IES Institute of Education Sciences

The Reliability and Consequential Validity of Two Teacher-Administered Student Mathematics Diagnostic Assessments

Regional Educational Laboratory Southeast

At Florida State University

REL 2020 039
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Evaluation and Regional Assistance at IES

The Reliability and Consequential Validity of Two Teacher-Administered Student Mathematics Diagnostic Assessments

Russell Gersten, Madhavi Jayanthi, Rebecca Newman Gonchar, Daniel Anderson, Samantha Spallone, and Mary Jo Taylor

September 2020

Several school districts in Georgia use two teacher-administered diagnostic assessments of student knowledge of mathematics as part of their multi-tiered system of support in grades K–8: the Global Strategy Stage (GloSS; New Zealand Ministry of Education, 2012) and the Individual Knowledge Assessment of Number (IKAN; New Zealand Ministry of Education, 2011), which comes in two formats (Counting Interview and Written Assessment). However, little is known about whether two teachers using the same assessment to assess the same student on two occasions within a short period of time assign the same Stage Score (interassessor reliability) or about how useful the teachers found the assessments (consequential validity).

Rather than relying on occasional testimonials from the field, decisions about using diagnostic assessments across the state should be based on psychometric data from an external source. Districts not currently using the GloSS and IKAN assessments have indicated an interest in using them, if they are proven to be reliable and valid diagnostic assessments, to assess students' understanding of mathematics and determine appropriate levels of instruction and intervention. This study found adequate interassessor reliability for the GloSS and for the IKAN Counting Interview but not for the IKAN Written Assessment. The IKAN Written Assessment requires additional attention to improve training so that reliability can be established. Teachers indicated that they found the screening data from the GloSS and IKAN assessments more useful for guiding decisions about student instruction and intervention than the screening data currently employed. Although teachers in the study's focus groups expressed strong support for both assessments, teachers reported in the study's survey that the GloSS is more useful than the IKAN because it assesses students' solution strategies, unlike most other mathematics assessments. Teachers also expressed some criticisms of both assessments; for example, they believed that the GloSS should include vocabulary familiar to students and that the IKAN Written Assessment should be untimed.

Why this study?

There is growing interest in using diagnostic assessments to assess students' mathematics understanding in order to determine appropriate content for instructional intervention (for example, Confrey, 2008; Ketterlin-Geller et al., 2019). Though several school districts in Georgia use two teacher-administered diagnostic assessments of student knowledge as part of their multi-tiered system of support in grades K–8—the Global Strategy Stage (GloSS; New Zealand Ministry of Education, 2012) and the Individual Knowledge Assessment of Number (IKAN; New Zealand Ministry of Education, 2011)—little is known about the reliability and validity of these assessments. Reliability and validity data on the assessments would provide the Georgia Department of Education with the confidence that students are being accurately identified for services, meaning students who need services are identified for services and students who do not need services are not.

For additional information, including the study instruments, methods, and supplementary analyses, access the report appendixes at https://go.usa.gov/xG4GW.

When used together, the GloSS and IKAN assessments furnish diagnostic information that teachers can employ to determine which students need intervention

and to address student strengths and deficits. The GloSS provides information on the strategies students use when solving mathematics problems. The IKAN is available in two formats, the IKAN Counting Interview, which is for students performing at lower levels on the GloSS, and the IKAN Written Assessment, which is for students performing at higher levels on the GloSS (see box 1 for definitions of key terms). Both formats provide information on students' number knowledge (magnitude comparisons, knowledge of the base 10 system, and meaning of decimals and fractions).

Although leadership at the Georgia Department of Education and at the regional and district levels supports using the GloSS and IKAN and views them as valuable for successful implementation of a multi-tiered system of support, there are no Georgia-specific data to back their use. The only psychometric evidence available on the

Box 1. Key terms

Consequential validity. Evidence of the social consequences that result from using a particular assessment, such as the benefits and unintended impacts of using the assessment (Messick, 1988, 1994) and the extent to which the assessment improves or hinders instructional practice (Behavioral Research and Teaching, 2017; Gersten et al., 1995; Shepard, 1997). In this study, teachers' perceptions of the benefits and unintended impacts (such as loss of instructional time) of using assessments are an initial gauge of consequential validity.

Fidelity of administration. The extent to which the people administering an assessment adhere to the protocol for administering it. This study used the number of protocol steps that teachers performed correctly when administering the assessments to determine fidelity of administration.

