Estimating Learner Ability with Item Response Theory

Stephen Gunter, PhD
Aptima, Inc.
Orlando, FL
sgunter@aptima.com

Jeffrey M. Beaubien, PhD
Aptima, Inc.
Woburn, MA
jbeaubien@aptima.com

ABSTRACT

Tests, simulations, and other modes of assessment are often used to capture large amounts of data from learners regarding the knowledge and skills that were trained. The data can come in various formats and from various assessment types such as knowledge checks, behavioral checklists, or rating scales. There is a wide range of tools that can be used to analyze the data captured from these assessments, each of which has their own assumptions, benefits, and costs. The use of appropriate analytical tools on these data becomes key to making accurate, data-driven decisions about learners. The purpose of this paper is to provide an overview of an analytical approach called Item Response Theory (IRT; Lord & Novick, 1968) that can estimate learner proficiency, discuss its benefits compared to traditional approaches, and describe how it can be used with an interoperable data format called the Experience Application Program Interface (xAPI; HT2Labs, 2019). We also provide practical guidance on how the reader can begin incorporating IRT-based analyses into their own efforts. All software code that was used in developing this paper is freely available from the first author upon request.

ABOUT THE AUTHORS

Dr. Stephen Gunter is a Senior Scientist in Aptima's Learning and Training Systems division. His expertise and work have been in designing, developing, and validating measurement methods for selection, certification, and diagnostic assessment. He has experience and expertise in situational judgment tests, surveys, classical test theory, item response theory, evidence-centered design, automatic item generation, and automated test assembly. He has worked with government agencies such as the Veterans Benefits Administration and with customers in the banking, call center, and customer service industry. He has a MS and PhD in industrial and organizational psychology from the University of Central Florida.

Dr. Jeffrey M. Beaubien is a Distinguished Principal Scientist in Aptima's Learning and Training Systems division. For the past 20 years, his work has focused on training and assessing leadership, teamwork, and decision-making skills. His research has been sponsored by the US Navy, the US Army, the US Air Force, and the Telemedicine and Advanced Technologies Research Center, among others. Dr. Beaubien holds a PhD in industrial and organizational psychology from George Mason University, a MA in industrial and organizational psychology from the University of New Haven, and a BA in psychology from the University of Rhode Island.

Estimating Learner Ability with Item Response Theory

Stephen Gunter, PhD
Aptima, Inc.
Orlando, FL
sgunter@aptima.com

Jeffrey M. Beaubien, PhD
Aptima, Inc.
Woburn, MA
jbeaubien@aptima.com

PROBLEM STATEMENT

Estimating a learner’s current level of proficiency is critical to making decisions about their readiness to perform specific tasks or duties. There is a wide range of analytical approaches that can be used. However, not all analytical tools are interchangeable, and using less effective tools, the wrong tools, or appropriate tools incorrectly can be problematic. For some positions where incorrect behaviors or ineffective decisions can result in injury or death, cost a lot of money, or cause critical delays in execution, the risk of over-estimating a learner’s proficiency can be quite high. The over-estimation would lead one to conclude that the learner has the required level of knowledge, skills, or abilities to perform the required tasks or make the appropriate decisions. However, the learner has less knowledge, skills, or abilities than is indicated by the estimation procedure and, therefore, has a higher likelihood of making an error or ineffective decision. The consequence of under-estimating a learner’s ability may also result in inefficient use of the learner’s time. For example, the learner may be assigned additional training that they do not need because they were estimated to have less knowledge and skill than they truly have. The overall consequence of this is a delay in getting proficient and ready individuals in their positions.

