is not the same as the t-test. In the t-score method two principles are combined.
- balancing for differences in the difficulty of the tests:
  this regards the average score on the test
- balancing for differences in the variation in scores of the test:
  this regards the standard deviation

The t-score procedure is simple.
For both Th and Pr, the score of each student is converted into a "standard" score:

$$standardscore = \frac{score - average\ score\ of\ the\ whole\ group}{standard\ deviation}$$

Now all students can be compared honestly in adding the standard scores for Th and Pr.

Be aware that students scoring less than (below) the average score of the whole group will have a negative standard score. The others, scoring better than the average, will have a positive standard score. In order to avoid the negative scores and too many decimals, usually the standard score is multiplied by 10 and to this number 50 is added.

$$standardscore = 10 * (\frac{score - average\ score\ of\ the\ whole\ group}{standard\ deviation}) + 50$$

It means that artificially the value of the average score is standardized to 50 and the value of the average standard deviation is standardized to 10.

The t-score corrects for differences in difficulty and for differences in standard deviation of two tests. That's the reason why the IBO ranking is based upon t-scores (also in 2004). Georgi got a t-score of 113,80 and You-Lin got a t-score of 113,64.

The influence of t-scores is small if students perform about the same on Pr and Th. For students performing very differently upon Pr and Th applying t-scores have more influence.

Look to the scores of the following 2 students participating in 2005.

Student name	Pr	Th	overall	t-score	medal
Elshad H	65,4 %	67,7 %	66,6 %	110,6	Silver
Sansabh M.	58,6 %	74,7 %	66,7 %	109,9	Bronze

We see that the overall score of Sansabh is slightly better than Elshad. But Elshad performs much better at the Pr test which, was more difficult than the Th test in that year.
So it was fair that Elshad had a silver medal while Sansabh, in spite of his slightly better overall score, had a bronze medal.

Balancing 4 Pr and 2 Th tests

In the IBO we usually have 4 Pr parts and 2 Th parts.
The 4 Pr parts usually have the same length in exam time and the same number of maximum points. So it is obvious we like to balance these four parts equally.
But if one part is much more difficult than the others or has a very big spread in results the balancing is not equally. For this reason it was decided that the t-score method should be applied for those 4 Pr parts. There is a simple way to do this. Just take the t-score of each of the 4 Pr parts and after that take the average (not the sum as this would lead to a big disbalance) of those four. This average is the overall t-score for the Pr part.

The Th test is different. Both parts normally differ in number of questions and number of points. It is obvious these two parts should not have the same weight. So for the Th test we just summarize the points for both parts and calculate the t-score of this total.
It means that the final ranking of the students in the IBO is based upon:
taking the average of the four t-scores of the practical exams and the t-score of the total student results on both theory parts. The final score is the sum of these two.