Focus group. Groups of people assembled to participate in a carefully planned discussion on a particular topic (Krueger & Casey, 2000). Participants share thoughts and feelings that cannot be gathered through a typical survey (Jayanthi & Nelson, 2002; Vaughn et al., 1996) or through a more formal interview (Gall et al., 2007). Three focus groups were held as part of this study.

Global Strategy Stage (GloSS). An assessment that provides information on the strategies students use when solving mathematics problems. The GloSS assessment is administered using a one-on-one interview format and takes 5–20 minutes. A single form of the assessment was used in this study. The GloSS identifies students' stages in three content area domains (addition and subtraction, multiplication and division, and ratios and proportions) and assigns students a Stage Score from 0 (one-to-one counting) to 8 (advanced proportional reasoning), according to the strategies they use to solve problems. Teachers sum the student's responses and match the sum to the ranges for each Stage Score in order to interpret the results.

Individual Knowledge Assessment of Number (IKAN). An assessment that provides teachers with information on students' Number Knowledge Stages across five increasingly abstract domains of arithmetic, including simple whole-number sequencing, multidigit operations, and operations with decimals and fractions. The IKAN is available in two formats: the Counting Interview and the Written Assessment. Students whose GloSS Stage Score is 0–3 receive the IKAN Counting Interview, which takes 5–15 minutes to administer. A single form of the assessment was used in this study. Students whose GloSS Stage Score is 4–8 receive the IKAN Written Assessment, a timed six-minute test that includes 40 items, divided across five Stage Scores (4–8). Students are required to respond to each item within the allocated time at a rapid pace, and responses to each item are scored as correct or incorrect. Teachers sum the student's responses and match the sum to the ranges for each Stage Score in order to interpret the results. Students receive a Stage Score based on the last stage in which they answered all items correctly. Teachers in this study all used Form 1 of the Written Assessment.

Interassessor reliability. Percentage of exact agreement or close (plus or minus one) agreement between two assessors who use the same assessment to assess the same student on two occasions within a short period of time. Exact agreement means that the two assessors assign the same Stage Score for the same student. Plus-or-minus agreement means that the two assessors assign either the same Stage Score or a Stage Score one above or one below the Stage Score assigned by the other. The study team decided before the study began to use plus-or-minus agreement to determine interassessor reliability for GloSS and exact agreement for IKAN (see appendix B for details).

GloSS and IKAN is suggestive and was collected by the New Zealand Ministry of Education (see appendix A). To provide state-specific data for Georgia, the Regional Educational Laboratory Southeast's Improving Mathematics Research Alliance conducted this external, exploratory study of interassessor reliability and consequential validity in three school districts (Fannin, Jefferson, and Walker County School Districts) in the state. Interassessor reliability indicates whether two teachers (assessors) administering the assessment on different occasions to the same student give the student the same Stage Score. If a student's Stage Score on the assessment varies by assessor, the assessment would not be a reliable means of determining gaps in a student's mathematics understanding or of identifying the kinds of interventions needed to help the student improve. In this study, consequential validity indicates whether teachers found the results of the assessments useful and what problems, if any, the teachers encountered in using the assessments.

Research questions

Four research questions guided this exploratory study of the reliability and validity of the GloSS and IKAN diagnostic assessments in Georgia:

1. What is the interassessor reliability of the GloSS diagnostic assessment when administered by two assessors to the same student within a one-week period?

2. What is the interassessor reliability of the IKAN diagnostic assessment when administered by two assessors to the same student within a one-week period?

3. What are teachers' perceptions of the usefulness of the GloSS and IKAN assessments for guiding intervention, an initial gauge of the assessments' consequential validity?

4. What was the teachers' level of fidelity in administering the assessments?

The analyses for research questions 1 and 2 identified the percentage agreement between two assessors testing the same student using the same assessment within a one-week period. Both assessments require that teachers sum the student's responses and match the sum to the ranges for each Stage Score in order to interpret the results.

The GloSS interassessor reliability calculation used the plus-or-minus-one agreement method, in which the two people assessing the same student within a one-week period must assign either the same Stage Score or a Stage Score one above or one below the other in order to be considered in agreement. For example, if one assessor assigned the student Stage 4 and another assessor assigned Stage 5, these assessors would be considered to be in agreement. This is appropriate because in order to score items on the GloSS, teachers must make complex inferences about the strategy a student used to solve a problem rather than scoring an answer as correct or incorrect. Interassessor reliability above 90 percent is considered adequate when using the plus-or-minus-one agreement method (Stemler, 2004). That means that the assessment would be considered to have adequate interassessor reliability if 9 out of 10 times two assessors assessing the same student assign the same Stage Score or a Stage Score one above or one below the other.