ESTIMATION OF LEARNER PROFICIENCY

Traditional Approach

A frequently used and traditional approach to estimating a learner’s proficiency is to count the number of correct responses that a learner makes on knowledge tests. Similarly, one might count the number of observed behaviors that are related to course objectives or skills that are being taught. Yet another approach is to calculate an average rating from a set of ratings made by an instructor on a grade sheet during training. These approaches take an overall view of performance and focus on an unweighted total score to estimate proficiency. If the total is higher than the pre-defined standard, then the learner passes the course or measurement event. If the total score is lower than the standard, then the learner does not pass and must engage in some sort of remediation.

The benefits of this approach are that it is a simple, straightforward calculation which can be used in instances where there is a small number of learners, and when all assessment items are weighted equally. However, this approach does not inherently consider various sources of information that can be useful for estimating a learner’s proficiency. For example, suppose at the end of a specific instructional module a knowledge test is administered to the learners. This test is composed of 10 multiple-choice items that are drawn from a bank of over 100 items. Common reasons for only using 10 items at a time from this large bank is to maintain item security by rotating items, to ensure that the entire content domain is covered, and to help ensure that the same assessment isn’t being administered repeatedly. Thus, each assessment will contain a somewhat different sample of items for each learner.

Now suppose that two learners have each answered 7 out of 10 questions correctly. Simply focusing on number-correct would lead to the same estimate of proficiency and to the same decision about learner readiness. However, if the two assessments were not at the same level of difficulty, then the estimates of learner proficiency and the decision about readiness would be incorrect. Answering 7 out of 10 questions correctly from a moderately easy assessment is not the same as answering 7 out of 10 questions correctly from a moderately difficult assessment. While this is a relatively stark example for illustrative purposes, the main point is that there is additional information contained in the assessment that can be useful for estimating learner proficiency. The first source is difficulty.

Two other important sources of information that are not inherently considered by the number-correct approach are discrimination and guessing. Discrimination is the ability of an assessment item or overall assessment to make distinctions between learners of different proficiency levels. The higher the discrimination, the better able the item or assessment can differentiate among learners with proficiency levels that are very close to each other. For example, in quality assurance one person may have the ability to detect deviations five degrees or more from the standard whereas another person may be able to detect deviations three degrees or fewer from the standard. Finally, guessing is the probability that a learner can guess the correct answer simply by chance. Learners who have far more knowledge or skill than is required for an item will be unlikely to attempt a guess. By contrast, a learner who has far less knowledge or skill than is required for an item will be more likely to attempt a guess.

Item Response Theory

Item Response Theory (IRT) is an alternative analytical tool that can incorporate this additional information about the assessment items into the proficiency estimate (Lord & Novick, 1968). IRT is a flexible family of psychometric models that estimate the probability that a person with a given level of a latent trait (i.e., ability or proficiency) will provide a particular output or response to an assessment item. As its name implies, IRT is focused on items and response patterns to those items, not necessarily on the overall assessment score.

This explicit focus on the characteristics of the items as well as the ability of the person provides IRT an important basis for use in various applications. IRT has been used on a range of measurement formats such as multiple-choice items (Birnbaum, 1968), forced-choice responses (Stark, Chernyshenko, & Drasgow, 2005), Likert-type rating scales (Andrich, 1978; Bock, 1972; Masters, 1982; Muraki, 1992; Roberts & Laughlin, 1996; Samejima, 1969), and situational judgment tests (Manniste, Pedaste, & Schimanski, 2019; Zu & Kyllonen, 2020). Moreover, IRT has been used in a range of testing programs including the Armed Services Vocational Aptitude Battery (ASVAB; Defense Manpower Data Center, 2012), United States Medical Licensure Examination (USMLE), and the Graduate Record Exam (GRE), to name only a few. Most notably, because the focus of IRT is on the items, it has been used in many computer adaptive testing (CAT) approaches.

