An Argument Based Approach to Evaluating the Reliability of First Contact

Amanda Ferster

Julia Shaftel

Alan Sheinker

Sookyung Shin

Patti Whetstone

University of Kansas

Author Note

The present publication was developed under grant 84.373X100001 from the U.S. Department of Education, Office of Special Education Programs. The views expressed herein are solely those of the author(s), and no official endorsement by the U.S. Department should be inferred.

Abstract

The purpose of the First Contact Survey is to collect fine-grain information regarding the students who participate in the Alternate Assessment based on Alternate Achievement Standards (AA-AAS). Educator ratings will be used to facilitate an understanding of the assessment population, provide insight into participation requirements across states, and inform the individualization of the Dynamic Learning Maps (DLM) online assessment system. Given these goals, it is imperative that the obtained ratings provide an accurate account of student characteristics and assessment needs. In order to evaluate the consistency of student ratings across educators, the DLM consortium administered a First Contact Reliability Pilot Survey. In contrast to simply reporting consistency indices, the consortium relied upon an *argument-based approach to reliability.* This process fostered our explication of acceptable index thresholds and permitted a focus on continuous improvement. The consortiums' application of an argument-based approach to reliability, the process of using the approach with the First Contact Survey, is the primary objective of this paper. Because indices and survey results are of specific interest to special educators, however, sample statistics and characteristics of the rated students rated are presented through tables and figures.

An Argument Based Approach to Evaluating the Reliability of First Contact

Since the reauthorization of the Elementary and Secondary Education Act, the No Child Left Behind Act of 2001 (NCLB, 2001), and its subsequent non-regulatory guidance (USDE, 2003; 2005), efforts have been made to better understand the population of students who are eligible to participate in the Alternate Assessment based on Alternate Achievement Standards (AA-AAS). Federal guidance reserves the AA-AAS for those students with ‘significant cognitive disabilities,’ yet maintains state flexibility in defining the criterion that constitutes this label. As a result, the population is currently ill defined. Researchers have developed two instruments in an effort to resolve this problem. The Learner Characteristics Inventory (LCI) (Kearns, Kleinert, Kleinert, and Towles-Reeves, 2006) and the Student Survey (Alternate Assessment Collaborative, 2004)--both of which strive to understand the characteristics of students that comprise this population. The Learner Characteristics Inventory, developed by the National Alternate Assessment Center (NACC) and field experts, consists of ten questions. It purports to address the areas of expressive communication, receptive language, vision, hearing, motor skills, engagement, health issues and attendance, reading, mathematics, and augmentative communication (AAC) systems. Although the LCI provides a window to the students who participated in the AA-AAS, its purpose was not to capture a fine-grain level of detail. For instance, Kearns et al. (2011) noted that although the researchers were able to determine the number of students relying on AAC for communication, the LCI does not collect data regarding the type of augmentative communication system used. The developers of the LCI authorize states to delete items in their entirety or add items that meet their unique needs. Alterations to inventory design, in an effort to gain specificity, however, do not permit an analysis generalizable across state agencies. The Alternate Assessment Collaborative, managed by the Colorado Department of

Education, developed the Student Survey with two components (Alternate Assessment Collaborative, 2004). The first survey consisted of 111 questions that aimed to collect in-depth student information in the four areas of demographics, assistive technology, communication, and mobility. The second component requested that teachers select a descriptive scenario that was most similar to the student they were assessing. Teachers had the opportunity to make clarifications if the scenarios did not resemble the target student under consideration. While the detail of the survey is commendable, the response burden of the instrument is high.

Currently, there are two consortiums engaged in developing alternate assessments linked to the Common Core State Standards, the Dynamic Learning Maps (DLM) Alternate Assessment Consortium and the National Center and State Collaborative (NCSC) Consortium. Now that a large-scale cooperative development opportunity exists, it is imperative to understand this population--both within and across the consortia. A well-defined population will aid test development efforts as well as inform consensus driven eligibility guidelines. With regard to the Dynamic Learning Maps project (DLM), a need exists for a more descriptive understanding of the population than what the previously developed inventories permit. Fine-grain information is necessary. That is, the assessment system defines routes of student learning, some of them influenced by the unique visual, auditory, or communication needs of the student. Students need not follow the same path to reach a specific target skill. Moreover, the mode of test delivery is computer based; thus, the test engine must work in concert with the students' assistive devices. Ultimately, test developers require a gauge of the variability of the students' current functioning and needs. The DLM team developed the First Contact Survey to meet this need.