In contrast, the IKAN scoring process for both the Counting Interview and the Written Assessment is more straightforward. Teachers mark a student's answers as either correct or incorrect, and the student must answer all problems at each developmental stage correctly to pass a particular stage. Because little interpretation is needed to accurately determine what students know, and thus their score, the exact agreement method was used for the reliability calculation. In this method, two assessors assessing the same student within a one-week period must assign the same Stage Score. Reliability above 70 percent is considered adequate for diagnostic assessments using the exact agreement method (Stemler, 2004) because this is not a high-stakes assessment, merely a tool

to help guide intervention. This means that the assessment would be considered to have adequate interassessor reliability if 7 out of 10 times two assessors assessing the same student assign the same Stage Score.

The analyses for research question 3 used data from a teacher survey and three focus groups (two in Jefferson County School District and one in Fannin County School District). The analysis of the survey responses was descriptive, noting the percentage of teachers who responded in each category of a Likert scale for each survey item. For the in-person focus groups the study team analyzed transcripts to identify themes. Issues explored included the adequacy of the one-day training on administering the assessments that was provided for teachers, the usefulness of the assessments, and any problems in administering the assessments. Data from the survey and the focus groups were used together to understand teachers' perceptions about the assessments.

To strengthen confidence in the study's evidence of reliability and validity, data were also collected to determine whether teachers followed protocols when administering the assessments (fidelity of administration). For research question 4 the number of protocol steps that teachers performed correctly when they administered the assessments was used to determine fidelity of administration. (The study's data sources, sample, and methods are summarized in box 2 and discussed in detail in appendix B.)

Box 2. Data sources, sample, and methods

Data sources. The study used data from three sources:

- Teacher scores for 60 students on two administrations of the Global Strategy Stage (GloSS) and the Individual Knowledge Assessment of Number (IKAN) assessments (in that sequence, as intended by the developers) within a one-week period in March 2019 (see appendix A for assessment descriptions and figures A1 and A2 for sample items from each assessment). For the GloSS, teachers assigned each student a GloSS Stage Score based on the strategies the student applied to solve problems using the four basic arithmetic operations (addition, subtraction, multiplication, and division), as well as problems involving ratios and proportions. Stage Scores range from Stage 0 (one-to-one counting) to Stage 8 (advanced proportional reasoning). On the IKAN, teachers score each student's answer for each item as correct or incorrect. The teacher converts the number of correctly answered items to a predetermined Stage Score.
- Teacher perceptions of the usefulness of the GloSS and IKAN diagnostic assessments collected through an online survey conducted in April 2019.
- Teacher perceptions of the GloSS and IKAN assessments collected during the focus groups (two in Jefferson County and one in Fannin County) conducted in May 2019.

Sample. The sample included 30 grade 1 and grade 3 teachers (6 on special assignment as mathematics instructional coaches) and 60 of their grade 1 and 3 students (32 in grade 1 and 28 in grade 3) from Fannin, Jefferson, and Walker County School Districts. Though the assessments are used across grades K–8, this study focused on grades 1 and 3 because the Georgia Department of Education is especially interested in early numeracy, and students in grades 1 and 3 could provide a range of early numeracy outcomes.

Methods. The exploratory study was conducted during the 2018/19 school year. One instructor facilitated a one-day training for teachers at two locations in the state. The training included the purpose of the assessments, the stages they cover, administration and scoring procedures, and video examples of teachers administering and scoring the assessments. Although the training was the same for all participants, it was a refresher for some participants and an introduction to the assessments for others.

To calculate the interassessor reliability of the GloSS and IKAN assessments (research questions 1 and 2), two different teachers tested the same students on two occasions within a one-week period using both the GloSS and IKAN assessments. Interassessor reliability for the GloSS and IKAN was calculated using percentage agreement. Agreement was defined as two scores within one point of each other ("plus or minus one") to calculate interassessor reliability for the GloSS because scoring the assessment requires making sophisticated judgments. Exact agreement was used to calculate interassessor reliability for the IKAN, as scoring is straightforward (correct or incorrect) and requires no inferences about student performance.