One of the benefits that underlies IRT's usefulness for various applications is how it represents the relationship between ability, the item characteristics, and the participants' responses. IRT seeks to predict a set of responses to items as a function of the person's ability and the characteristics of those items. As shown in (1) and depicted in ***Figure 1***, the prediction takes the form of a logistic function and is called an Item Response Function (IRF). The IRF depicted in ***Figure 1*** is for the three-parameter logistic (3PL) model because three parameters are estimated: item difficulty (beta, β), discrimination (alpha, α), and guessing (*c*).

$$P\left(x_j = 1 \,\middle|\, \theta, \beta_j, \alpha_j, c_j\right) = \; c_j + \left(1 - c_j\right) \frac{e^{\alpha_j(\theta-\beta_j)}}{1+\, e^{\alpha_j(\theta-\beta_j)}} \qquad (1)$$

As depicted in ***Figure 1***, the IRF shows that as the person's ability or proficiency increases, they have a higher probability of providing a given response (e.g., answering an item correctly), where ability is depicted by θ (theta) on the x-axis and the probability of a correct response is on the y-axis. This relationship makes intuitive sense. If a learner's proficiency is far below what is required for the assessment item, they should have a very low probability of providing a correct response, except by guessing. If a learner's proficiency is far above what is required, they should have a very high probability of providing a correct response.

The 3PL IRF below can be interpreted in the following ways. First, difficulty (β) is also called item location because it refers to where on the x-axis the IRF is located in a left-to-right orientation. An IRF that is located to the far left represents a relatively easy item. An IRF that is located to the far right represents a relatively difficult item. The item shown in ***Figure 1*** has a difficulty of 0.71, which is depicted by the black vertical line. The value of 0 on the x-axis represents average. Thus, this item has a slightly higher-than-average difficulty. Second, the discrimination parameter (α) refers to the slope of the IRF, which is depicted by the green line. The steeper the slope of the line, the higher the discrimination, where higher discrimination is a desirable characteristic. The example item below has a discrimination of 1.71. Finally, the guessing (*c*) parameter is the point where the IRF intersects the y-axis, depicted by the purple line. Guessing is defined as the probability that someone with infinitely low ability will be able to provide a correct response. The guessing parameter for this item is 0.22.

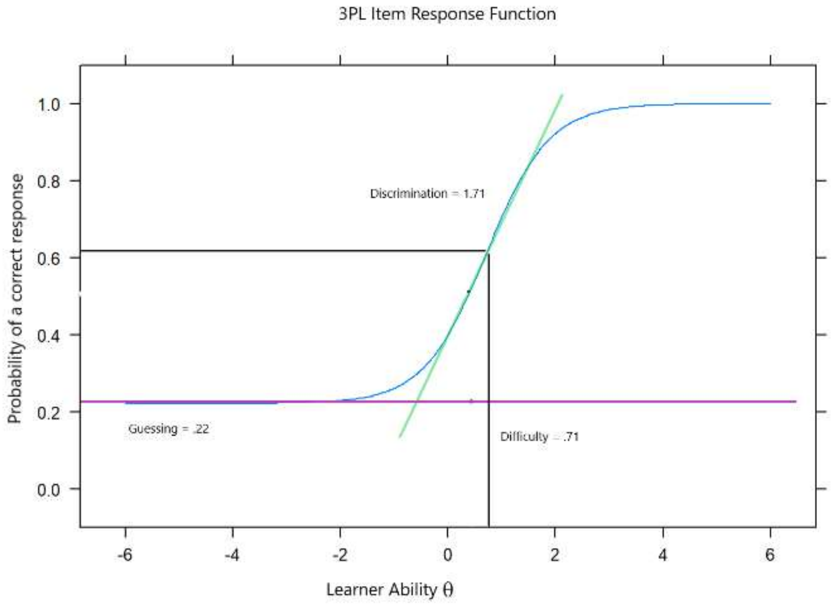

Figure 1. 3 Parameter Logistic (3PL) Item Response Function (IRF)

In addition to considering the characteristics of the assessment item, another important benefit of IRT is that it puts both people and items on the same scale. In other words, the representation of the learner's ability uses the same number line as the assessment of item difficulty. This allows for easy comparison between people and items. Flexibility is another important benefit of IRT. The flexibility of IRT comes from three important areas: 1) the number of abilities in the model; 2) the number of item characteristics that are estimated, and 3) the number of possible responses (i.e., scores). Each of these areas can be restricted or loosened to create different IRT models. In this paper, we will focus only on the 3PL model. Interested readers can refer to de Ayala (2009) or to Embretson and Reise (2000) for more detailed information about other IRT models.