In order to meet the survey objectives, educator classifications and ratings must accurately depict student characteristics, needs, and broad academic abilities. A study of the *consistency* across educator ratings facilitates the evaluation of whether inferences can be drawn when one rater completes the operational survey. Because the overarching goal is to ensure accurate ratings, the Dynamic Learning Map team relied on J. Parkes (2007) *Reliability as an Argument* approach. While his original publication was devised for measurement, it provides a clarifying organizational structure and encourages modifications in the spirit of continuous improvement. Therefore, this approach best served our needs.

The First Contact Survey

The First Contact Survey is a web-based inventory comprised of approximately 65 items. One educator that has extensive knowledge of each student participating in the AA-AAS will complete the operational instrument. The survey collects information regarding rater and facility characteristics, student demographics, special education placement, sensory perception, motor skills, expressive and receptive language, computer access, use of assistive technology, use of augmentative and alternative communication devices, academic skills, and engagement with and attention to instruction. Although the survey covers numerous domains, with many items designed as cross-tabular rating scales, the instrument employs skip logic to reduce response burden. The operational administration, the *First Contact Census Survey*, is a current data collection effort. The window opened on November 1, 2012 and will remain open to May 1, 2013.

Depicted visually in Figure 1 are the varied goals of the First Contact Census Survey (i.e., goals of the operational administration). Initially, the data collection effort will facilitate our fine-grain understanding of the AA-AAS population. Next, the results will influence students' Personal Needs Profile. The 'Personal Needs Profile' (PNP) is part of the 'Accessible Portable Item Protocol Standard' (APIP), the standard the next-generation assessments are implementing to ensure that the accessibility needs of students with disabilities are met. Here, individual student characteristics, garnered through the ratings, will provide insight into the profile. The connection of First Contact to the PNP will reduce response burden for primary educators, yet maintain educator control by permitting him/her to modify any setting established through First Contact. Once the primary educator confirms the profile, the information will interact with assessment delivery in two respects. The PNP will merge with Accessible Portable Item Profile (APIP) tags to administer items appropriate for the student *and* the academic skill ratings of First Contact will initiate the student into a specific region of the map. That is, the *initial item* that a student receives within an embedded assessment will coincide with the teacher's rating of student ability (i.e., as the student answers additional items, item presentation will be based on prior item response).

Item Development & Usability Study

Following best design practice, the First Contact development team solicited feedback from field experts during item development. Each item was evaluated with an eye toward domain coverage, textual clarity, and inclusive response options. In an effort to further evaluate the comprehensiveness of the instrument, open-ended comment fields concluded each component (e.g., after queries regarding a student's vision and visual aids, each participant was presented with an open-ended item, "comments on the child's vision").

In the winter of 2012, the DLM team piloted the instrument in an effort to evaluate the usability of the survey. Through the usability study, participants provided feedback related to the format of the items and the ease of item progression throughout administration. Researchers used the usability pilot results to modify instructions and item design within the online platform, Qualtrics.

First Contact Reliability As An Argument

Data Source

During the summer of 2012, the DLM team administered the *First Contact Reliability Pilot* to collect and evaluate the consistency of student ratings. Seven of the 13 DLM partner states participated in the Reliability Pilot. Six states permitted educators to conveniently sample the students to rate at the school level. Several educators from the seventh state, North Carolina, participated in the survey during a DLM professional learning module.

Administration details regarding the pilot were shared with volunteers via DLM partner meetings and electronic communications. During administration, *two professionals*—both with extensive knowledge of a specific AA-AAS participating student--completed the survey. The Reliability Pilot administration window closed on September 7, 2012. Three unique student identification fields facilitated rater pairing.