To determine the consequential validity of the GloSS and IKAN assessments (research question 3), teachers who administered the assessments were surveyed about their perceptions of the usefulness of the GloSS and IKAN diagnostic assessments

for informing instruction. All 30 teachers who participated in this study completed the survey. The survey included 22 items: 21 Likert scale items and one general open-ended question. The Likert scale items asked teachers how much they agree or disagree with a variety of statements about the usefulness of the assessments. Because the GloSS and IKAN assessments make up a single diagnostic system, some survey items were posed about the system as a whole and others were posed for the GloSS and IKAN assessments separately. The study team also conducted three focus groups with the teachers to discuss how they used the GloSS and IKAN assessments in practice. Survey responses were tabulated, and focus group transcriptions were coded.

Over a two-week period in March 2019 study team members conducted site visits in two districts to assess fidelity of assessment administration in a randomly selected 25 percent of GloSS and IKAN administration sessions. Two observers used checklists to assess the accuracy of teachers' assessment administration procedures. Fidelity of assessment administration was calculated as the percentage of procedures on the checklist that were implemented correctly, excluding procedures marked as not applicable. The study team noted which items were not implemented with fidelity.

See appendix B for more details on the sample and methods.

Findings

This section first presents the results for interassessor reliability and fidelity of assessment administration, followed by the results for consequential validity based on the teacher survey and focus groups. Demographic data are summarized in appendix B.

Interassessor reliability was adequate for the Global Strategy Stage assessment

The GloSS was scored with adequate interassessor reliability. Interassessor reliability for the GloSS was 92 percent across grades 1 and 3 when calculated using the plus-or-minus-one agreement method (table 1). It was 91 percent for grade 1 students and 93 percent for grade 3 students. In other words, assessors assigned the same Stage Score or a Stage Score one above or one below the other's Stage Score for 91 percent of the grade 1 assessment administrations and 93 percent of the grade 3 assessment administrations.

Table 1. Interassessor reliability was 92 percent for the Global Strategy Stage for grades 1 and 3 combined, 71 percent for the Individual Knowledge Assessment of Number (IKAN) Counting Interview, and 58 percent for the IKAN Written Assessment, 2019

Assessment	Number of students	Percent agreement	
		Exact agreement	Plus or-minus one agreement
Global Strategy Stage (GloSS)			
Grade 1	32	na	91
Grade 3	28	na	93
Grades 1 and 3	60	na	92
Individual Knowledge Assessment of Number (IKAN)			
Counting Interview			
Grade 1[a]	35	71	na
Written Assessment			
Grade 3[b]	24	58	na

na is not applicable.

a. Three grade 3 students are included in this calculation because their mathematics knowledge was not high enough to take the Written Assessment.

b. Three grade 3 students with low GloSS scores took the Counting Interview and not the Written Assessment. One grade 3 student had incomplete IKAN reliability data because the first teacher administered the Counting Interview, and the second teacher administered the Written Assessment.

Source: Authors' analysis of primary data collected for the study in 2019.

The GloSS was also administered as intended, meaning that the assessors administered the assessment according to the recommended protocols. This finding lends confidence to the interassessor reliability findings because it is likely that other teachers who administer the assessment as intended will reach the same reliability level. Fidelity of the GloSS administration was calculated for 25 percent of the total 120 assessment administrations (30 students were administered the assessment twice, once by each assessor). On average, teachers correctly completed 90 percent of the 12 steps required to accurately administer the GloSS (median of 92 percent, with a range of 64–100 percent). The item on the GloSS that had the lowest fidelity of administration (67 percent) was applying the decision rules correctly to determine whether to continue with the next set of more difficult addition and subtraction problems (see table C3 in appendix C).

Interassessor reliability was adequate for the Individual Knowledge Assessment of Number (IKAN) Counting Interview but not for the IKAN Written Assessment

Interassessor reliability was 71 percent for the IKAN Counting Interview, which is adequate, and 58 percent for the IKAN Written Assessment, which is not adequate (see table 1). That means that the assessors assigned the same Stage Score for the same students 71 percent of the time for the Counting Interview (which is intended for students performing at lower levels on the GloSS, mostly grade 1 students in this study) and 58 percent of the time for the Written Assessment (which is intended for students performing at higher levels on the GloSS, mostly grade 3 students in this study).

