

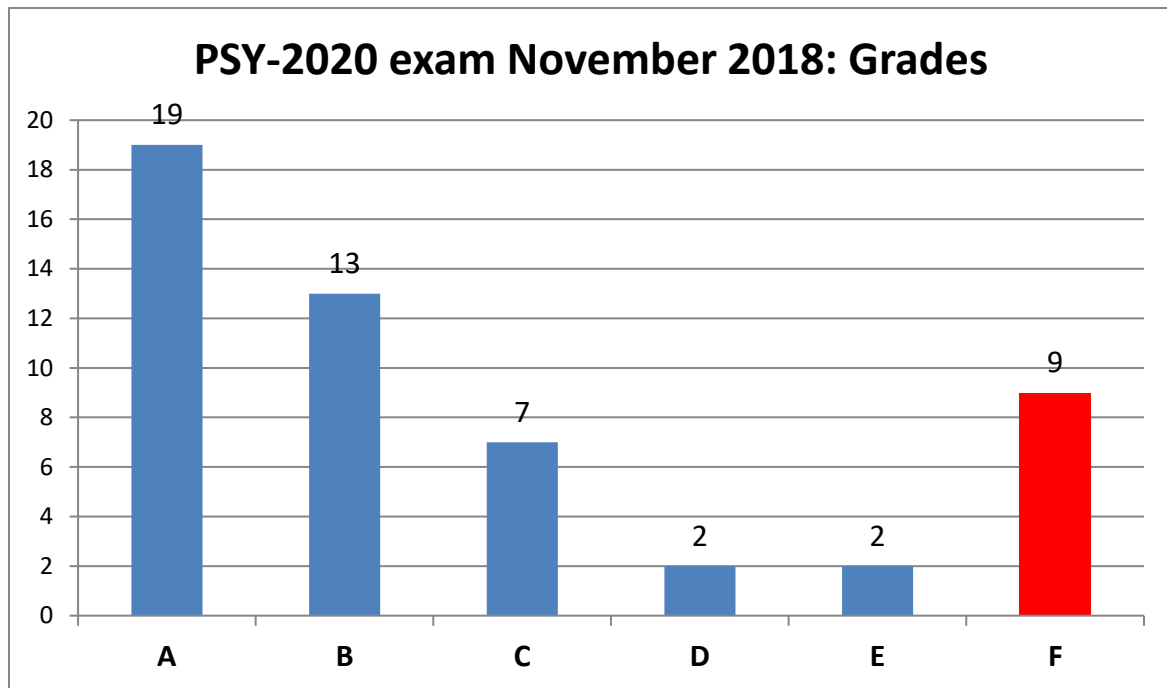
PSY-2020 Autumn 2018, Grading and results

“PSY-2020 The psychology of attitudes” is an English-taught course for second- and third-year Bachelor students in psychology. The course is open to international exchange students, and to students from other study programs. Its 12 two-hour lectures are based on a course textbook, research articles, and classroom exercises.

The electronic exam (November 20, 2018, four hours school-type) had 15 main questions (some with sub-questions). Together, the questions covered all sessions of the course. The required answers differed in length between a few words as the minimum, and circa one page as the maximum. Detailed, written instructions for the examiners defined the normatively correct answer to each question, as well as answer variants that should also count as correct. Each question had a certain number of maximum points that could be achieved, from 2 points (for simple definitions) over 4-6 points (for questions that required more thought and/or own formulations) to 10 points (for the detailed description of a full study or program of research). The maximum number of points for each question was listed in the exam; the examiners gave partial points for partly correct answers. The maximum number of points for the whole exam was 90 points. 75 points or more counted as an A. For each 15 points less, one grade was subtracted (B = 60-74, C = 45-59, and so on).

The course teacher served as the internal examiner in the examination committee. The external examiner (a native speaker of Norwegian) was a psychologist working elsewhere. The examination committee received the exams in an anonymous form. The examiners first graded all exams separately, using the written guidelines for this exam and the university's general "karakterskala" that defines the meaning of each grade verbally. Then they compared their ratings for each exam and determined the final grades.

Summarized results across the whole class are shown on the next page.



All "F" exams: No exam was available for grading (most likely, due to illness on the day of the exam)

The grades for submitted exams cover almost the full range of the grading scale.
Most exams received an A, B, or C.

Where do the differences in the grades come from?

As usual, the general picture is that they come mostly from the questions that required to actually formulate an answer in one's own words.

Simple reproductive questions ("what is the definition of ...") were answered correctly most of the time. Some questions asked for short (ca. half a sentence) examples to illustrate a concept. Here the answers used examples that sometimes came from the course materials and sometimes came from everyday life; in either case, these answers were usually correct. One question asked to apply a theory to a novel real-world problem, by filling in the words "few" and "many" into the word-gaps of pre-formulated sentences. Here most answers were correct too.

Other questions, in contrast, required long self-formulated answers: to describe, in one's own words, a complete study, or a set of critique against a given research instrument. Here points were lost. That was due to manifest knowledge gaps, as shown by answers that left out relevant parts or filled missing parts from (plausible but incorrect) everyday knowledge. The question to spontaneously construct research instruments stood out positively: A much higher percentage of answers was both correct and complete this year than in previous years.

All exams missed at least a few points underway, for answers that were incomplete, unclear to one or both examiners, or simply wrong. That, by itself, was not necessarily a problem: Missing up to 15 points across all 15 answers still allowed an "A." However, missing even more points then started to impact the grade.

Were there differences between the graders?

All examiners followed the same, detailed, written instructions. Below are statistical comparisons of the two examiners' independent grades (that means, from before they compared their grades and resolved the few differences). The 9 missing exams ("F") are not considered here; with these included, the results would be the same or even stronger.

The following two tables show that, on average across all exams, both examiners gave grades with the same mean and standard deviation. The difference between examiners is tiny (circa one tenth of a grade), and is not statistically significant. That means that both examiners were equally "lenient" or "strict":

Paired Samples Statistics

	Mean	N	Std. Deviation
examiner_1	2,05	42	1,10
examiner_2	1,93	42	1,07

Note. Means on a scale from 1 = "A" to 5 = "E".

Paired Samples Test

	Paired Differences		t	df	Sig. (2-tailed)
	Mean	Std. Deviation			
examiner_1 - examiner_2	,12	,50	1,53	41	,13

The next table shows an exam-by-exam comparison between the internal and the external examiners' initial grades, expressed as a correlation coefficient. What we see is a near-perfect match:

Paired Samples Correlations

	N	Correlation	Sig.
examiner_1 & examiner_2	42	,89	,000

In sum, there were no differences between examiners.

Were all exam questions “good”?

The following analysis of the exam’s internal consistency is based on the question-by-question points from the 43 “A to E” exams.

The left-hand table shows that Cronbach's Alpha is .85 – no matter whether we give equal weight to each of the 15 exam questions, or take into account that different questions had different numbers of maximum points. That is a good reliability score.

The right-hand table shows, separately for each exam question, what Cronbach's Alpha would be if that question had not been asked. If a question differs from the others, then removing that question would increase Cronbach's Alpha noticeably. However, as the table shows, we can remove any question, and Cronbach's Alpha would stay practically unchanged.

Reliability Statistics		
Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.85	.85	15

Item-Total Statistics	
	Cronbach's Alpha if Item Deleted
q1	.85
q2	.84
q3	.86
q4	.86
q5	.85
q6	.84
q7	.83
q8	.85
q9	.84
q10	.83
q11	.85
q12	.83
q13	.85
q14	.84
q15	.84

In sum, all questions were "good" and contributed in the same way to the total grade.