The total number of valid student ratings, 758, represents a 50% response rate of the number originally intended. Researchers were conservative in their merge process, pairing cases only when confident of ratings by two educators. The total number of merged student records (i.e., two unique ratings per student) was 299. This represents approximately 79% of the valid data set. Table 1 depicts the number of merged records by DLM partner state.

Reliability Perspective

The research staff conducting the First Contact reliability study believes that researchers should not neglect the overarching values and purpose, guiding initial development, from the evaluation of the instrument's technical properties. Too often researchers analyze data according to conventional statistical methods without thought. Moreover, there is a tendency to treat recommended thresholds, such as consistency and stability coefficients, as absolutes. An argument-based approach to validity (Kane, 1992) has served to encourage an ongoing evaluation of validity and a deep understanding of measuring instruments according to predetermined goals. Treating reliability in a similar fashion does not diminish the technical standards of the evaluation. Instead, it provides focus to a reliability study—forcing researchers to concentrate on the need for specific conventional indices. If traditional approaches cannot support or refute claims, the researcher must cast a wider methodological net. This may take the form of unique approaches to reliability that may advance understandings and ultimately serve to improve the instrument.

Application of the Reliability as an Argument Approach

Following Parkes (2007) framework, the reliability argument for the First Contact Survey is defined through six components. The components, *Value Base, Purpose & Context, Replication, Required Tolerance, Evidence, and Judgment,* develop the argument from the aspects of reliability that are most valued, given the nature of the instrument, to the final culmination of evidence in support of a decision. Embedded in the framework are the reasons why specific reliability indices are important, how the indices interact with context, what level of reliability will be accepted as supportive evidence, and the types of data that constitute the multiple pieces of evidence. Table 2 depicts a synthesis of the First Contact Reliability Components.

The *Value Base* of First Contact relates to the importance of accurate item level response—each item adds to the description of the AA-AAS population and specific items pre-populate the student level PNP. The development team *values* that the information obtained best describes the student. If the rater is integrally familiar with the student, response should not be a function of the rater. Consistency across raters is revered over stability across time. Student demographics should be stable. However, stability across academic variables negates the premise of learning or growth.

As previously described, the *purpose* of the First Contact Survey is multi-faceted. It describes the AA-AAS population in the aggregate and informs the online assessment system at the individual level. The diverse goals of the instrument directly influence the prioritization of reliability indices according to *context*. That is, assumptions are developed based on the interaction of context and purpose—the assumptions enable a prioritization of indices. Reliability indices vary according to what constitutes or defines a replication. *Replication* was defined as two ratings per student. As depicted in Table 2, through continuous evidence collection efforts, the definition of replication will expand to incorporate an evaluation of stability and accuracy. The first inter-rater consistency assumption was that practitioners of the same profession (i.e., both educators) would rate students more similarly over professionals serving in disparate roles (i.e., an educator and a therapist). The DLM team felt that agreement indices were likely to diverge if raters did not have the same opportunity to observe the student within a similar academic setting. A second assumption was that rater agreement would digress according to student grade-band. As a student progresses to the secondary educational level, his/her nominal demographics should remain consistent, yet his/her performance and attention to instruction may legitimately vary according to his/her relationship or preference toward a specific educator. Prior to analysis, the DLM team stated that if evidence supported the assumptions, the 'tolerance for error' would be different according to context; the evaluators set pre-determined requirements regarding tolerance. *Required Tolerance* declares the index threshold corresponding to each condition. Alternately stated, the DLM team stated, prior to analysis, that exact agreement, Kappa statistics, and intraclass correlations would be lower for the group of students rated by only one educator and a support staff member and for high school students irrespective of type of rater—the lower indices were acceptable under these conditions.

The *evidence* supporting the over-arching reliability claim took the form of statistics. Qualitative rater response provided additional support through the qualification of specific numerical ratings. Rater agreement was analyzed through agreement descriptive statistics--percent of exact, adjacent, and discrepant ratings defined by zero, one, and greater than or equal to two discrepancy points, respectively. Cohen's Kappa (Cohen, 1960) provided chance-corrected inter-rater agreement indices. Kappa statistics measure the degree of agreement over what is expected by chance alone. However, influencing Kappa is the prevalence and balance within the table (Jelles, Coen, Beenekon, Lankhorst, Sibbel, & Bouter, 1995; Sim & Wright, 2005). Researchers attended to the symmetry of each table and evaluated the recommended Kappa thresholds in conjunction with the agreement descriptive statistics (Sim & Wright, 2005). The researchers considered Kappa values greater than .60 as acceptable. If the value was calculated using academic skill variables or within a context described above, we permitted values in the moderate range, .41-.60. Intraclass correlations were also evaluated, the indices provide an index of the variance attributable to the students and in this instance absolute agreement among raters (McGraw & Wong, 1996). Table 3 presents sample indices.

