A Rapid Review of Interim Assessment Use

Nathan Dadey

Calvary R. Diggs

The National Center for the Improvement of Educational Assessment

Interim Assessment Research Synthesis

Interim assessments are a broad class of assessments that are currently used in schools to guide a variety of educational practices (Perie, Marion, & Gong, 2010). Although some schools used these assessments prior to the No Child Left Behind Act ([NCLB] 2001), interim assessments were rapidly adopted following the act's passage. NCLB (2001) expanded the number of end of year tests designed to assess schools. It held schools accountable for the educational attainment and proficiency of students in third through eighth grades. Interim assessments were adopted based on schools' desire to quickly and effectively support student achievement on end of year tests (Shepard, 2010). The claims made by users and developers of interim assessments cited the formative assessment literature-base (Burch, 2010; Shepard, 2010; Troy, 2011), which suggests that the frequent and strategic collection of student data can be used to support instruction (e.g., Black & Williams, 1998), inform programmatic decisions, and predict student performance on criterion assessments (e.g., Hintze & Silberglitt, 2005). Unfortunately, the link between formative and interim assessment is unclear, given that they are different (Perie et al., 2010). It is also unclear if interim assessments can be effectively utilized for predictive and evaluative uses, in addition to formative or instructional uses (Shepard, 2010).

Despite the lack of evidence regarding the validity of the claims made regarding this brand of assessments, the policy context created an environment that beckoned the rapid adoption of interim assessments (Shepard, 2010). In fact, the purchase of interim assessments can be quite expensive for districts, and some suggest that administrators will prioritize interim assessments when making budget decisions (Herman, 2007). Thereby sacrificing other potential supports, resources, and staff that could support student and school achievement for a brand of

assessments predicated on a theory of action and specific practices that, to-date have not been supported by the research literature in a clear synthesis. With the rise of Multi-tiered systems of support, data-based decision-making, assessment literacy, and formative assessment, it is likely that educational assessments – including interim assessments – will remain embedded in school, district, and state practice. However, it is imperative that this research be synthesized.

The trend to adopt interim assessments continues to this day, and there is now a critical mass of research on interim assessments. Given that there has not been a systematic review of assessment that can be identified as 'interim,' the purpose of this study was to (a) identify the current literature on interim assessments, (b) summarize available literature on interim assessments, (c) understand how interim assessments are used, and (d) gather evidence regarding the efficacy of those uses.

Theoretical Framework

Many scholars have focused on describing and understanding interim assessments. Perie et al.'s (2010) seminal article regarding interim assessment is frequently cited to understand and conceptualize just what interim are and how they differ from other classes of assessment designated for particular uses, such as formative and summative. The purpose of their seminal article was to provide a definition of interim assessments as well as a framework for users to incorporate in their purchase and development plans. In addition, the authors provided a historical context regarding the rise of interim assessments due to the limitations of tests designed and used for summative purposes.

To begin, Perie et al. (2010), like other authors (e.g., Brown, Steege, & Bickford, 2014; Christ & Keller-Margulis, 2014), create an illustrative dichotomy between formative and summative assessments. In short, summative measures are intended for evaluation,

accountability, and end of year grades. Their use is restricted to one time of measurement, and their stakeholders and are often at the statewide level. The scores are used for high-stakes decisions. Some examples of summative assessments are end of year, grade, or course tests; final exams; and the drop-out rates in high schools following the implementation of a student engagement program; or measures of generalization given at the conclusion of programmatic intervention.

In contrast, formative assessments are used for instructional planning and decision-making. Stakeholders of formative assessments are often classroom teachers. Unlike summative assessments, formative assessments are administered in rapid-cycle, and have little use beyond the classroom context. Formative assessments may be standardized or informal, and are used to inform low-stakes decisions such as assigning students to groups and gauging student understanding of a particular concept.

Situated between formative and summative assessments, exists interim assessments. Interim assessments do not a particular purpose. Instead they are contextualized entirely by their uses. Interim assessments are used several times a year or once mid-year. They are often scheduled assessments that are administered at key decision-points. Rarely is the scheduling determined by classroom teachers. Instead, school or district administration determine the testing windows in which they may be administered. Interim assessments may – and are often purported – to be used for the classroom, but they are designed for stakeholders at the school and district levels.

The above is a brief overview that situates interim assessment as a class of assessments that are between summative and formatives uses. Other authors have summarized the literature

and theoretical distinctions of formative (e.g., Black and William, 1998; Brown et al., 2014), summative, and interim (Perie et al., 2010).

Characterizing Use. In the writing above, one may note that there is a distinction that remains somewhat ambiguous and convoluted relating to the distinction between an assessment's class (i.e., formative, interim, summative) and the intended interpretation and use of information that the assessment provides (e.g., formative, summative). The above distinguishes the classes of assessment, which provide some information. However, it is also valuable to understand the intended use of assessment (Kane, 2006). The interpretation and use of an assessment's score for a particular purpose is what is validated, not the assessment's affiliation with a particular class. Though, it is important for broad classification purposes to understand the broad class, in the case of interim assessments – where class and use are fluid, it is important to understand how assessments are used.

Kane's (2006) argument-based approach to validity is one way for defining assessment use as it relates to validity. The work posits that an assessment is neither valid or invalid. Instead, the scores derived from the assessment may be used for particular purposes, which should be validated. Under this approach, the 'claim' is the nomenclature used to denote 'use.' In this paper, the two may be thought of synonymously. Kane's (2006) approach requires that a clear statement is provided regarding the intended use, and that the evidence to support the claims is evaluated (Cook, Brydges, Ginsburg, & Hatala, 2015). This particular approach to validity compliments the previous descriptions of broad assessment categories. For example, assessments may be characterized broadly as formative, summative, or interim. However, the important aspect is what the results of the assessment are used for, not the general label provided. In the case of interim assessments, focusing on the use may provide further clarity.

Kane's (2006) proposed a theoretical approach regarding the process of validating a score and its various claims. The purpose of the work was not to attempt to define assessment use. Instead it aimed to describe how users and developers could work to evaluate the plausibility of a particular interpretive claim. As such, it is necessary to examine other typologies that aim to characterize assessment use that may be applied to interim assessments.

There are several authors that attempt to characterize the use of assessment (Crane, 2010; Perie et al., 2010; Salvia, Ysseldyke, & Bolt, 2011). Salvia et al (2011) provide a seven purposes of assessment: screening, progress monitoring, planning and modifying instructional, allocating resources, special education eligibility determination, program evaluation, and accountability (see source for specific descriptions). Salvia et al.'s (2011) framework for assessment use is intended to represent all uses of educational assessment. It is worth noting this particular work was written from a perspective of school psychology and special education, which compliments but is different than the programmatic perspective of general education departments that tend to adopt or develop interim assessment programs. Crane (2010) cites eight purposes of assessment: diagnosis, prediction, preparation, placement, student evaluation, school intervention, promotion and graduation, and local accountability. Their approach is used to support the understanding and evaluation of educational assessment programs, with a focus on interim assessment. Finally, there is Perie et al.'s (2009) typology which has already been introduced. Under this approach, the authors specifically classify interim assessment use. They assert that interim assessments broadly address at least one of three purposes: instructional, evaluative, or predictive.