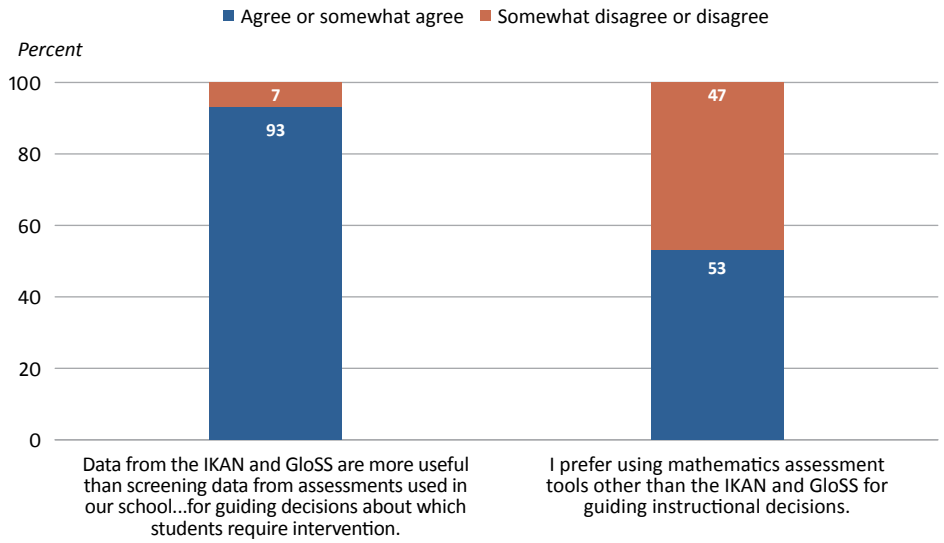
Fidelity of administration was also higher for the IKAN Counting Interview than for the IKAN Written Assessment. Fidelity was, on average, 96 percent across the 14 items for the Counting Interview (median of 100 percent, with a range of 64–100 percent) and 85 percent across the 7 items for the Written Assessment (median of 100 percent, with a range of 43–100 percent; see table C4 in appendix C). Confidence in the accuracy of the assessment administration and thus in the reliability of the results was higher for the Counting Interview than for the Written Assessment. For the Counting Interview fidelity was assessed on 26 percent of the total 70 assessment administrations (35 students were administered the assessment twice, once by each assessor). For the Written Assessment, fidelity was assessed on 25 percent of the total 48 assessment administrations (24 students were administered the assessment twice, once by each assessor). The item on the IKAN Written Assessment with the lowest fidelity (67 percent) was identifying the Stage Score for the fractions domain (see table C3 in appendix C).

Although the IKAN Written Assessment was administered according to the protocol, with 85 percent fidelity, the two assessors achieved exact agreement just 58 percent of the time. That means that although the teachers were able to follow directions in administering the assessment, they did not assign the same Stage Score for the same student 42 percent of the time when they scored the students' responses. This finding indicates a problem with scoring this assessment.

The consequential validity of the assessments was mixed—teachers considered the data useful but also identified several concerns with training, administration, and scoring

About 93 percent of teachers indicated on the survey that they found the data from the GloSS and IKAN assessments more useful than data from other school assessments for screening students and determining which students require intervention, but only 47 percent preferred GloSS and IKAN for instructional decisions (figure 1). One teacher in the focus group commented that the assessment results are not just "filed away" but are used to better understand a student's mathematics level and are especially useful at the beginning of the year, when teachers are still getting to know their students. (The percentage of responses for each question on the survey and a list of the themes from the focus group transcripts are in tables C1 and C2 in appendix C.)

Figure 1. Teachers reported that the Global Strategy Stage (GloSS) and the Individual Knowledge Assessment of Number (IKAN) assessments were useful for screening students but less useful in guiding instructional decisions, 2019