With respect to *judgment*, the DLM team evaluated results according to our original assumptions. Neither of our assumptions related to differential indices by group were correct. That is, the highest rater agreement was realized between a primary educator and a paraprofessional. With regard to grade level of the students, indices were generally consistent across grade-band (i.e., elementary, middle, or high). Agreement within academic ratings, however, was lower than more observable student characteristics. While evidence suggested the overall consistency across raters was acceptable, the team evaluated the results with an eye toward improvement for the First Contact Census Survey. The team developed action-steps including strengthening rater requirements (i.e., only primary educators may rate students on the Census Survey), developing an administration fact sheet and video, releasing the full survey to our state partners to share with educators prior to administration, and the team modified several items with respect to embedded definitions and/or design. Subsequently, the research team shared an extended presentation with our state partners. We briefly described the reliability as an argument approach, reviewed indices that supported or refuted our assumptions, and discussed our specific action steps that would improve the First Contact Census Survey data collection effort and experience. Our state partner feedback led us to one additional action-step, to release the survey data to each state on a monthly basis. As a final step in the approach, the DLM team discussed alternate data sources and *future research* that will assist in supporting or refuting our First Contact claims.

Scholarly Significance

This study complements both the methodological and special education literature. According to Parkes (2007) applications of the 'reliability as argument' approach are needed to demonstrate the utility of the conceptualization. Without clear examples, practitioners are less apt to frame their studies within an unfamiliar structure. This study provides an example of the process using a different type of instrument, a survey, yet the approach was exceedingly useful. Results facilitated action steps; these steps were further developed through a group discussion of results with our stakeholders.

The results also informed our understanding of the degree to which professionals are able to classify the fine-grain needs of students with significant cognitive disabilities. Overall, educational staff are consistent in their ratings. As expected, raters diverged within the ratings of academic skills. Understanding academic achievement, however, is the primary goal of the assessment system.

Footnote

The results describing the population of students who participate in the Alternate Assessment based on Alternate Achievement Standards will be available once the Dynamic Learning Maps First Contact Census Survey window closes this spring. Graphics of the preliminary student characteristics based on the rating of only the primary educator are shown in figures 2 through figure 4. Please note, these results are based on only 299 students whereas the anticipated *N* for the First Contact Census Survey is approximately 100,000. We present the results, here, based on a suggestion garnered through the review process. Dynamic Learning Map staff presented the descriptive information at the TASH conference (November, 2012).

References

Almond, P., & Bechard, S. (2005). *An in-depth look at students who take alternate assessments: What do we know now?* Denver, CO: Colorado Department of Education.

Alternate Assessment Collaborative (2004). *Student Survey*. Denver, CO: Colorado Department of Education.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46. doi: 10.1177/001316446002000104

Cohen, J. (1968). Nominal scale agreement with provision for scaled disagreement or partial credit, *Psychological Bulletin, 70*(4), 213-220.

Jelles, F., Bennekom, C., Lankhorst, G.J., Sibbel, S.P., & Bouter, L.M. (1995). Inter and intra-rater agreement of the rehabilitation activities profile. *Journal of Clinical Epidemiology, 48*(3), 407-416.

Kane, M. T. (1992). An argument-based approach to validity. *Psychological Bulletin, 112*(3), 527-535.

Kearns, J., Kleinert, H., Kleinert, J., & Towles-Reeves, E. (2006). Learner characteristics inventory. Lexington, KY: University of Kentucky, National Alternate Assessment Center.

Kearns, J., Towles-Reeves, E., Kleinert, H., Kleinert, J., Thomas, M. (2011). Characteristics of and implications for students participating in alternate assessments based on alternate academic achievement standards. *Journal of Special Education*, *45*, 3-14.