Together, the typologies provide various ways of classifying interim assessment use. Each approach has unique advantages and disadvantages. Each typology further clarifies use: Crane (2010) and Salvia et al. (2011) provide more discrete categories, while Perie et al. (2009)

developed broader classifications. Each is able to describe a variety of applications of interim assessments. Unfortunately, the areas of use broadly denote specific claims in ways that reflect the precision of Kane's (2006) interpretive argument. As such, regardless of the framework used to characterize use, it will be necessary to examine specific claims made about use in the literature. It is the difference between an instructional (Perie et al., 2009), school intervention (Crane, 2010), and instructional planning and modification (Salvia et al., 2011) compared to – for example – using assessment results to identify if a particular series of lessons need to be retaught. As such, an existing typology is important as a way to initially conceptualize use; however, there are specific uses under each category that should be stated and clarified. Unfortunately, no such typology currently exists.

Defining interim assessments. Collectively, interim assessment is a term used to describe a large range of assessments that fall somewhere between summative and formative uses for assessments. Of the three typologies, Perie et al. (2009) will be adopted for the theoretical framing of interim assessment use given that is seminal in the scholarly literature on interim assessment. As such, a more complete and thorough definition of interim assessment will be described using their framework.

Interim assessments are assessments that fall between "classroom-level, low-stakes, high-frequency formative assessment and state-level, high-stakes, low-frequency summative assessment" (Crane, 2010, p. 4). The term "interim assessment" was adopted by Perie, Marion, & Gong in 2010 to serve as a catch-all for this distinct kind of assessment, which had been previously referred to not only as interim, but also, *benchmark, predictive, periodic, district, local*, and, perhaps most challenging, *formative* assessment. To provide clarity around the term, Perie et al. (2010) provide the following definition for interim assessment:

Assessments administered during instruction to evaluate students' knowledge and skills relative to a specific set of academic goals in order to inform policymaker or educator decisions at the classroom, school, or district level. The specific interim assessment designs are driven by the purposes and intended uses, but the results of any interim assessment must be reported in a manner allowing aggregation across students, occasions, or concepts (p. 6, emphasis added).

Based on the above definition, we conclude that there are at least three criteria to identify an assessment as interim – (a) it must be given during instruction, and presumably not at the end of the year, (b) it must inform a specific decision and (c) it must allow for aggregation (Perie et al. (2010) provide two similar criteria in their elaboration of this definition). Thus interim is a term meant to describe a specific intersection of assessment administration, decision making and assessment design.

In short, Perie et al. (2009) describe interim assessments as measures that assess current knowledge compared to some standard, can be aggregated for greater use beyond the teacher, and be used for a particular purpose. These assessments are used for three clear areas of use. The first is instructional, which is defined interim assessments are used to "adapt instruction and curriculum to better meet student needs" (p.7). Under instructional uses for interim assessments, education professionals use the data to adjust and improve instruction, identify student strengths and weaknesses, and understand student conceptual and procedural errors with the ultimate aim of improving student learning.

Next, evaluative uses are, "designed explicitly to provide information to help the teacher, school administrator, curriculum supervisor, or district policymaker learn about curricular or instructional choices and take specific action to improve the program, affecting subsequent teaching and thereby, presumably, improving the learning" (p. 8). Under evaluative uses, results

are used to enforce quality standards for the curriculum and pacing within and across schools, understand student performance across settings separate from differences in grading standards across settings, understand the efficacy of instructional programs and initiatives administered at the school or district levels, measure student growth, understanding student mastery and weaknesses of particular concepts, adapt the curriculum for future academic years; and to understand the alignment between the test, curriculum, and instruction. Perie et al. (2009) note that evaluative uses do not aim to directly intervene for students experiencing educational difficulties. Instead, they are used to evaluate the efficacy of teachers, programs, or strategies.

Finally, predictive uses are, "designed to determine each student's likelihood of meeting some criterion score on the end-of-year tests" (p. 8). These uses of interim assessments are tied understanding how a student may perform on summative assessments. Summative assessments could be end of year or grade tests, high school exit exams, or success with postsecondary curricula (Perie et al., 2009).

Thus, in this paper, interim assessments are defined as mid-cycle assessments given strategically for instructional, evaluative, or predictive purposes. Perie et al (2009) provides a variety of examples of these uses: understand what a student knows and can do, predict performance on summative measures, compare or assess effectiveness of instructional strategies, understand student achievement at the district level, understand knowledge gaps at the classroom level for individual students, determine if students are on track to pass summative assessment, to identify corrective feedback to help students pass a summative assessment, motivate and provide feedback to students about their learning, provide information to modify future instruction of similar material, monitor pacing of delivering the curricula, provide a thorough examination of students' understanding, and determine whether to move onto the next instructional unit. The

previous examples would be clarified as specific uses (i.e., claims) of interim assessments, and each could be characterized as instructional, predictive, or evaluative. It is an assumption of this paper, that both are needed to under interim assessment use at both a high and practical level.

Although this paper relies heavily on the Perie et al (2009), there are other seminal and valuable scholarly and professional articles on interim assessments. They discuss the policy context which gave rise to interim assessments, considerations in adopting and using interim assessments, and exploring nuances in the roles of the vendor and the purchaser of interim assessments (e.g., Burch, 2010; Shepard, 2010;). Although this research exists, there is not currently a review that examines the use and efficacy of interim assessments for particular purposes.

Delimitations to the current definition. Interim assessments are a broadly defined category intended to encapsulate a diverse class of assessments. The definition of interim assessments provided above, if followed strictly, results in investigating related areas such as formative assessment (Black & Williams, 1998; Kingston & Nash, 2011), data-driven instruction (e.g., Jung et al., 2018; van Geel et al, 2017), and multi-tiered systems of supports (e.g., Eagle et al., 2015; Stormont & Reinke, 2013). Each of these areas represent an extensive literature-base with assessments that could certainly meet the criteria for interim. However, given that research reviews exist for such studies, the time constraints of the review, and the timeliness of the topic, the primary focus of this literature review was to describe the body of research on interim assessments unlikely to be captured by previous research programs. In summary, to avoid confusion, capture research that has perhaps been excluded and overlooked in other reviews, and to describe assessments that are uniquely interim, this review provides a first, targeted rapid review (Grant & Booth, 2009)

Purpose

Primarily in response to the pressures imposed by federal accountability policy (i.e., the No Child Left Behind Act of 2001), interim assessments were rapidly adopted by school districts throughout the country in the early 2000s despite limited evidence of their utility (Shepard, 2010). Although this trend of adoption continues to present day, there is now a critical mass of research on interim assessments. To date, however, there has not been a systematic review of assessments that can be identified as 'interim.' The purpose of this study was to (a) identify literature on interim assessment use, (b) summarize available literature on interim assessments, (c) understand how interim assessments are used, and (d) gather evidence regarding the efficacy of those uses. The goals and objectives of this research were met through conducting a rapid review (Grant & Booth, 2009) of the peer reviewed and grey literature on interim assessment use.

