Program obfuscation is increasingly popular among malware creators. Objectively comparing different malware detection approaches with respect to their resilience against obfuscation is challenging. To the best of our knowledge, there is no common empirical framework for evaluating the resilience of malware detection approaches w.r.t. behavior obfuscation. We propose and implement such a framework, called FEEBO that obfuscates the observable behavior of malware binaries targeting Microsoft Windows operating systems. To assess the framework’s utility, we use it to obfuscate known malware binaries and then investigate the impact on detection effectiveness of different popular behavior-based malware detection approaches. We find that the obfuscation transformations employed by FEEBO can affect the accuracy of such detection approaches significantly. FEEBO is therefore an effective and fair way to test the degree of resilience of behavior-based malware detection approaches against behavior obfuscation.