Screening in systematic reviews within the social sciences can be a complex endeavor. Often, the intervention of interest in such reviews is not a particular named treatment, but rather a broader concept, approach, or method believed to exert an effect on an array of different outcomes through various more or less well-defined mechanisms. This inherent complexity of interventions targeting e.g. learning processes requires screeners to be aware of and familiar with topic matter terminology in order to judge whether a given intervention is in fact within the scope of the review question. Our review on the effects of testing frequency is a case in point. In this review, the intervention (student testing) was a type of learning strategy that is ubiquitous in education and used in a variety of ways and for very different purposes. Testing can be used as a formative tool, e.g. to promote retention of academic content, adjust instructional strategies, and uncover student needs for remediation or more intensive support. In most school systems, testing is also used summatively for assigning grades, determining graduation or certification, and for school accountability assessment. Important to note here is that the distinction between formative and summative testing is not clear-cut as tests can serve both purposes simultaneously and can have more or less stakes attached to them, both from a student and from a school perspective. It follows that testing is not a uniform type of intervention, but a multi-facetted phenomenon encompassing a variety of approaches and a heterogeneous terminology (tests are not just called tests, but may also be referred to as e.g. quizzes, progress-monitoring, curriculum-based measures, and retrieval practice). Judging the eligibility of particular interventions therefore requires subject matter familiarity.

Further adding to the complexity of the screening process is the fact that interventions involving tests within education are often designed as multi-component treatments involving more than just a change in testing regimen. As an example, curriculum-based measurement (CBM) often entails not only testing, but also decision rules regarding required instructional changes based on measurement results, professional development, particular feedback procedures and types of learning material etc. As we were interested in the particular effects of testing students more or less (and not more broadly the effects of instructional changes informed by test results), we had to pay close attention to the issue of co-interventions, that is, the existence of other intervention components with the potential to confound the effects of testing. As it turned out, judging the confounding potential of intervention components – and whether these components were balanced between intervention and control groups - was far from simple. Therefore, we felt it necessary to include more studies than usual in order to ensure that studies were read and assessed in detail by several reviewers. As a consequence, a number of studies retained for risk of bias assessment turned out to not live up to the full text inclusion criteria (these studies were reported on the “Excluded with reasons” list).

Finally, we were faced with what would seem a rather basic challenge: determining the testing frequencies of treatment and control/comparison groups and the nature of the practice and outcome tests used. For a study to be eligible, we had to have information about the number of tests received by students (and there had to be a variation in testing frequency between groups). In some cases, this information was not readily available – or the information provided was hard to interpret (e.g. treated students may receive a particular set of tests whereas controls received “business-as-usual”, but if this “business-as-usual” also involved testing to an unknown extent, determining what the comparison was in terms of testing frequency became challenging). Also, there were instances where the tests used as part of the intervention (the practice tests) were identical to the tests used for ascertaining the effect of intervention (that is, the same tests were used both for practice and as outcomes). In these cases, treated students would have an unfair advantage over control students simply because they were exposed to the exact same material on multiple occasions. All these complexities put together made the screening process for this review more difficult than usual, even for review members with extensive experience within the field of educational research.