Method

Description of the Review and Aims

This study was a rapid review, which is described by Grant and Booth (2009). Rapid reviews are conducted to understand current issues in policy and practice. As the name suggests, such approaches aim to provide a systematic review of the literature under time constraints. Given this approach, the literature review process (i.e., search strategies, inclusion of grey literature, coding and extraction of key variables, and quality appraisal) may be limited based on the researchers' judgment. These restrictions are explicitly reported.

The current literature review is a rapid review given that the resources funding this scholarship as well as the final product were both time-sensitive and conducted over approximately eight weeks, including conceptualization. The current paper reports on the

findings of the rapid review. Articles were obtained using a systematic search procedures that will be described in "Identification of Studies." However, the majority of the findings in this rapid review ([b], [c], and [d] of the study purpose) are based on the recommendations of an expert in educational measurement who conducted a forward and backward citation search of Perie et al. (2009) using Google Scholar during summer of 2019. The efforts from the rapid review were also used to lay the groundwork for future reviews of interim assessments by documenting the variety of research available relevant interim assessments ([a] of the study purpose), while clearly articulating the inclusion and exclusion criteria. As such, a more comprehensive search was conducted; however, those findings were not coded and evaluated in the current review.

Identification of Studies

Three methods were utilized to gather a comprehensive body of research for inclusion. The first method utilized an expert in the design, scaling, and use of educational assessments for accountability. They were tasked with gathering research relevant to interim assessment to guide the current study as well as recommend articles for consideration in including. They used their own knowledge of the field as well as Google Scholar in conjunction with forward and backward citation search of Perie et al. (2010). That search resulted in a total of 43 articles that were identified for further examination.

The second method was a traditional systematic literature search using Academic Search Premiere (ASP) and ERIC via EBSCO host on June 28, 2019. Two sets of search terms were applied to both databases in order to retrieve relevant scholarly articles, dissertations, theses, and research reports. The first set of search terms broadly related to *terminology* (interim assessment, benchmark, formative assessment, predictive assessment, standards-based assessment, predictive

validity, instructional assessment, evaluative, unit test, periodic, quarterly), *typologies of use or evidence* (assessment, validity, evaluation, implementation), and *instructional settings* (classroom, elementary school, middle school, junior high, high school, primary school, secondary school, teacher, administrator, district). The first set of search terms were required to appear in either the title or the abstract. The search was conducted iteratively and modified. The key studies identified in the first search method were used to validate the comprehensiveness of this systematic literature search using ASP and ERIC. Four iterations were conducted before arriving at the current search terms.

In addition, a second search was conducted in ASP and ERIC via EBSCO on July 1, 2019. This set of search terms related to specific interim assessments endorsed on the assessment pages of state departments of education and public instruction (Smarter Balanced, MAP, i-Ready, AIMS Web, DIBELS/Acadience, FAST, Istation, Phonological Awareness Literacy Screening , STAR Early Learning, NGSS, STAR, Terra Nova, Leap 360, Brigance Early Childhood Screens III, STAAR Interim Assessments). Both the general and targeted assessment sets of search terms are available in Appendix A.

The final search method were requests to interim assessment vendors for technical reports related to use of relevant interim assessments. The vendors of the assessments targeted in the second database search were contacted. All studies were then pooled together and underwent screening to filter out unrelated research documents (e.g., medical, preschool, university programs). Then the inclusion criteria were applied to the remaining articles.

In conducting literature syntheses, the accuracy, replicability, and comprehensiveness of the search procedures are often a key consideration, which often requires attention to the inclusion of grey literature and considering publication bias. Grey literature is research that goes

unpublished for various reasons, while publication bias reflects the tendency of statistically significant results to be published and those without to go unpublished (Card, 2015). Grey literature can come in the form of dissertations, research and technical reports, conference papers, and unpublished manuscripts (Harris, Hedges, & Valentine, 2009). The current rapid review as well as the general procedures included dissertations as well as research and technical reports. The broader review also incorporated technical reports and unpublished manuscripts.

Criteria for Inclusion

The inclusion criteria were designed to identify research relevant to interim assessment use. The criteria were (Gate 1) studies were written in the English language, conducted in the United States, and published since 2000; (Gate 2) conducted in K-12 settings; (Gate 3) an assessment was a key component of the study and functioned as an independent variable; (Gate 4) the focus of the assessment was academic in nature (e.g., language arts, mathematics, social studies, science); (Gate 5) used an assessment that was administered by school or affiliated staff for school use (i.e., not administered by parents, clinics, outside consultants); (Gate 6) the assessment can be described as interim – (a) multiple measurement points or used midway through the term, (b) assessment data were used for a purpose that may be broadly described as evaluative, instructional, or predictive, and (c) data may be aggregated or disaggregated and still serve a purpose; (Gate 7) the interim assessment is/was commercially available or was developed by a school or system(s); and (Gate 8) the study evaluates an interim assessment using a traditional experimental or quasi-experimental design, observational methods, or recollections.

The inclusion criteria were designed to identify studies and assessments that fall under our current typology of an, “interim assessment.” In the current conceptualization, interim is used to denote assessments that are not clearly used for formative or summative purposes.

However, there are various ways that each of these concepts could be defined, which, in turn, would provide a different approach to the inclusion and exclusion process. This study aims to provide a wide, scoping search of the literature. Exclusion criteria for both summative and formative assessments were defined. Focal assessments that were clearly used for formative purposes (i.e., teacher administered, informal, conducted frequently, and data cannot be aggregated and meaningfully generalized beyond the individual classroom) were designated as formative assessment and excluded using either the formative assessment, typically at Gate 6. Likewise, if the focal assessment was clearly used for summative purposes (e.g., an end of the year state test) the study was typically excluded at Gates 3 or 6.

Inclusion and Coding Procedures

All studies eligible for inclusion were compiled in a single Microsoft excel document, and filtering criteria were applied to screen out clearly ineligible studies and duplicates. This resulted in 4,059 articles eligible for inclusion in the review. Each article was then assessed for agreement with the study inclusion criteria, resulting in a total of 125 studies included. The first and second authors selected 20% of the eligible articles and applied the inclusion criteria to them. The inter-rater agreement will be reported in the final paper, as this process is in progress. [The inter-rater agreement was ___ %, prior to resolving all agreements. The most common area of disagreement occurred at Gate 6, which was also one of the most common gates for exclusion. The overall inclusion process is represented in Figure 1. One can observe the entire high-level search and inclusion process conducted.

A similar process was conducted for the expert identified and forward and backward citation search articles. The exception is that the first and second authors worked closely together to determine which articles met inclusion and exclusion criteria. Those articles were used to

refine the criteria before applying them to the larger body of research. As a result, all 43 articles were reviewed for inclusion by the first and second authors, and exclusion decisions were only made after both authors came to agreement. The policy during this process was to include an article, unless there was compelling evidence that it did not contribute to the purpose and goals of the rapid review. In total, out of the 48 studies that were recommended, 20 were included and coded in the present review.