NO CHILD LEFT BEHIND Act of 2001, 20 U.S.C. § 6319 (2008).

Parkes, J. (2007). Reliability as an argument. *Educational Measurement: Issues and Practice, 26*(4), 2-10.

Sim, J. & Wright, C.C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85,* 257-268.

Towles-Reeves, E., Kearns, J., Kleinert, H., & Kleinert, J. (2009). An analysis of the learning characteristics of students taking alternate assessments based on alternate achievement standards. *Journal of Special Education, 42*, 241–254.

U.S. Department of Education. (2003). *Standards and Assessment: Non-regulatory guidance.* Washington, DC: Author.

U.S. Department of Education. (2005). *Alternate achievement standards for students with the most significant cognitive disabilities: Non-regulatory guidance.* Washington, DC: Author.

Table 1

Reliability Study Matched Student Ratings by State

	N Merged Student Ratings	
State	N	%
Iowa	31	10.4
Kansas	56	18.7
Michigan	73	24.4
North Carolina	7	2.3
Utah	60	20.1
Washington	17	5.7
West Virginia	55	18.4

Table 2

First Contact Reliability as an Argument

Claim: Response to the First Contact Survey is Reliable		
	Definition	First Contact
Value Base	Statements that depict which values are important. The reliability argument is constructed to best demonstrate evidence related to the values	Each First Contact item or query provides information. The value is the information that is collected accurately depicts student characteristics, needs, and general functioning in an academic setting. Reveres rater agreement (consistency) over stability. While demographics remain stable, stability in learning characteristics is antithetical. Most accurate rating is made by the person(s) who best knows the student.
Purpose & Context	Describes why the collected information must be reliable, how the evidence will be collected, what type of evidence will be shown, and in which context the evidence will be evaluated and to what degree	The First Contact Census results are used in the aggregate to describe the population and at the individual level to inform PNP and the online assessment system. The current primary educator will maintain his/her right to modify a pre-populated PNP field. *The classifications and ratings must accurately represent the student's characteristics and needs.* *Rater agreement must be collected and evaluated.* Agreement indices may differ according to the professional role of the rater Assumption 1: Agreement across primary educators will be stronger than agreement across non-educational staff Agreement indices may differ across student grade-bands Assumption 2: Agreement will be stronger for elementary students than for