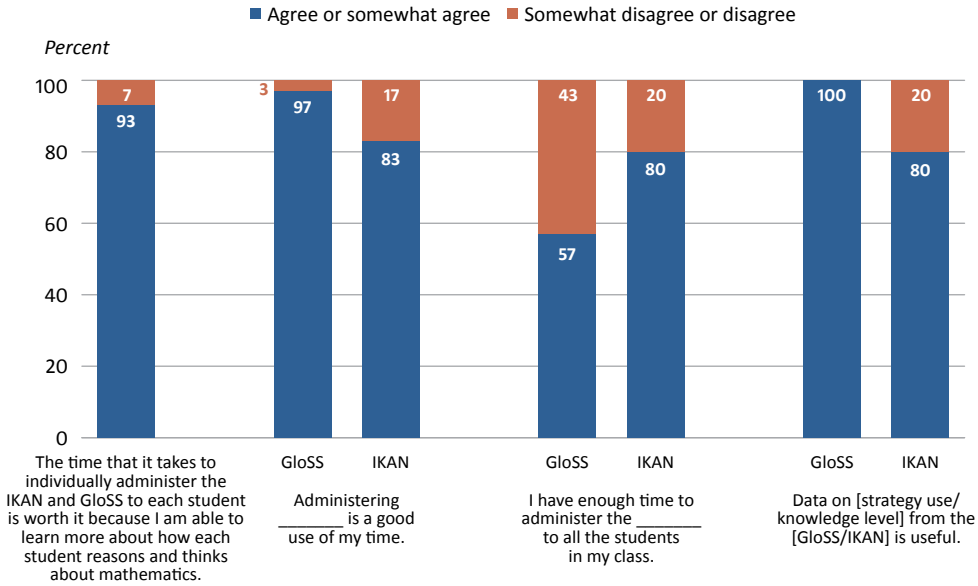


Note: *n* = 30.

Source: Authors' analysis of primary data collected for the study in 2019.

Overall, 93 percent of teachers noted on the survey that the time spent administering the two assessments was worth it because they were able to learn more about how students reason and think about mathematics (figure 2). One teacher commented in the open-response survey item that "Administration of GloSS and IKAN gives a teacher valuable information on a student's level of strategy usage and level of fluency with numeracy." Another teacher stated in the survey, "Both tests gave me valuable insight as to how my students are seeing and reasoning

Figure 2. Teachers indicated that administering the Global Strategy Stage (GloSS) and Individual Knowledge Assessment of Number (IKAN) assessments was a good use of their time, 2019



Note: *n* = 30.

Source: Authors' analysis of primary data collected for the study in 2019.

mathematics, as well as number knowledge level." However, only 57 percent of teachers reported on the survey that they felt that they had sufficient time to administer the GloSS individually to all their students, and teachers in the focus groups also expressed reservations about the time it took to administer the GloSS as a one-on-one interview. Despite these misgivings, however, most teachers—97 percent for the GloSS and 83 percent for the IKAN—reported on the survey that they felt that administering the GloSS and IKAN was a good use of their time.

Teachers in the focus groups also reported that the assessments enabled them to see how their students were thinking, which helped the teachers determine what instructional steps to take next and how to group their students for instruction. Some teachers in the focus groups noted that they used the data to differentiate instruction or to develop interventions to help their students advance. One teacher in the focus group remarked, "I can see what strategies they are using so I know where to take them next." Another added, "When my students took the IKAN Counting Interview and had trouble with counting, I created the hundreds puzzle as an intervention by taking a hundreds chart and cutting it up so that the students could tell me what number comes before and what number comes after."

Teachers in the focus groups rated the GloSS and the IKAN Counting Interview as equally useful and the IKAN Written Assessment as less useful. One teacher referred to scoring the IKAN Written Assessment as "an exercise in guessing" and remarked that scoring it was not a good use of time because the information gleaned about what her students know and do not know is inadequate. Teachers in the focus groups also commented that the pacing of the timed IKAN Written Assessment was too fast and that students became frustrated by trying to write their answers within the allotted time. Teachers thought that the assessment would be improved by removing the timing element. Most teachers in the focus group expressed support for the GloSS, but some teachers voiced complaints about the pictures and vocabulary used; they felt that the pictures were misleading and that some of the vocabulary in the questions was unfamiliar to students and confused them (for example, "lamington" and "lollies").[1]

Teachers reported in the survey that they found the GloSS to be more useful than the IKAN for a range of purposes: for identifying skills and concepts in which students are weak (GloSS: 87 percent; IKAN: 73 percent); determining placement within a multi-tiered system of support (GloSS: 70 percent; IKAN: 57 percent); and modifying mathematics instruction (GloSS: 77 percent; IKAN: 67 percent; figure 3).