Each article designated for inclusion was coded for features relevant to (a) research design, (b) characteristics of the sample, (c) characteristics of the interim assessment, and (d) characteristics of assessment use. In total, these four broad categories represented 48 unique codes. A codebook was developed based on these criteria and refined through iteratively coding a single article (Abrams et al, 2014). In addition, an excel sheet was created with dropdown options to minimize disagreements. However, the broad categories of use were populated based on the coders' analysis of discrete uses described in each study.

Discrete uses within study were classified as instructional, evaluative, or predictive. Discrete uses were also subsumed into researcher-developed categories to provide a more detailed understanding of how the interim assessment was used. In other words, this study also utilized a grounded approach (Hook, 2015) to develop clear claims regarding assessment use that contained more specificity than Perie's et al. (2009) broad categories. Of note, the categories developed under the grounded approach were not mutually exclusive. As such, the multiple discrete uses could be subsumed under one use statement. In addition, a discrete use could – though, rare – appear in multiple use statements. All authors iteratively coded Abrams et al (2014) until agreement was reached regarding the coding procedures.

Results

The current study is a rapid review on interim assessments used in education. The findings are discussed in relation to the four purposes of the review. The first purpose was to (a) identify a body of literature on interim assessments and their use. As such, a total of 107 studies were identified through systematic search criteria (k = 107) and forward and backward citation searching by a scholar familiar with the research domain (k = 20).

Summary of the Literature

The second purpose of this research was to (b) summarize available literature on interim assessments. This was done by using the 20 articles researcher identified articles from forward and backward citation searches (Abrams et al., 2014; Blanc et al., 2010; Bulkley et al., 2010; Burch, 2010; Christman et al., 2009; Clune & White, 2008; Davidson & Frohbieter, 2011; Diaz-Bilello, 2011; Goertz, 2009; Halverson, 2010; Jones, 2013; Konstantopoulos et al.., 2016; Kulp, 2017; Lai, 2009; Lange, 2014; Medford, 2014; Olah et al., 2010; Ross, 2012; Shepard et al., 2011; Underwood, 2010).

General information relevant to each of the included studies is presented in Table 1. At a high level, it can be observed that the majority of the studies were qualitative in nature (k = 17). The remaining three studies used quantitative methods, while one other (Underwood, 2010) used both quantitative and qualitative methods. As such, the majority of articles were focused on describing characteristics of interim assessment use to better understand how teachers, administrators, and systems use them (Abrams et al, 2014; Blanc et al, 2010; Bulkley et al, 2010; Burch, 2010; Christman et al, 2009; Clune & White, 2008; Davidson & Frohbieter, 2011; Goertz, 2009; Halverson, 2010; Jones, 2013; Kulp, 2017; Lange, 2014; Medford, 2014; Olah et al, 2010; Ross, 2012; Shepard et al, 2011; Underwood, 2010).

Of the articles, nine were dissertations. Thus, the majority of studies were dissertations (peer reviewed research [k=7], research reports [k=5]). The primary purpose of this research question is to provide readers with an understanding of the context and goals of the included studies. As such, trends are noted.

Interim assessments were coded for various categories. Of the 20 studies, 7 were developed in-house by the school system or district, 4 purchased directly from vendors, 5 jointly developed assessments, and 4 were unreported. All interim assessments were broad measures of the curriculum (as opposed to narrow measures of the curriculum). However, the studies regarding the benchmark and quarterly assessments administered in Pennsylvania (Blanc et al, 2010; Christman, 2009; Olah, 2010; Goertz, 2009) did document that the instructional period was modified to flow with the administration of interim assessments; however, it was unclear if this material on the interim assessments clearly and exclusively tested content from that instructional sequence. However, 11 studies reported alignment to the curriculum – or mixed evidence of alignment, and only 4 failed to report evidence regarding alignment. The trends indicate that all studies developed in-house or in collaboration with a vendor were designed to be linked to the curriculum in some capacity. Although some interim assessments that were purchased directly from the vendor reported alignment, two studies reported clear that the curriculum and interim assessments were not aligned.

There was some variability in the use of interim assessments for student grades or in other ways to motivate students to perform at their best on them (k =6). Abrams' et al (2014) findings suggest that grades and other motivators are used outside of elementary schools to keep students motivated during the test. In addition, there was some variability in the presence of initial professional development. Half of the studies reported that staff received initial training

for the administration of interim assessments, while fewer (k =7) reported continued consultation to support interpreting and using findings during the school year for instructional purposes as well as for evaluative purposes. Lastly, there was little variability in response format: all interim assessments utilized at least multiple-choice responses. And all studies included mathematics interim assessments.

Characterizing Use

The third purpose was to (c) understand how interim assessments are used. Use was characterized in two ways. The first approach relied on Perie et al (2009) for classification. The second utilized the grounded approach to further describe interim assessment within instructional, predictive, and evaluative classifications. This information is represented in Tables 2, 3, and 4. They describe the five most frequent uses of interim assessments for each broad category, provide examples of use from a specific study, and indicate whether a study that provides relevant evidentiary insight has been conducted. For additional information regarding use, the Excel document used to create these categories is available as supplemental material or available upon request.

Instructional uses were the most frequent. They occurred within all studies included in this rapid review (k = 20). On average, a study included 8.6 uses (SD = 5.7). They broadly related to changes in instruction, identifying and proving remediation for students, grouping, and developing plans for support. Table 2 reports the five most frequent instructional claims for instructional uses. However, there were a total of 24 uses identified.

The most frequent instructional use were denoted by broad claims to modify or improve instruction. This category included statements lacking precision regarding an assessment used for an instructional purpose. For example, Clune and White (2008) administered a survey to teachers

and asked if they used interim assessments to modify instruction. The next claim that was most frequent (k= 13) was that assessments were used to identify which students needed additional support. The instructional uses were primarily teacher directed, and decisions were most often made based on the professional judgment of the user (k = 8), if the process for decision-making was reported at all (k = 7).

Evaluative uses were also quite common. They occurred within 16 of the 20 studies. On average, a study included 4.3 (SD = 4.4). This category included discrete uses that related to program and teacher evaluation, understanding and modifying the scope and sequence of the curriculum, monitoring student progress and growth, and resource allocation at the staff, school, and programmatic levels. Table 3 provides an overview of these findings. Unlike the instructional category, the evaluative category only contained four claims regarding use.

The most frequent was monitoring programmatic, teacher initiatives, and school performance for local or federal accountability (k = 15). For example, the sample reported in Underwood (2010) used interim assessments to understand how their school was performing in comparison to other schools. The least frequently identified use was using the data to monitor trends by subgroups, grades, or schools (k = 1).

Predictive uses were the least common (k = 8), and the least variable in nature. They occurred in only half of the included studies. Within study, predictive uses had an average occurrence of 0.55 (SD = 0.6) per study, which practically was only one within study predictive use at most. The findings for predictive uses are represented in Table 4. As can be observed, there was only one predictive use: forecasting student achievement on a summative assessment (k = 8).