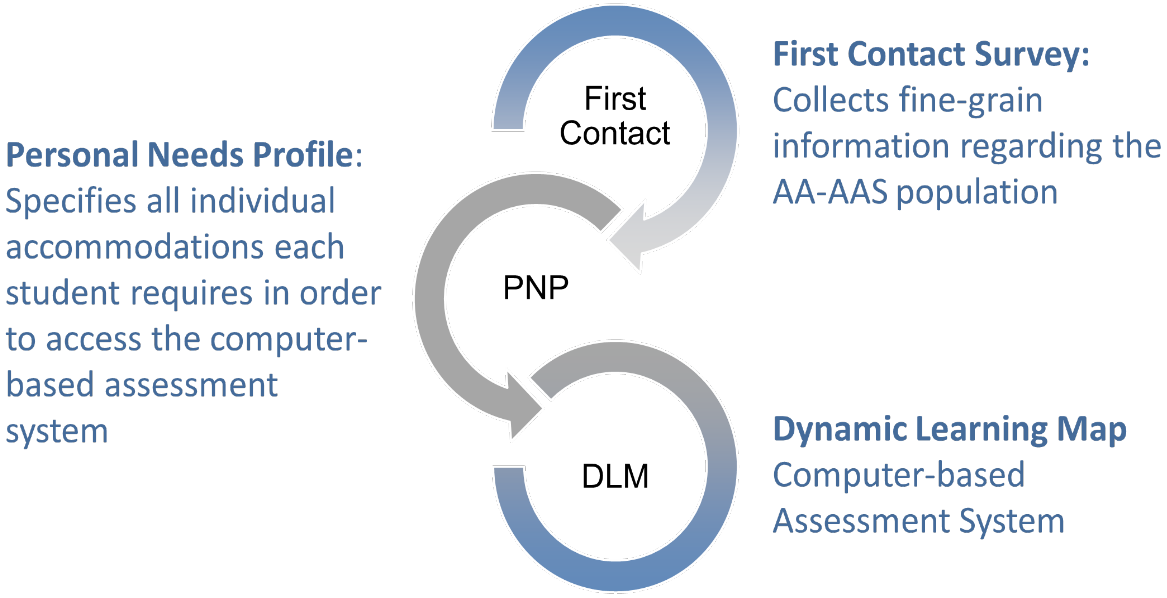
		high school students. Older children may intentionally react differentially dependent on educator preference.
Replication	Describe what constitutes a replication. It may be conceptual and overlap with traditional validity considerations	Ratings by two professionals within the current study Future Sources: Census ratings over consecutive years (applicable to demographics) Parent Ratings for students participating in observational laboratories Comparison of academic ratings to actual assessment performance (cross with validity)
Required Tolerance	Specify the 'tolerance of error' according to the purpose and context	Tolerance for error is lower for student aids and needs; higher for academic skill ratings as academic skills are better defined through an assessment system Tolerance for error is lower when both raters have intricate knowledge of the student
Evidence	Describe the evidence that will support the over-arching reliability claim	Qualitative comments serving to confirm or refute item response Descriptive rater agreement (percent exact, adjacent, and discrepant) Cohen's Kappa for nominal classifications Cohen's Kappa & Intraclass correlations for ordinal ratings
Judgment	Synthesize the above steps into an argument, share the inclusive argument with stakeholders, and render a judgment	In brief, evidence suggests that response to the First Contact Survey is reliable. However, specific results suggest actions may be put into place to improve confidence in response
Action Steps	Define any continuous improvement action steps garnered through the reliability as an argument process	Tighten rater requirements; only primary educators may complete the survey

		Expand the number of items requiring a mandatory response Generate an administration reference Fact Sheet Develop a brief online professional development video to reinforce administration Provide the full survey to state partners prior to administration Several items modified with respect to embedded definitions, separation of context, and web-based design
Potential Future Study	Describe plans for future evaluations that will continue to support or refute the argument	Evaluate the concordance of parental response to primary educator response with regard to the First Contact Census Survey Evaluate the relationship between First Contact Census Results and assessment performance

Table 3

Reliability Indices for Select Variables

	Rater Agreement % in Each Category			Absolute Agreement among Ratings			Agreement in Classification Above Chance
Item	Exact	Adjacent	Discrepant	ICC	Lower Bound	Upper Bound	Kappa
Highest Level of Understanding Instruction	63.1	28.7	8.2	0.579	0.498	0.651	0.453
Approximate Instructional Reading Level	73.2	20.7	6.1	0.899	0.862	0.911	0.667
Expressive Communication with Speech	78.4	18.7	2.9	0.715	0.633	0.781	0.596



First Contact Survey: Collects fine-grain information regarding the AA-AAS population

Personal Needs Profile: Specifies all individual accommodations each student requires in order to access the computer-based assessment system