The benefits of IRT in comparison to traditional methods are notable. However, there are drawbacks to IRT that practitioners, instructors, and analysts must consider before implementing IRT in their training and assessment program. The first is complexity. While IRT is flexible, it is also more complex, and many of the parameter estimation procedures may not immediately be approachable to those without a background in psychometrics. The second major drawback is sample size, which may render the approach inapplicable to small sample studies. Many factors have an effect on the appropriate sample size (de la Torre & Hong, 2010; Harwell & Janosky, 1991; Seong, 1990; Stone, 1992). Generally, the more parameters to be estimated, the larger the sample size is required. The general rule of thumb based on research suggests a sample size approaching 1000 for a 3PL model (Baker, 1998; Drasgow, 1989; Harwell & Janosky, 1991; Hulin, Lisak, & Drasgow, 1982; Stone, 1992). As such, IRT is particularly useful in cases where the same assessment items are used repeatedly, such as a standardized end-of-course completion exam. It is also extremely useful in cases where there is a large bank of potential assessment items with the need to develop equivalent "alternate forms" of the test, for example in cases where the learners are highly motivated to cheat.

Interoperable Data Storage and Exchange

As with any proficiency estimation procedure, numbers representing scores and responses are the main source of data for IRT models. If the responses are transformed to numbers, then IRT can be applied. However, before IRT can be applied, the data must be in a particular tabular format where rows represent the learners, columns represent the assessment items, and cells contain the numbers that represent scores or responses. This is often referred to as "tidy" data (Wickham & Grolemund, 2016). Different sources of data or assessment modalities may not produce data in a standardized and easily accessible format. What is needed are data in particular formats and structures that make themselves readily available for the analytical tools. Recently, there has been development of data standards that allow for any number of training and performance assessment tools. A primary example, and the focus of this paper, is the Advanced Distributed Learning initiative's (ADL) Experience Application Programming Interface (xAPI).

xAPI is a data specification that defines how a learner's experiences, responses, behaviors, and/or performance should be structured for documentation and exchange. ADL has created specifications regarding appropriate statement properties and syntax (Advanced Distributed Learning Initiative, 2017). Learner experiences and responses are captured from a simulation, multi-player game, mobile or wearable device, computer-based knowledge assessment, or tablet-based instructor rating and then serialized according to JSON (JavaScript Object Notation) rules. The xAPI-formatted learner records are then stored in a Learner Record Store (LRS).