Evidence Supporting Use

The final purpose of the present rapid review was to (d) gather evidence regarding the efficacy of identified uses of interim assessments. Konstantopoulos et al. (2016) was the only experimental study that investigated interim assessment use. The authors examined the effectiveness of two interim assessment systems compared to control using non-online based interim assessment systems in elementary and middle school. Specifically, the researchers used the interim assessments to improve TerraNova in grades K-2 and ISTEP+ in grades 3-8. Results were obtained using regression, and suggest null effects on ISTEP+ and negative effects on mClass for the treatment group. Stated limitations were lack of fidelity of treatment implementation, no modeling was done at the class level, model misspecification due to insufficient variables. As such, Konstantopoulos et al. (2016) findings suggest that the adoption of interim assessments do not lead to academic gains at the school level that are significantly greater than control schools without those interim assessments.

A similar finding was reported by Jones (2013). In this dissertation, Jones (2013), used path analysis to uncover patterns in the data structure relevant to interim assessment use. Analyses were restricted to 8th grade general education students who were native English speakers. Results were obtained using path analysis and suggest small negative relationships between number of benchmark tests and student outcomes on the end of year state summative assessment in both reading and math. As such, these findings suggest that the adoption of multiple interim assessment programs – greater than the average – was associated with lower achievement school-level achievement scores, on average.

Two other studies reported evidence regarding the psychometric properties of the assessments (Diaz-Bilello, 2011; Underwood, 2010). Diaz-Bilello (2011) Conducted a validation

study of a locally developed interim assessment. The dissertation was used to explore unidimensionality, local independence, item fit, relationship with the state test, diagnostic consistency, and use for teacher compensation. Evidence suggested support for unidimensionality, local independence, item fit, and predictive relationships with the state summative assessment. However, Diaz-Bilello (2011) found unconvincing evidence for instructional purposes. Specifically, assigning students to summer school based on their ability grouping, as when confidence intervals were taken into account, many students score could be in either performance category. Further unconvincing evidence was found for evaluative use – specifically, for informing teacher compensation; the scores on the interim assessment were discrepant with growth on the state summative assessment, suggesting differences in compensation based on the measure used – with less growth being observed on the interim measure.

Finally, Underwood (2010) examined the relationship between interim assessment scores and scores on the Florida summative assessment for students in 10th grade. Specifically, the researchers used correlational analyses to understand the relationship to outcomes and interviews with school principals from the seven high schools that were incorporated in the analysis to understand interim assessment use. Empirical results suggested a strong positive correlation between interim assessment and summative assessment scores (r = .74) in 10th grade. These findings suggest that interim assessments can be related to summative scores.

Discussion

The purpose of this study was to identify the body of literature classified under the term of interim assessment as well as characterize and evaluate the various uses of interim assessments. Specifically, this study was a rapid review of the interim assessments described in

20 studies, consisting of both published and grey literature. Findings highlight that although used broadly for instructional, evaluative, and predictive purposes, there are a variety of more discrete uses of interim assessments, particularly for instructional use. Unfortunately, current findings did not reveal evidence supporting the use of interim assessments for instructional decisions, as findings were null or negative (Diaz-Bilello, 2010, Jones, 2013; Konstantopoulos et al., 2016). However, findings did support basic validity evidence of the interim assessments in the form of predictive validity and unidimensionality (Diaz-Bilello, 2010).

The findings have led to several lessons. To begin, describing use is fairly common in the research. The majority of studies were (or included) qualitative or survey methods to characterize how schools utilized interim assessment data. Shepard (2010) called for more research investigating the use of interim assessments. It appears that her and the field's desire for additional research led to many valuable findings about interim assessment use in schools. Fascinatingly enough, however, about three-quarters of the way through the coding process, new discrete uses were not observed. They could primarily be subsumed under an existing use. This could mean that there is only so much variation in how interim assessments are used within predictive, evaluative, and instructional purposes.

Despite the exploration of use, there is still some ambiguity around interim assessment use. As observed in the instructional and evaluative categories, there can a lack of precision in the claims made about use. Sometimes it was difficult to classify a use as instructional or evaluative because the intention was unclear (e.g., understand student performance). For example, Konstantopoulos et al. (2016) investigated use; however, it was a broad application of use. Specific uses were not examined at the experimental level. Only the presence of specific interim assessments were manipulated.

This reflected a somewhat common theme in the literature, including the qualitative research: interim assessments can be used for change. However, the mechanisms for how that change occurs are fairly unclear. The qualitative research tried to understand what those mechanisms were by asking users. The experimental study investigated change. However, there is no unifying theory unique to interim assessments that suggest why and how the presence leads to improvements at the district, programmatic, teacher, or student levels. The only link that has been somewhat investigated are the correlational analyses conducted by Diaz-Bilello (2011) and Underwood (2010), which suggested associations between summative and interim assessment scores.

In addition, using interim assessments for instructional purposes is the most common in the research, yet few studies have investigated the efficacy of using these programs. Adopting interim assessments did not lead to improvement in academic achievement when compared to schools already engaged in assessment practices with different measures (Konstantopoulos, et al., 2016). In addition, Jones (2013) found that having more interim assessments than average is not associated with better student outcomes. These findings suggest that the interim assessment alone is not the intervention. It likely can be used to facilitate other practices, but the mechanisms that are needed to incur treatment validity are still unknown and unresearched.

Limitations and Future Direction

As revealed by the searches of databases and requesting grey literature, there is a vast body of research on interim assessments still ready for synthesis. The purpose of this rapid review was to identify that body of literature and begin the process of classifying use and summarizing evidence. However, rapid reviews of the literature have their own unique strengths and weakness, as do other methods (Grant & Booth 2009). In the context of the current review,

the comprehensiveness of the review is its most notable limitation. In other words, it is still unclear the extent to which findings are generalizable to the larger context of interim assessment research. It will be important to contextualize these findings within future reviews of interim assessment use. Despite this, the current review has the potential to have captured the majority of discrete interim assessment use. To draw from economics, Pareto's law of the vital few, could be used to suggest that the most important and frequent uses associated with effects, outcomes, and a lack-thereof, have already been captured by the current review. Regardless of whether this paper achieved saturation of interim assessment uses, future research is needed in this area.

Another limitation is that study quality was not directly investigated. This was done due to the nature of the review and research project; however, a minimal investigation of study quality may be gleaned from the findings. More specifically in the form of levels of evidence (e.g., Ackley, Swan, Ladwig, & Tucker, 2008). Some could argue that the designation between qualitative, non-experimental, quasi-experimental, and experimental methods is an indicator of evidence. The generalizability and level of experimental control is certainly variable within designs and is a reason to include comprehensive quality indicators in reviews; however, general differences based design are observable within the current rapid review. Future research may aim to understand the quality of studies, and if that is related to findings – particularly, clearly articulated claims regarding interim assessments.