Dynamic Learning Map Computer-based Assessment System

Figure 1. The Goals of the First Contact Census Survey

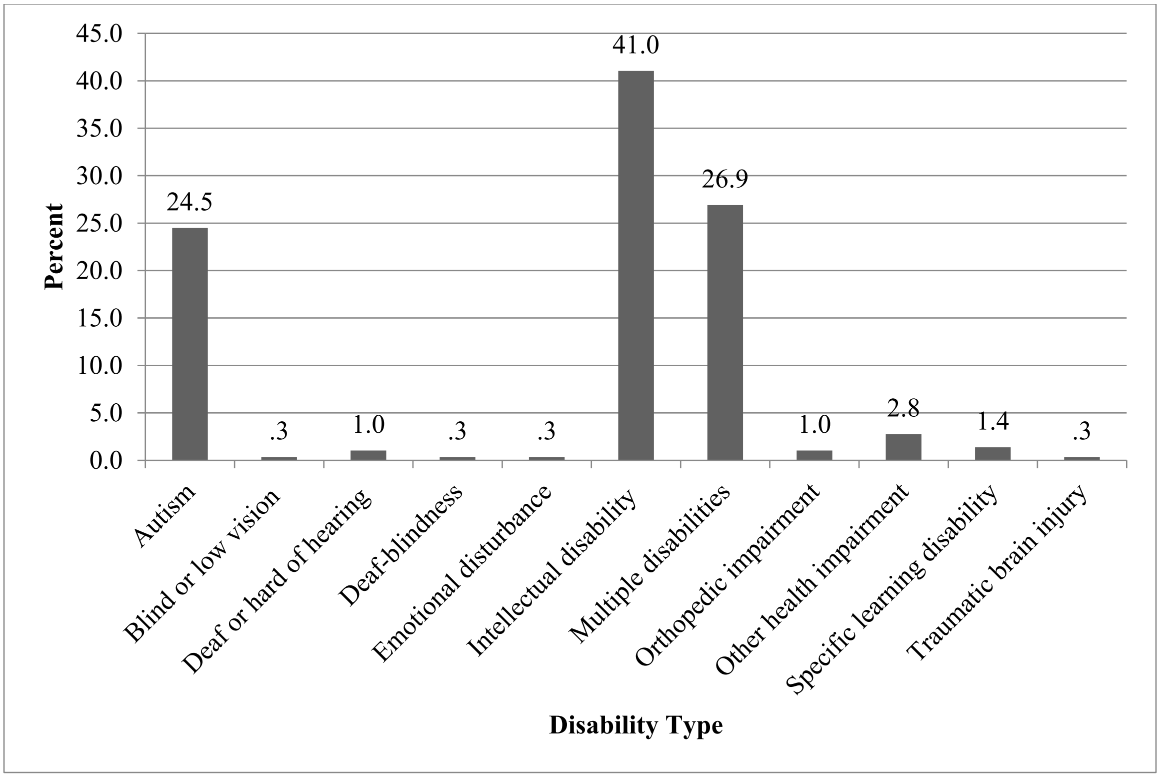


Figure 2. Primary Disability of the Students Rated by a Primary Educator within the First Contact Reliability Study Survey.

