Uvažujme kostičkový svět níže. Agent (žlutý) se pohybuje světem pomocí akcí (N-North, W-West, E-East, S-South, a speciální akce D-Depart v terminálních stavech Exit). Reward/odměnu dostane pouze v případě opuštění cílového stavu (zelená a červená políčka). Předpokládejme discount factor $\gamma = 1$.

3		-30	120
2			
1	-70	-30	120
	1	2	3

Agent začíná v levém horním rohu. Vyzkouší několik trénovacích epizod, níže v tabulce. Každý řádek v tabulce trénovací epizody je n-tice (s, a, s', r).

Episode 1	Episode 2	Episode 3	Episode 4	Episode 5
(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0
(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0
(2,2), S, (2,1), 0	(2,2), N, (2,3), 0	(2,2), E, (3,2), 0	(2,2), N, (2,3), 0	(2,2), S, (2,1), 0
(2,1), Exit, (D), -30	(2,3), Exit, (D), -30	(3,2), S, (3,1), 0	(2,3), Exit, (D), -30	(2,1), Exit, (D), -30
		(3,1), Exit, (D), 120		

Vypočťete Q hodnoty níže pomocí **přímé evaluace** (direct evaluation/estimation) z trénovacích epizod:

Q((1,2), E) = _____ Q((2,2), E) = _____ Q((3,2), N) = _____