A final limitation to note is that the current study did not code for non-use of interim assessments. It was evident in coding the qualitative that some teachers chose to administer the interim assessment because it mandated by their administration; however, they did not use the results. Likewise, some teachers were never trained on how to use the results to improve instruction, while some did not believe in the validity of the claims to improve instruction. Thus,

there were cases where interim assessments were intentionally not used. This primarily happened at the instructional level. Districts and schools, however, tended to have evaluative and predictive expectations for score use. Future studies may want to adopt and expand the current coding scheme to include this phenomenon.

One final future direction regards efficacy. Studies should investigate the relationship between use and efficacy, perhaps utilizing the greater body of research literature and conduct a review of reviews to understand what evidence exists to support or discourage the actions that are taken based on interpretations of interim assessments data.

References

Abrams, L. M., McMillan, J. H., Wetzel, A. P. (2015). Implementing benchmark testing for formative purposes: Teacher voices about what works. *Education Assessment, Evaluation, and Accountability, 27,* 347-375.

Ackley, B. J., Swan, B. A., Ladwig, G., & Tucker, S. (2008). *Evidence-based nursing care guidelines: Medical-surgical interventions.* (p. 7). St. Louis, MO: Mosby Elsevier.

Black, P. & William, D. (1998). Inside the black box: Raising standards through classroom assessment. *Phi Delta Kappa,* 1-13.

Blanc, S., Christman, J. B., Liu, R., Mitchell, C., Travers, E., & Bulkley, K. E. (2010). Learning to learn from data: Benchmarks and instructional communities. *Peabody Journal of Education, 85,* 205-225.

Brewer, S. M. (2014). *Computerized Benchmark Assessments: The Influence on Student Achievement Scores* (Doctoral dissertation). Retrieved from ProQuest Dissertations and Theses database. (UMI No. 3621651).

Brown, R., Steege, M. W., & Bickford, R.. (2014). Responsive assessment and instruction practices. *Academic Assessment and Intervention,* 117.

Bulkley, K. E., Christman, J. B., Goertz, M. E., & Lawrence, N. R. (2010). Building with benchmarks: The role of the district in Philadelphia's benchmark assessment system. *Peabody Journal of Education, 85*(2), 186-204.

Burch, P. (2010). The bigger picture: Institutional perspectives on interim assessment technologies. *Peabody Journal of Education, 85*(2), 147-162.

Card, N. A. (2015). *Applied meta-analysis for social science research*. Guilford Publications.

Chappuis, S. & Chappuis, J. (2007). The best value in formative assessment. *Educational Leadership, 65,* 14-19.

Cho, V., & Wayman, J. C. (2009, April). *Knowledge management and educational data use*. Paper presented at the Annual Meeting of the America Educational Research Association, San Diego, CA.

Christ, T. J., & Keller-Margulis, M. A. (2014). CBA: Curriculum-based Assessment. *Academic Assessment and Intervention*, 117.

Christman, J. B., Neild, R. C., Bulkley, K., Blanc, S., Liu, R., Mitchell, C., & Travers, E. (2009). Making the Most of Interim Assessment Data. Lessons from Philadelphia. *Research for action*.

Clune, W. H., & White, P. A. (2008). *Policy effectiveness of interim assessments in Providence Public Schools* (Working Paper No. 2008-10). Retrieved from http://www.wcer.wisc.edu/publications/workingPapers/ papers.php

Cook, D. A., Brydges, R., Ginsburg, S., & Hatala, R. (2015). A contemporary approach to validity arguments: A practical guide to Kane's framework. *Medical Education, 49*, 560-75.

Crane, E. W. (2010). Building an Interim Assessment System: A Workbook for School Districts. *Council of Chief State School Officers*.

Davidson, K. L. & Frohbieter, G. (2011). *District adoption and implementation of interim and benchmark assessments*. (CRESST Report 806). Los Angeles, CA: University of California, National Center for Research on Evaluation, Standards, and Student Testing (CRESST).

Diaz-Bilello, E. K. (2011). *A validity study of interim assessments in an urban school district* (Unpublished doctoral dissertation). University of Colorado at Boulder, CO.

Eagle, J. W., Dowd-Eagle, S. E., Snyder, A., & Holtzman, E. G. (2015). Implementing a multi-tiered system of support (MTSS): Collaboration between school psychologists and administrators to promote systems-level change. *Journal of Educational and Psychological Consultation*, *25*(2-3), 160-177.

Gersten, R., Chard, D., Jayanthi, M., Baker, S., Morphy, P., & Flojo, J. (2009). A Meta-analysis of Mathematics Instructional Interventions for Students with Learning Disabilities: A Technical Report. Los Alamitos, CA: Instructional Research Group.

Goertz, M. E., Olah, L. N., & Riggan, M. (2009). *From testing to teaching: The use of interim assessments in classroom instruction*. (Research Report No. 65). Retrieved from Consortium for Policy Research in Education website: https://www.cpre.org/testing-teaching-use-interim-assessments-classroom-instruction-0

Grant, M. J., & Booth, A. (2009). A typology of reviews: an analysis of 14 review types and associated methodologies. *Health Information & Libraries Journal*, *26*(2), 91-108.

Halverson, R. (2010). School formative feedback systems. *Peabody Journal of Education*, *85*(2), 130-146.

Harris, C., Hedges, L., & Valentine, J. (2009). Handbook of research synthesis and meta-analysis. *Russell Sage Foundation, New York.*

Herman, J. L, & Choi, K. (2008). *Formative assessment and the improvement of middle school science learning: The role of teacher accuracy* (Research Report No. 740). Retrieved from Center for Research on Evaluation Standards and Student Testing Website: http://cresst.org/wp-content/uploads/R740.pdf

Herman, J. O. A. N. (2017). Interim assessments in brief. *Oakland, CA.*

Hintze, J. M., & Silberglitt, B. (2005). A longitudinal examination of the diagnostic accuracy and predictive validity of R-CBM and high-stakes testing. *School Psychology Review*, *34*(3), 372.

Hook, N. (2015). Grounded theory. In *Game Research Methods* (pp. 309-320). ETC Press.

Jones, K. D. (2013). *The myth of benchmark testing: Isomorphic practices in Texas public school districts' use of benchmark testing* (Doctoral dissertation). Retrieved from Texas State University Library.

Jung, P. G., McMaster, K. L., Kunkel, A. K., Shin, J., & Stecker, P. M. (2018). Effects of Data-Based Individualization for Students with Intensive Learning Needs: A Meta-Analysis. *Learning Disabilities Research & Practice*, *33*(3), 144-155.

Kane, M. T . (2006). Validation. In R. L . Brennan (Ed.), *Educational measurement* (4th ed., pp. 17–64). Westport, CT: American Council on Education & Praeger.

Kingston, N., & Nash, B. (2011). Formative assessment: A meta-analysis and a call for research. *Educational measurement: Issues and practice*, *30*(4), 28-37.

Konstantopoulos, S., Miller, S. R., van der Ploeg, A., & Li, W. (2016). Effects of interim assessments on student achievement: Evidence from a large-scale experiment. *Journal of Research on Educational Effectiveness*, *9*, 188-208.

Kulp, C. I. (2017). *An embedded case study of the implementation of a school division's benchmark assessment system* (Doctoral dissertation). Retrieved from William and Mary Libraries.