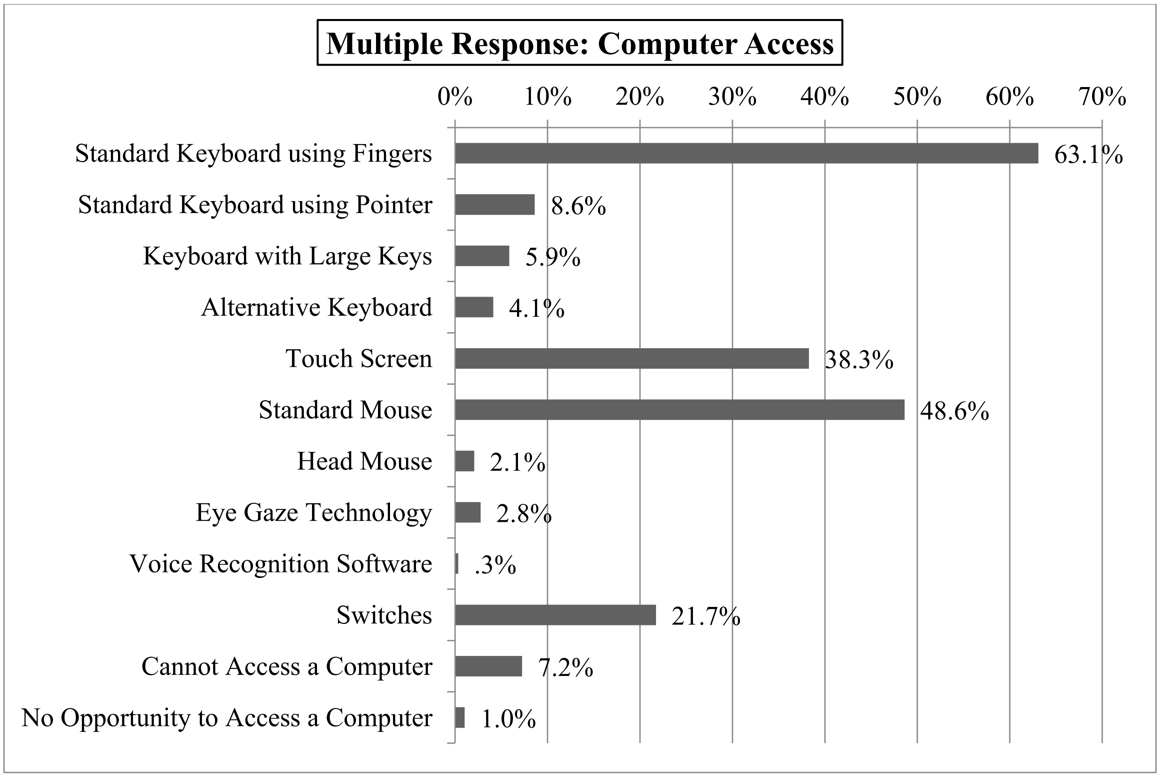


Figure 3. Computer Access of the Students Rated by a Primary Educator within the First Contact Reliability Study Survey.

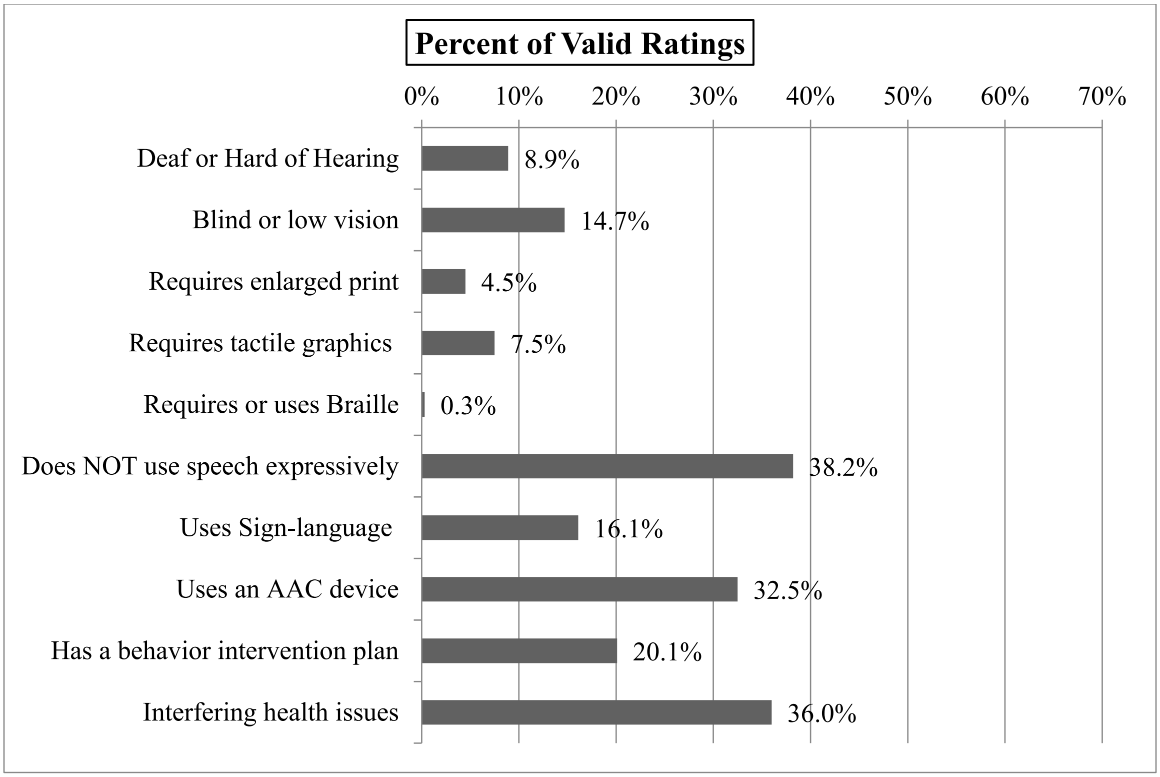


Figure 4. Selected Characteristics of the Students Rated by a Primary Educator within the First Contact Reliability Study Survey. The characteristics were selected due to their relation to the student Personal Needs Profile