xAPI statements have several important components that makes them highly useful and relevant for learner analytics. Each statement contains properties about the learner (*actor*), what the learner did (*verb*), and what the learner interacted with (*object*). A benefit of the xAPI statements is that they are extensible and can include useful contextual information such as metadata, attachments, and date/time stamps which are useful for longitudinal assessments of learner growth (Andrade & Tavares, 2005; Embretson, 1991). An example xAPI statement that is applicable for IRT is shown below in ***Figure 2***. In this simple example, learner Stephen Gunter (actor = "sgunter@example.com") answered (verb = "answered") a single item on a knowledge test (object = "QID100") correctly (success = "true"). Additional information is also contained in the xAPI statement. For example, it indicates that the English text of question ID 100 was "Which of the following animals can run the fastest on land?" and that the four possible response options are: "Saluki," "Springbok," "Ostrich," and "Northwestern Wolf."

```
{
  "actor": {
    "mbox": "mailto:sgunter@example.com",
    "openid": "http://sgunter.openid.example.com",
    "account": {
      "name": "Stephen Gunter",
      "homePage": "http://example.com"
    },
    "name": "Stephen",
    "objectType": "Agent"
  },
  "verb": {
    "id": "http://adlnet.gov/expapi/verbs/answered",
    "display": {
      "en-US": "answered"
    }
  },
  "object": {
    "id": ".../DeclarativeKnowledgeAssessment/MultipleChoiceExample.html",
    "definition": {
      "name": {
        "en-US": "QID100"
      },
      "description": {
        "en-US": "Which of the following animals can run the fastest on land?"
      },
      "type": "http://adlnet.gov/expapi/activities/cmi.interaction",
      "interactionType": "choice",
      "choices": [
        {
          "id": "A",
          "description": {
            "en-US": "Saluki"
          }
        },
        {
          "id": "B",
          "description": {
            "en-US": "Springbok"
          }
        },
        {
          "id": "C",
          "description": {
                  "en-US": "Ostrich"
               }
            },
            {
               "id": "D",
               "description": {
                  "en-US": "Northwestern Wolf"
               }
            }
         ],
         "correctResponsesPattern": [
            "B"
         ]
      },
      "objectType": "Activity"
   },
   "result": {
      "score": {
         "scaled": 1,
         "raw": 1,
         "max": 1
      },
      "success": true,
      "completion": true,
      "response": "Springbok",
      "duration": "PT1M12S"
   }
}
```

Figure 2. Example xAPI Learner Record

Properly formatted xAPI learner records contain all the information that is needed for IRT analysis. First, the *actor* is required because this specifies for IRT that this is a unique learner who is responding to some type of assessment item. Second, the *object* is required because it identifies the unique assessment item and contains information about the item itself. Third, the statement also provides the specific response that was made (i.e., correct response pattern = "B," the English description is "Springbok") and its score (i.e., scaled score = "1" or success = "true"). The beauty of xAPI learner records is the common structure allows for automated parsing and transformation processes to be conducted. This then renders that data contained within the statement to be readily accessible for analytical tools such as IRT. In the following sections, we describe how to extract tidy data quickly and easily from xAPI learner records and how to analyze them using 3PL IRT models. Full, documented software code in the open-source R programming language is available directly from the first author upon request.

METHOD

To begin, we used ADL's xAPI statement generator (Advanced Distributed Learning Initiative, n.d.) to generate a small set of xAPI learner records for 11 learners who each completed two assessment items. The generated xAPI statements are available from the first author. We did not generate xAPI statements for hundreds of learners and dozens of assessment items —as would be typical for many IRT analyses—because we only needed to demonstrate a repeatable process that would work with any number of people and assessment items. Thus, a small set of people and items was sufficient.

The main information that is needed for IRT that must be parsed and transformed is the *actor*, *object*, and *score*. We used the open source *jsonlite* package (Ooms, 2014) within the open source R statistical application to parse the xAPI statements. The parsing function in *jsonlite* focuses on opened and closed curly brackets, the key:value pairs within opened curly brackets, and the commas that are used to delimit multiple key:value pairs. The function marches through the statement at the first level of curly bracket indentation before moving to lower levels of indentation. For example, in ***Figure 2***, the function parses "actor:mbox" (email address), "actor:openid," (alternate identifier), "actor:name" (nickname) and "actor:objecttype" (agent) before parsing actor:account:name (full name) and actor:account:homepage (homepage) because the latter two indented within the former. Most of the information in the xAPI statements can easily be parsed in this way into a tidy R data frame.

However, special attention must be paid to nested statements such as “choices” (i.e., response options), which are nested underneath object:definition:choices. The parsing function sees the nested choices as a data frame and does not parse them as unique elements. This limitation is not problematic for the 3PL model described in this paper because it focuses only on the assessment item scores, not the specific option choices. However, the nested data frame for the choices would need to be parsed if the analyst wanted to focus on the specific response options.

