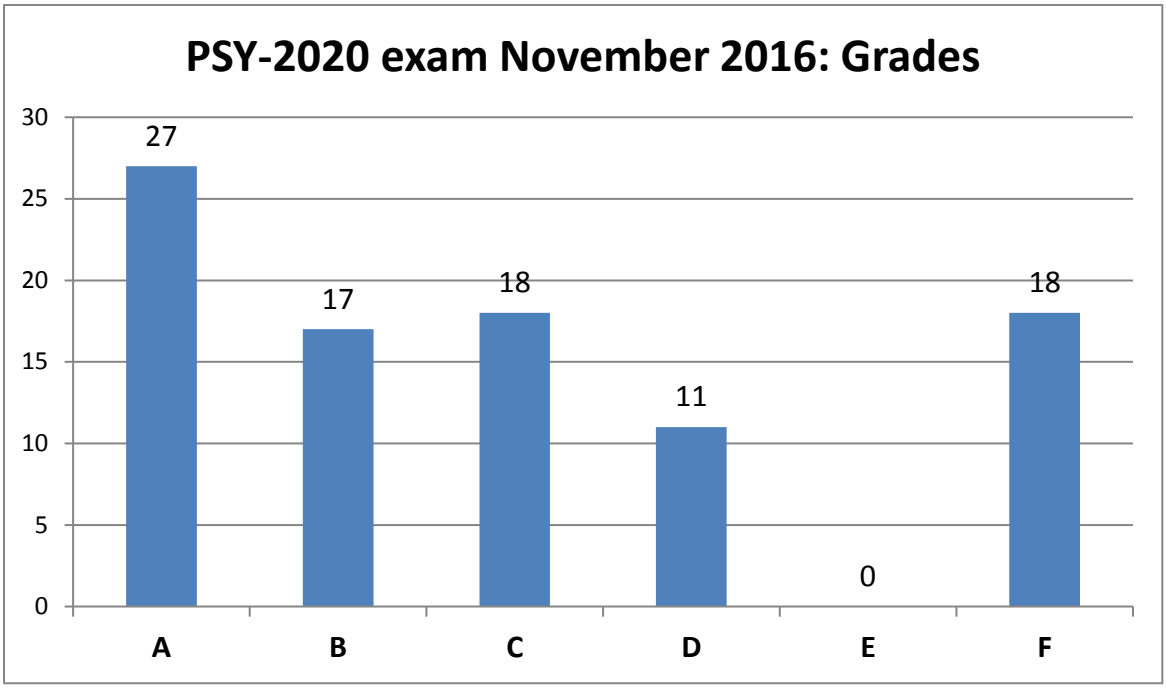
PSY-2020 Autumn 2016, Grading and results

“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 twelve two-hour lectures are based on a course textbook, research articles, and classroom exercises.

The electronic exam (November 10, 2016, 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).

Two examination committees graded the exams. In each committee, the course teacher served as the internal examiner. The external examiners were psychologists working elsewhere (both of them were native speakers of Norwegian). The examination committees 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 pages.



"F": not submitted (12), withdrawn (1), blank exam (1), or too few correct answers (4)

Statistics		
Grade		
N	Valid	91
	Missing	0
Mean		2,93
Std. Deviation		1,82

On the scale from 1 = "A" to 6 = "F", the mean corresponds closely to a straight "C" (which would be 3.0).

Grade				
		Frequency	Percent	Cumulative Percent
Valid	A	27	29,7	29,7
	B	17	18,7	48,4
	C	18	19,8	68,1
	D	11	12,1	80,2
	F	18	19,8	100,0
	Total	91	100,0	

The grades cover the full range of the grading scale. Almost 70% of all exams received an A, B, or C. That is a very good result.

Where do the differences in the grades come from?

Not unexpectedly, 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. In one question you were 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, a whole research program, or a set of critique against a given research instrument. Here points were lost. Partly 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. Another reason for lost points was a lack of ability or motivation to formulate answers such that they are complete. For example, one question asked to "describe and explain three alternative explanations" for a given effect. Some answers listed correct keywords, but did not explain them (one or two sentences of explanation each would have been enough), and therefore received fewer points than the possible maximum for this question. Finally, several exams could not spontaneously produce usable instruments to measure attitudes towards a novel attitude object, although we had a dedicated session on constructing research materials where we had practiced that.

In sum: The exams missed (partial) points at several places, for answers that were incomplete, unclear, or wrong. Missing up to 15 points across all 15 answers still allowed an "A", but from then on, the loss impacted the grade.

Were there differences between the graders or the grading committees?

All examiners followed the same, detailed, written instructions. The following table shows, separately for the two committees, the correlation between the internal and the external examiners' initial grades. What we see is a near-perfect match:

Correlations

Examiner team			Grade internal examiner
Exams 1-46	Grade external examiner	Pearson Correlation	,979
		Sig. (2-tailed)	,000
Exams 47-91	Grade external examiner	Pearson Correlation	,993
		Sig. (2-tailed)	,000

The following tables show that both committees gave final grades with the same average and standard deviation -- the difference between the committees is tiny (less than 1% of a grade), and is not anywhere close to statistical significance. That means that both committees were equally "lenient" or "strict":

Group Statistics

	Examiner team	N	Mean	Std. Deviation
Grade	Exams 1-46	46	2,935	1,90
	Exams 47-91	45	2,933	1,75

Note. Means on a scale from 1 = "A" to 6 = "F"

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means		
		F	Sig.	t	df	Sig. (2-tailed)
Grade	Equal variances assumed	,725	,397	,004	89	,997
	Equal variances not assumed			,004	88,672	,997

In sum, there were no differences between graders or between grading committees.

Were all exam questions “good”?

The following analysis of the exam’s internal consistency is based on the question-by-question points from the 78 exams (out of 91) that were (a) actually submitted, and (b) not withdrawn.

The left-hand table shows that Cronbach's Alpha is .91 if we give equal weight to each of the 15 exam questions, or .92 if we take into account that different questions had different numbers of maximum points. These are excellent reliability scores.

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 still be the same (or would become a tiny bit worse).

Reliability Statistics

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
,91	,92	15

Item-Total Statistics

	Cronbach's Alpha if Item Deleted
q1	,911
q2	,908
q3	,906
q4	,905
q5	,907
q6	,899
q7	,903
q8	,910
q9	,903
q10	,900
q11	,905
q12	,896
q13	,904
q14	,903
q15	,905

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