Lai, E. R. (2009). *Interim assessment use in Iowa elementary schools* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses. (UMI Number: 3390174).

Lange, T. M. (2014). *Interim assessment data: A case study on modifying instruction based on benchmark feedback.* (Doctoral dissertation). Retrieved from Liberty University Libraries.

Medford, R. S. (2014). *An analysis of teachers' classroom instructional activities based on NWEA measures of academic progress (MAP) data.* (Doctoral dissertation). Retrieved from Gardner-Webb University Libraries.

No Child Left Behind (2001). No child left behind act of 2001. *Publ. L*, 107-110.

Oláh, L. N., Lawrence, N. R., & Riggan, M. (2010). Learning to learn from benchmark assessment data: How teachers analyze results. *Peabody Journal of Education*, *85*(2), 226-245.

Perie, M., Marion, S., & Gong, B. (2009). Moving toward a comprehensive assessment system: A framework for considering interim assessments. *Educational Measurement: Issues and Practice*, *28*(3), 5-13.

Ross, M. (2012). *A phenomenological study of teacher and administrator experiences in the analysis and interpretation of student assessment data.* (Doctoral dissertation). Retrieved from ProQuest Dissertations & Theses.

Salvia, J., Ysseldyke, J., & Bolt, S. E. (2012). *Assessment: In special and inclusive education*. Cengage Learning.

Shepard, L. A. (2010). What the marketplace has brought us: Item-by-item teaching with little instructional insight. *Peabody Journal of Education*, *85*(2), 246-257.

Stormont, M., & Reinke, W. M. (2013). Implementing Tier 2 social behavioral interventions: Current issues, challenges, and promising approaches. *Journal of Applied School Psychology*, *29*(2), 121-125.

Troy, T. (2011). Comprehensive Assessment Systems: Purposes and Implementation. Research Watch. E&R Report No. 11.10. *Wake County Public School System*.

Underwood, M. (2010). *The relationship of 10th-grade district progress monitoring assessment scores to Florida comprehensive assessment test scores in reading and mathematics for 2008-2009* (Unpublished doctoral dissertation). University of Central Florida, FL.

van Geel, M., Visscher, A. J., & Teunis, B. (2017). School characteristics influencing the implementation of a data-based decision making intervention. *School effectiveness and school improvement*, *28*(3), 443-462.

Table 1. *Study features and descriptions*

	Study	Design	Unit	N Units	Assessment Domain	Grade Range	Study Purpose
1	Abrams et al (2014)	Qual	Teacher	67	Math, ELA, Science, SS	K-8	Understand how elementary and middle school teachers (a) think about and use interim assessments and to (b) understand the barriers and facilitators to effective use
2	Blanc et al (2010)	Qual	School	10	--	3-8	Understand how ten elementary school districts in Philadelphia utilized interim assessments
3	Bulkley et al (2010)	Qual	School	6	Math	3-8	Examine Philadelphia's use of benchmark assessments, expectations of district leaders, and the supports to facilitate effective use
4	Burch (2010)	Qual	District	--	--	--	Explore the dynamics between schools and private organizations in administering interim assessments for the purposes of increasing efficacy, compliance, and equity
5	Christman et al (2009)	Mixed	Teacher	--	--	3-8	Report on findings from Philadelphia's interim assessment use
6	Clune & White (2008)	Qual	School	22	Math, ELA	K-12	Provide a qualitative perspective of the implementation period of Providence Public School District (PPSD) in their usage of quarterly (interim) assessments that were developed for math and language arts at every grade from 2004 to 2007
7	Davidson & Frohbieter (2011)	Qual	School	NRP	Various	6-8	Investigate why interim assessments were adopted (intended uses) as well as their applications in middle 10school mathematics classrooms in districts that had previously implemented interim assessments
8	Diaz-Bilello (2011)	Quant	District	NRP	Math, LA, Science	3-8	Evaluate the validity evidence supporting the interpretation of interim assessments for evaluative, instructional, and predictive purposes in Denver Public Schools
9	Goertz (2009)	Qual	Teacher	48	Math	3-5	Explore how elementary school mathematics teachers use interim assessments to inform their instructional practices.
10	Halverson (2010)	Qual	School	1	Math	3-5	Describe how school teachers and principals approach data-driven decision-making
11	Jones (2013)	Quant	District	1	NRP	8	Examine the relationships between district characteristics, interim assessment use, and student outcomes on the Texas end of year summative assessment
12	Konstantopoulos et al. (2016)	Quant	Student	6249	LA, Math	K-8	Examine the effectiveness of two interim assessment systems compared to control using non-online based interim assessment systems in elementary and middle school

13	Kulp (2017)	Qual	Teacher	9	ELA, Math, Science, SS	K-5	Examine the relationships between district characteristics, interim assessment use, and student outcomes on the Texas end of year summative assessment
14	Lai (2009)	Qual	School	144	ELA, Math	K-5	Understand elementary school use of interim assessments
15	Lange (2014)	Qual	Teacher	9	NRP	9-12	Examine VA high school teachers' use of interim assessment for instructional and evaluative purposes
16	Medford (2014)	Quant	Teacher	11	ELA, Math	K-5	Examine the instructional utility of MAP for teachers at a rural school in NC that had been using MAP for (over) five years.
17	Olah et al (2010)	Qual	Teacher	25	Math	3-5	Understand how elementary school math teachers in three Philadelphia schools meeting AYP, analyze, plan instruction, and use benchmark data
18	Ross (2012)	Qual	Teacher	10	NRP	K-12	Understand the process of analyzing and using student assessment data
19	Shepard et al (2011)	Qual	Teacher	30	ELA, Math	6-8	Investigate specific examples of how middle school mathematics teachers use data from interim and benchmark assessments
20	Underwood (2009)	Mixed	School	7	ELA, Math	10	Examine the relationship between interim assessment scores and scores on the FCAT for students in 10th grade

ELA = English Language Arts. NRP = Not reported. SS = Social Studies

Table 2. *Five Most Frequent Instructional Uses Evident in the Research on Interim Assessments*

	Use Description	*k*	Example	(Quasi-) Experimental Study Conducted?
1.	Broad claim to modify or improve instruction	14	A total of 86% of teacher reported modifying instruction based on interim assessment results (Clune & White, 2008).	Yes
2.	Identify students to provide additional support	13	Results were used to identify students for supplemental instruction (e.g., software, working with volunteers, afterschool tutoring; Shepard et al [2011]).	No
3.	Identify what content to reteach	10	The administration hoped teachers would reteach with new strategies (Bulkley et al., 2010).	No
4.	Improve score on the summative assessment	10	Guide schoolwide improvement efforts to meet AYP (especially in low-performing schools; Bulkley et al., 2010).	Yes
5.	Differentiate instruction	9	Identifying students with similar patterns of performance on the assessment and using that to constructs groups to differentiate instruction (Blanc et al., 2010)	Yes

Note. Table created based on coding a total of 20 studies featuring interim assessments.