The parsing function creates a large data frame that contains all the information from the xAPI statement. As noted above, much of this information is not used by the 3PL model. Therefore, we conducted a few extra steps to transform the data frame into one that only contains the relevant IRT information. First, we ensured that the parsed information was in tidy format, then renamed each column (i.e., variable name) to something more interpretable for our analysis. We then selected only the information that is relevant to the 3PL model: the unique learner, the two questions, and the score for each learner on each question. Finally, we spread the data from a single column that contained responses for both questions to a separate column for each question. The resulting data frame is shown in ***Table 1***. It includes 11 learners (sgunter, mdupree, jrice, lriley, etc.) and two assessment items (question ID 100 and question ID 101). Finally, the data are coded in binary format, with values of “1” indicating a correct response and values of “0” indicating an incorrect response. In the example below, learner “sgunter” correctly answered both questions, while learner “mdupree” incorrectly answered the first question, but correctly answered the second one. The transformation from xAPI formatted learner records to “tidy” data was relatively straightforward and could easily be completed by anyone with a basic background in the R open source statistical package. It can also parse thousands of xAPI-formatted learner records, not just the 11 learners in this simple example.

Table 1. **Data Frame for IRT Analysis**

LearnerID	QID100	QID101
http://sgunter.openid.example.com	1	1
http://mdupree.openid.example.com	0	1
http://jrice.openid.example.com	1	0
http://lriley.openid.example.com	1	0
http://lrselmon.openid.example.com	1	1
http://bstoops.openid.example.com	0	0
http://jwashington.openid.example.com	0	1
http://rcalmus.openid.example.com	1	0
http://tlehman.openid.example.com	0	1
http://apeterson.openid.example.com	0	0
http://gmccoy.openid.example.com	1	1

As noted previously, IRT works best with large numbers of learners. Therefore, while the previous R code was sufficient to show how easily it is to parse properly formatted xAPI learner records, we needed a larger data set to demonstrate the 3PL model. To do this, we used the open source *mirt* R package (Chalmers, 2012) to simulate data for 2000 learners across 10 items. We developed a set of realistic difficulty, discrimination, and guessing values and used the default ability levels. Within the *mirt* package, we used the simdata function to simulate responses for all 2000 learners. Next, we used the mirt function to estimate the item parameters and learner ability levels. We will describe the basic process of estimation but not fully cover topics such as model fit, item fit, and person fit because it is beyond the scope of this paper. Interested readers can refer to de Ayala (2009) or Embretson and Reise (2000) for more detailed information.

The basic process of estimating item parameters and learner abilities is called calibration. Calibration is a two-step process where the parameters and abilities are estimated and then scaled so they share a common scale. The creation of a common scale is necessary for assessment items and learners to be directly comparable. Part of the first step is to assess how well the model fits the data. Typical reasons for the model not fitting the data include multi-dimensionality, aberrant responding, or poorly functioning assessment items.

Next, we retrieved the estimated item parameters, ability levels, expected test scores, and item response functions (IRFs) from the coef, fscores, and expected.test, itemplot functions, respectively. The item parameters from coef and IRFs from itemplot can be very useful sources of information about the overall functioning of the assessment items during initial stages of assessment development. The ability levels from fscores and expected test scores from expected.test would be the primary information that can be used by practitioners or instructors to assess the current levels of learner ability based on the information the IRT model used in the estimation process. The estimated learner abilities can be used, for example, on feedback reports or instructor or learner dashboards to show current progress towards accumulating knowledge, skill, or ability in targeted areas.

Figure 3 depicts the 3PL IRF for the first item in our simulated test. The IRF depicts a difficulty value of -0.24, a discrimination value of 1.07, and a guessing probability of .13. The results indicate that the first item is slightly easier than the average item because its value of 0.24 units is below 0. Also, the results indicate that the item discriminates reasonably well. We want the discrimination value to be at least 1.0. Finally, the results for guessing also show there is not a very high tendency for our learner to guess the correct answer by chance (13%). Overall, this item is functioning reasonably well.