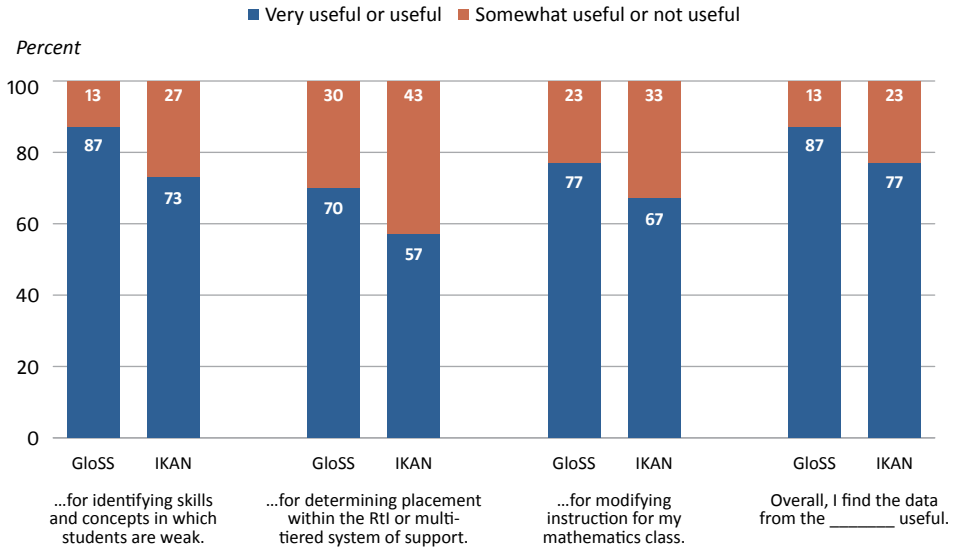
In the focus groups, teachers reported that they used data from the assessments for the following purposes:

- Identifying which students are struggling with or are gifted in mathematics.
- Identifying specific needs and deficits of students who are struggling and gaps in student learning, even for gifted students.
- Determining the most appropriate intervention for students.
- Grouping students for instruction.
- Monitoring student progress.
- Understanding students' thinking and strategy use.
- Adjusting instruction to students' needs.

A teacher in the focus group commented that she used the assessments once a month to monitor progress; another used them at key points to determine whether students should remain in the same tier of instruction or move to another. One teacher voiced the feelings of many others: "Basically, I take the data and I form the small group. From the small group, I taught the skill that they were lacking on from the data I gathered from the test." Another teacher in the focus group said she used the assessments to monitor students' progress in the intervention: "When I gave it this second time, they showed a lot of growth." In two of the three focus groups, teachers mentioned using the data to identify not just students who were struggling but also areas in need of improvement

1. The vocabulary reflects that the assessment was created by New Zealand's Ministry of Education, and the version that was updated to American English vocabulary was not used for this study.

Figure 3. Teachers found the Global Strategy Stage (GloSS) assessment to be more useful than the Individual Knowledge Assessment of Number (IKAN) assessments, 2019



RtI is response to intervention.

Note: *n* = 30.

Source: Authors' analysis of primary data collected for the study in 2019.

for gifted or high-achieving students. One teacher remarked, "For those students, it is still important to identify how they are thinking so that they can be accelerated, while still recognizing that they too sometimes have underlying misconceptions that need to be addressed."

About 90 percent of teachers indicated on the survey that the assessment training was adequate. In contrast, some teachers in the focus groups said that the training was inadequate while others mentioned that the training was of high quality but minimal, particularly for teachers who were new to the assessments. Some teachers indicated that additional practice with assessment administration and scoring would be beneficial. One teacher in the focus group commented, "Since I had never used these assessments before, I found the one training insufficient to make me confident in administering and scoring them even though the training was well done." Another teacher in the focus group observed: "The training in February was helpful for me because I have given these assessments for several years. However, I do not think it is enough training for new teachers." Five teachers also commented in the open-response survey item that they needed more training in scoring the assessments and using the data to inform instruction.

Several teachers in the focus groups also expressed frustration with efforts to link students' strengths and deficits to specific interventions. Many teachers in the focus groups mentioned having difficulty using the assessment developer's website for support in linking the data to an intervention. One teacher remarked that she spent a lot of time "navigating the [assessment developer's] website" but she "didn't know where to go; how to go through it…to find just the right intervention."

Implications

Because the GloSS and IKAN Counting Interview assessments were shown to have adequate interassessor reliability and consequential validity (teachers gained insights into their students' mathematical thinking and found the assessments to be useful in guiding decisions about interventions), the study findings can inform the Georgia Department of Education's recommendation to other districts on using the assessments for diagnostic purposes. However, several

factors limit the generalizability of the findings even though student scores covered the range of possible stages and the sample was diverse and was similar in mathematics proficiency and family income to the overall population of grade 1 and 3 students in Georgia (see appendix B). The study included only three school districts and only two grade levels. In addition, the racial/ethnic background of students in the study is not representative of that of the state as a whole; the percentage of White students was twice that of the state average in two of the three districts. Therefore, caution should be taken when generalizing findings to other grade levels and other student populations.

