Android malware authors have increasingly relied on techniques to hinder dynamic analysis of their apps by hiding their malicious payloads or by scheduling their execution based on complex conditions. Consequently, researchers devise different approaches to bypass such conditions and stimulate the malicious behaviors embedded within the Android malware. Despite the availability of different behavior stimulation approaches and dynamic analysis tools that implement them, they are seldom empirically evaluated to assess their applicability and effectiveness. In this paper, we survey the literature to identify different behavior stimulation approaches and assess the performance of three tools implementing them against four datasets of synthetic and real-world malware. Using the obtained results, we highlight significant limitations of such analysis tools, including their instability and their inability to stimulate scheduled behaviors even in automatically generated synthetic malware. Those limitations enable simple approaches based on the random manipulation of an app's User Interface (UI) to outperform more sophisticated behavior stimulation approaches. We aspire that our results instigate the adoption of more rigorous evaluation methods that ensure the stability of newly-devised analysis tools across different platforms and their effectiveness against real-world Android malware.