Table 3. *Evaluative Uses Evident in the Research on Interim Assessments*

	Use Description	*k*	Example	Empirical Study Conducted?
1.	Monitor programmatic initiatives and/or teacher and school performance for local or federal accountability	15	Evaluating effects of specific organizational changes or effects of recent curricular changes (Lai, 2009)	Yes
2.	Allocate resources to support programs, schools, staff	11	Results informed professional development selection and other resources (Davidson & Frohbieter, 2011)	No
3.	Understand effectiveness of scope and sequence to support modification for future years	7	Identified weaknesses in the current curricular scope and sequence (e.g., fractions, geometry, measurement; Clune & White, 2008)	No
4.	Use the data to monitor trends by subgroups, grades, or schools.	1	Principals looked for gradewide trends (Bulkley et al, 2010)	No

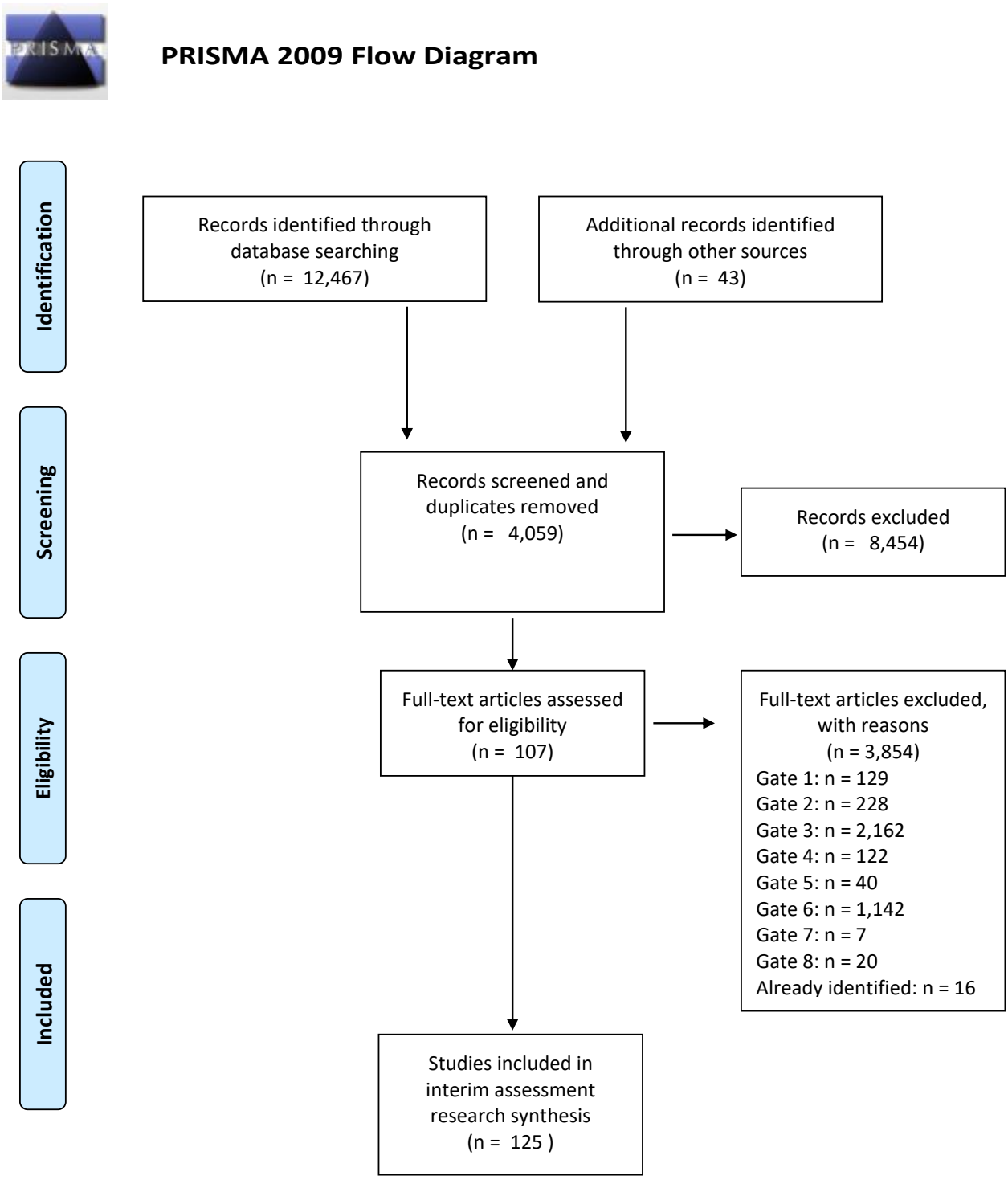
Note. Table created based on coding a total of 20 studies featuring interim assessments.

Table 4. *Predictive Uses Evident in the Research on Interim Assessments*

	Use Description	*k*	Example	Empirical Study Conducted?
1.	Forecast performance on end of year summative assessment	8	For low performing schools, benchmarks would be used to monitor and support the students who were close to reaching proficiency in order to meet AYP (Bulkley et al., 2010).	No

Note. Table created based on coding a total of 20 studies featuring interim assessments.

Figure 1



From: Moher D, Liberati A, Tetzlaff J, Altman DG, The PRISMA Group (2009). *P*referred *R*eporting *I*tems for *S*ystematic Reviews and *M*eta-*A*nalyses: The PRISMA Statement. PLoS Med 6(7): e1000097. doi:10.1371/journal.pmed1000097

For more information, visit www.prisma-statement.org.

Appendix A

June 28, 2019 broad search using ERIC and Academic Search Premiere via EBSCO host.

(Interim assessment OR Benchmark OR formative assessment OR predictive assessment OR standards-based assessment OR predictive validity OR instructional assessment OR evaluative OR unit test OR periodic OR quarterly) AND (assessment OR validity OR evaluation OR implementation) AND (classroom OR elementary school OR middle school OR junior high OR high school OR primary school OR secondary school OR teacher OR administrator OR district)

July 1, 2019 targeted search using ERIC and Academic Search Premiere via EBSCO host.

Interim Comprehensive Assessments OR Interim Assessment Blocks OR Smarter Balanced Interim Assessments OR Smarter Balance OR Northwest Evaluation Association Measure of Academic Progress OR NWEA MAP OR NWEA Measure of Academic Progress OR Northwest Evaluation Association MAP OR i-ready reading OR i-ready math OR i-ready diagnostic OR i-ready standards mastery OR i-ready instruction OR curriculum associates i-ready OR AIMSweb OR DIBELS OR Acadience reading OR Dynamic Indicators of Basic Early Literacy Skills OR formative assessment system for teachers OR fastbridge OR Imagination Station reading OR Imagination station math OR istation OR Phonological Awareness Literacy Screen* OR renaissance star early literacy OR star early literacy OR next generation science standards OR STAR reading OR STAR math OR STAR 360 OR terranova test OR terranova assessment OR leap 360 OR Brigance Early Childhood Screens OR State of Texas Assessments of Academic Readiness OR Texas STAAR OR Readiness Improvement Success Empowerment