The Georgia Department of Education might wish to consider some options for collecting additional information to support the selection and use of these assessments in Georgia. Given the current study's findings about the IKAN Written Assessment, the Georgia Department of Education might consider several steps to validate its ongoing use in the state. It might want to map the scoring of both assessments to the Georgia or Common Core Standards; conduct an additional evaluation of the IKAN Written Assessment using the version of the assessment updated in American English and perhaps using more contemporary psychometrics, such as Item Response Theory; and examine whether the use of these diagnostic assessments is associated with improvements in grade 3 performance on state tests. Mapping of scores could help teachers understand how to use the data to address key mathematical ideas in the state standards. An evaluation of the IKAN Written Assessment could address whether the ordering of items and stages is consistent with curricula and standards in Georgia. The introduction of topics or skills may occur in a different order in New Zealand's curriculum than in Georgia's curriculum. Studying whether the use of these diagnostic assessments is associated with improvements in grade 3 performance could inform the adoption of the IKAN Written Assessment for use as a diagnostic assessment in other districts in Georgia.

Although this study was exploratory, the Georgia Department of Education might want to use the information gathered to improve teacher training in the use of the GloSS and IKAN assessments. For example, more training (initial, follow-up, online, or ongoing) could improve teachers' accuracy in interpreting the strategy levels on the GloSS and in administering and scoring the IKAN Written Assessment. A two-day training (rather than the one-day training offered in this study) assessment that includes instruction on scoring the assessments as well as practice administering them might improve interassessor reliability. Teachers need training in applying the rules for scoring the IKAN Written Assessment consistently. The training might also include time for teachers to practice matching student responses to Strategy Stages on the GloSS. The Georgia Department of Education might want to devote some professional development time to help teachers understand the rationale for the assessments and how teachers might use the assessment results to inform instruction, a desire expressed by teachers in the focus groups. Finally, the department might consider providing guidance for districts and schools in scheduling these assessments to reduce the amount of instructional time that is lost.

References

Behavioral Research and Teaching. (2017). *Consequential validity survey results for the Oregon Extended Assessments*. University of Oregon.

Confrey, J. (2008). Teaching teachers to use data to inform issues of equity and instruction. In P. Ernest (Ed.), *Philosophy of mathematics education journal: Special issue on justice* (n.p.). Cambridge University Press.

Gall, M., Gall, J. P., & Borg, W. R. (2007). *Educational research: An introduction* (8th ed.). Pearson.

Gersten, R., Keating, T. J., & Irvin, L. K. (1995). The burden of proof: Validity as improvement of instructional practice. *Exceptional Children, 61*(6), 510–519. https://doi.org/10.1177/001440299506100602.

Jayanthi, M., & Nelson, J. S. (2002). *Savvy decision making: An administrator's guide to using focus groups in schools*. Corwin Press.

Ketterlin-Geller, L. R., Shivraj, P., Basaraba, D., & Yovanoff, P. (2019). Considerations for using mathematical learning progressions to design diagnostic assessments. *Measurement: Interdisciplinary Research and Perspectives*, *17*(1), 1–22. https://eric.ed.gov/?id=EJ1206894.

Krueger, R., & Casey, M. A. (2000). *Focus groups: A practical guide for applied research* (3rd ed.). Sage Publications.

Messick, S. (1988). The once and future issues of validity: Assessing the meaning and consequences of measurement. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 33–46). Lawrence Erlbaum.

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher, 23*(2), 13–23. https://doi.org/10.3102/0013189X023002013.

New Zealand Ministry of Education. (2011). *Individual Knowledge Assessment for Numeracy (IKAN).*

New Zealand Ministry of Education. (2012). *Global Strategy Stage (GloSS).*

Shepard, L. S. (1997). The centrality of test use and consequences for test validity. *Educational Assessment: Issues and Practice, 16*(2), 5–24. https://doi.org/10.1111/j.1745–3992.1997.tb00585.x.

Stemler, S. E. (2004). A comparison of consensus, consistency, and measurement approaches to estimating interrater reliability. *Practical Assessment, Research & Evaluation, 9*(4), 1–11. https://doi.org/10.7275/96jp-xz07.

Vaughn, S., Shay Schumm, J., & Sinagub, J. (1996). *Focus group interviews in education and psychology*. Sage Publications.

Acknowledgments

The authors would like to thank Robin Schumacher, Pamela Foremski, and Christopher Tran for assisting with the research and preparation of the report.

REL 2020–039

September 2020

This report was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-17-C-0011 by the Regional Educational Laboratory Southeast administered by Florida State University. The content of the publication does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.