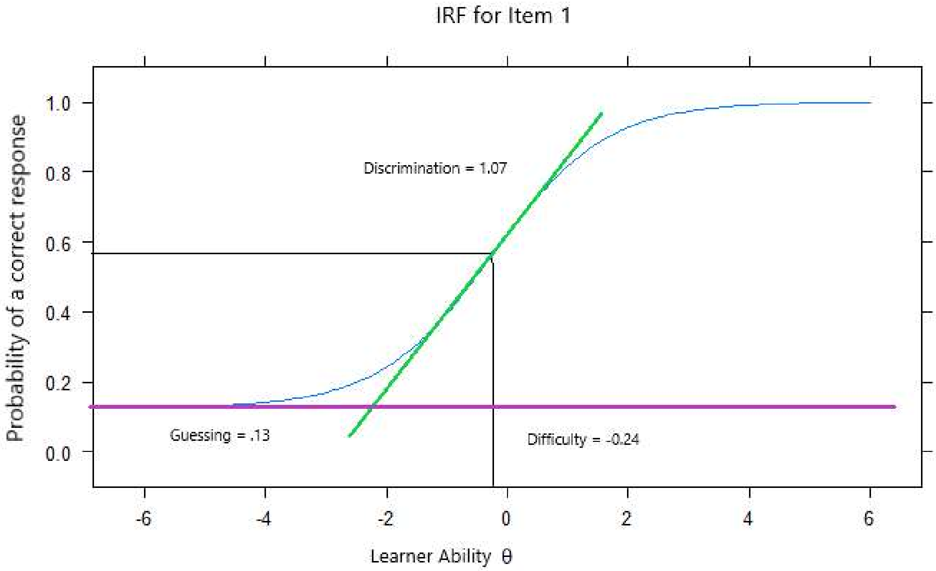

Figure 3. 3PL IRF for Item 1

In ***Table 2*** below, we have provided the learner ability estimates from the stimulated data and used the LearnerIDs from the xAPI statements we generated earlier. Recall, the ability estimate is on the same scale as the difficulty value, which allows practitioners and instructors to directly compare items and learners. The results indicate that the learners demonstrated a range of ability. Learners with high positive ability levels are likely to be successful at easy-to-moderately difficult items, whereas learners with low or negative ability levels are likely to be successful only on easy items. Instructors and practitioners can order the learners by ability level, develop a cut score, and use that cut score to pass or fail the learners. They can also use the ability and difficulty values to help determine which would be the best items to give the learner in subsequent knowledge tests.

Table 2. Learner Proficiency Estimates

LearnerID	Ability
http://sgunter.openid.example.com	1.55
http://mdupree.openid.example.com	0.81
http://jrice.openid.example.com	-0.68
http://lriley.openid.example.com	0.60

http://lrselmon.openid.example.com	-1.52
http://bstoops.openid.example.com	-1.05
http://jwashington.openid.example.com	0.81
http://rcalmus.openid.example.com	-0.28
http://tlehman.openid.example.com	-1.10
http://apeterson.openid.example.com	0.43
http://gmccoy.openid.example.com	-0.91

CONCLUSION

In conclusion, the standardized nature of the xAPI statements and JSON serialization makes data gathered from them readily accessible for use in statistical models such as IRT. Practitioners can use free, open-source software to quickly and easily parse and transform the properly formatted xAPI learner records into tidy data. Moreover, they quickly generate IRT models to estimate learner ability independent of item difficulty and guessing. The ability level estimates that account for this additional information may prove to be more useful and informational than a simple number-correct score, which does not consider this additional information. As noted previously, fully documented R software code is freely available from the first author upon request.

