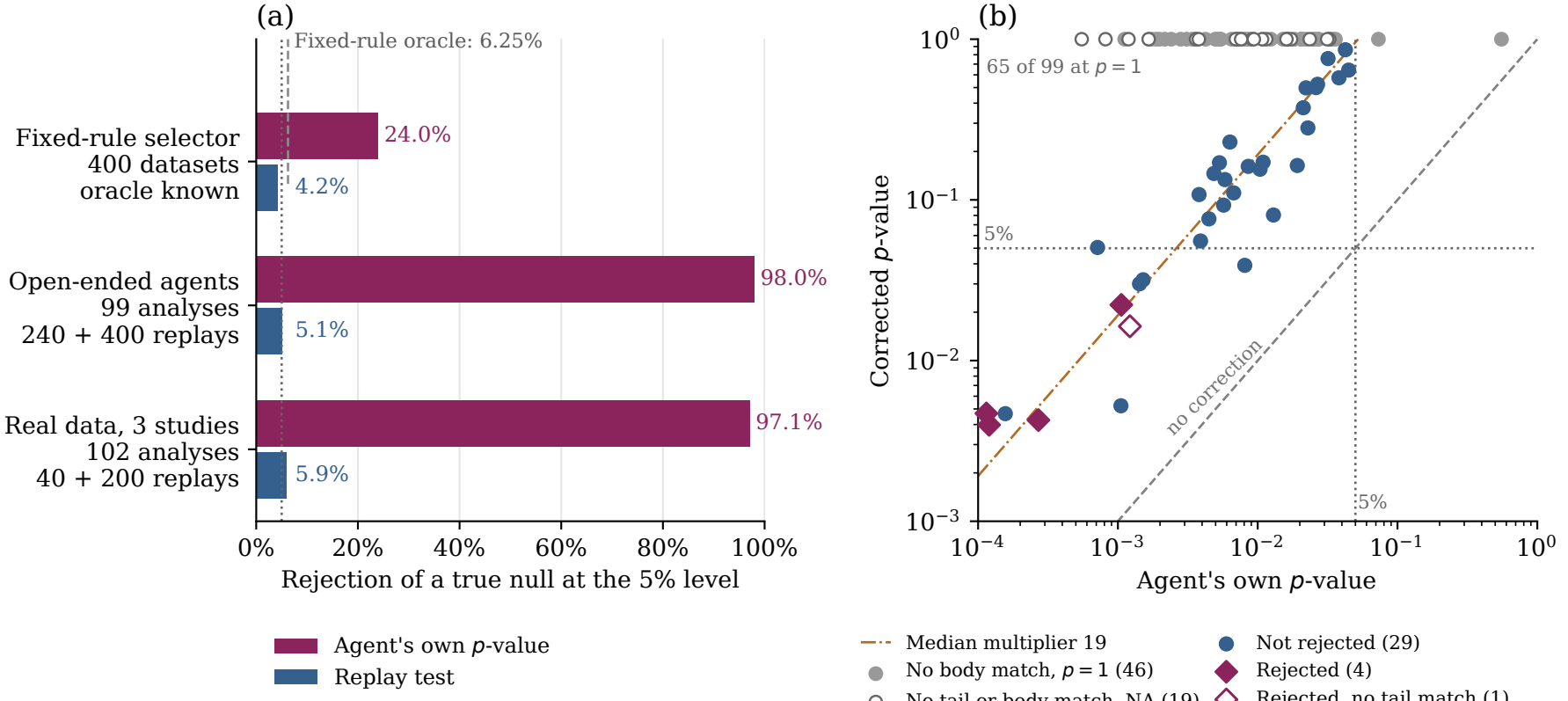

(a)
Fixed-rule oracle: 6.25%
Fixed-rule selector
400 datasets
oracle known
24.0%
4.2%
Open-ended agents
99 analyses
240 + 400 replays
98.0%
5.1%
Real data, 3 studies
102 analyses
40 + 200 replays
97.1%
5.9%
0%
20%
40%
60%
80%
100%
Rejection of a true null at the 5% level
Agent's own p-value
Replay test
(b)
65 of 99 at p = 1
5%
no correction
5%
Corrected p-value
Agent's own p-value
Median multiplier 19
No body match, p = 1 (46)
No tail or body match, NA (19)
Not rejected (29)
Rejected (4)
Rejected, no tail match (1)