REFERENCES

Advanced Distributed Learning Initiative. (2017). *xAPI Statement Generator*. Available: https://adlnet.github.io/xAPI-Workshop/Tools/StatementGenerator/.

Advanced Distributed Learning Initiative. (2017). *xAPI Specification*. Available: https://github.com/adlnet/xAPI-Spec/blob/master/xAPI-About.md.

Andrade, D. F., & Tavares, H. R. (2005). Item response theory for longitudinal data: Population parameter estimation. *Journal of Multivariate Analysis, 95*, 1-22.

Andrich, D. (1978). A rating formulation for ordered response categories. *Psychometrika, 43*, 561-573.

Baker, F. B. (1998). An investigation of the item parameter recovery characteristics of a Gibbs sampling procedure. *Applied Psychological Measurement, 22*, 153-169.

Birnbaum, A. L. (1968). Some latent trait models and their use in inferring an examinee's ability. *Statistical theories of mental test scores*.

Bock, R. D. (1972). Estimating item parameters and latent ability when responses are scored in two or more nominal categories. *Psychometrika, 37*, 29-51.

Chalmers, R. P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29.

de Ayala, R. J. (2009). *The theory and practice of item response theory*. United Kingdom: Guilford Publications.

Defense Manpower Data Center (2012). CAT-ASVAB forms 23-27 (Technical Bulletin No. 4). Seaside, CA: Author.

de la Torre, J., & Hong, Y. (2010). Parameter estimation with small sample size: A higher-order model approach. *Applied Psychological Measurement, 34*, 267-285.

Drasgow, F. (1989). An evaluation of marginal maximum likelihood estimation for the two-parameter logistical model. *Applied Psychological Measurement, 13*, 77-90.

Embretson, S. (1991). A multidimensional latent trat model for measuring learning and change. *Psychometrika, 56*, 495-515.

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. United Kingdom: Erlbaum.

Harwell, M. R., & Janosky, J. E. (1991). An empirical study of the effects of small datasets on varying prior variances on item parameter estimation in BILOG. *Applied Psychological Measurement, 15*, 279-291.

Hulin, C. L., Lisak, R. I., & Drasgow, F. (1982). Recovery of two- and three-parameter logistic item characteristic curves: A monte carlo study. *Applied Psychological Measurement, 6*, 249-260.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Manniste, T., Pedaste, M., & Schimanski, R. (2019). Situational judgement test for measuring military tactical decision-making skills. *Military Psychology, 6*, 462-473.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, 159-176.

Ooms, J. (2014). The jsonlite package: A practical and consistent mapping between json data and r objects. *arXiv preprint arXiv:1403.2805*.

R Core Team (2018): R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Roberts, J. S. & Laughlin, J. E. (1996). A unidimensional item response model for unfolding responses from a graded disagree-agree response scale. *Applied Psychological Measurement, 20*, 231-255.

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph Supplement*, No. 17.

Seong, T. (1990). Sensitivity of marginal maximum likelihood estimation of item and ability parameters to the characteristics of the prior ability distribution. *Applied Psychological Measurement, 14*, 299-311.

Stark, S., Chernyshenko, O. S., & Drasgow, F. (2005). An IRT approach to constructing and scoring pairwise preference items involving stimuli on different dimensions: The multi-unidimensional pairwise preferences model. *Applied Psychological Measurement, 29*, 184-201.

Stone, C. A. (1992). Recovery of marginal maximum likelihood estimated in the two-parameter logistic response model. An evaluation of MULTILOG. *Applied Psychological Measurement, 16*, 1-16.

Wickham, H., & Grolemund, G. (2016). R for data science: Import, tidy, transform, visualize, and model data. Boston: O'Reilly Media.

Zu, J., & Kyllonen, P. C. (2020). Nominal response model is useful for scoring multiple-choice situational judgement tests. *Organizational Research Methods, 23*, 342-366.