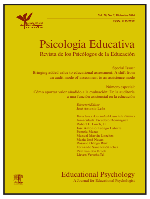
ELSEVIER

Psicología Educativa

www.elsevier.es/psed



Empirical recovery of argumentation learning progressions in scenario-based assessments of English language arts

Peter W. van Rijn[a], E. Aurora Graf[b], and Paul Deane[b]

[a]*Educational Testing Service Global, Amsterdam, The Netherlands*
[b]*Educational Testing Service, Princeton, U.S.A.*

ARTICLE INFORMATION

Manuscript received: 08/05/2014
Revision received: 28/08/2014
Accepted: 16/10/2014

Keywords:
Learning progressions
Writing assessment
Item response theory
Parallel forms

ABSTRACT

We investigate methods for studying learning progressions in English language arts using data from scenario-based assessments. Particularly, our interest lies in the empirical recovery of learning progressions in argumentation for middle school students. We collected data on three parallel assessment forms that consist of scenario-based task sets with multiple item formats, where students randomly took two of the three assessments. We fitted several item response theory models, and used model-based measures to classify students into levels of the argumentation learning progression. Although there were some differences in difficulty between parallel tasks, good agreement was found among the classifications of the parallel forms. Overall, we managed to recover empirically the order of the levels in the argumentation learning progression as they were assigned to tasks of the assessments by the theoretical framework.

Replicación empírica de las progresiones de aprendizaje de la capacidad para argumentar en una evaluación basada en escenarios de competencias de lecto-escritura

Palabras clave:
Progresiones de aprendizaje
Evaluación de la escritura
Teoría de la respuesta al ítem
Formas paralelas

RESUMEN

En este trabajo se investigan métodos para estudiar las progresiones de aprendizaje de competencias de lecto-escritura utilizando evaluaciones basadas en escenarios. En particular, nos interesa poder replicar las progresiones en el aprendizaje de la capacidad para argumentar en estudiantes de enseñanza secundaria obligatoria. Se han recogido datos aplicando tres formas paralelas de una prueba que consiste en conjuntos de tareas basadas en escenarios con preguntas de distinto formato; cada estudiante respondió a dos de estas tres formas, que fueron asignadas aleatoriamente a cada uno de ellos. Se han ajustado a los datos varios modelos de teoría de respuesta al ítem y se han utilizado medidas basadas en esta teoría para clasificar a los estudiantes en los niveles correspondientes de la progresión en el aprendizaje de la capacidad para argumentar. Aunque se han detectado algunas diferencias en las tareas de las formas paralelas, se ha encontrado un grado razonable de acuerdo entre las clasificaciones realizadas en base a las formas paralelas de la prueba. En general, se ha replicado empíricamente el orden de los niveles de la progresión en el aprendizaje de la argumentación, tal y como fueron asignados los niveles a las tareas de la prueba en el marco teórico.

*Correspondence concerning this article should be addressed to Peter W. van Rijn. ETS Global. Strawinskylaan 929. 1077XX Amsterdam, The Netherlands. E-mail: pvanrijn@etsglobal.org

DOI: http://dx.doi.org/10.1016/j.pse.2014.11.004

Learning progressions have received considerable attention over the last decade because of their intended rationale to guide student learning. They describe the change in a students' level of sophistication for key concepts, processes, strategies, practices, or habits of mind (Smith, Wiser, Anderson, & Krajcik, 2006). Examples of learning progressions can be found in many subjects, for example in English language arts (Song, Deane, Graf, & van Rijn, 2013), mathematics (Arieli-Attali, Wylie, & Bauer, 2012; Carr & Alexeev, 2011; Clements & Sarama, 2004), and science (Alonzo & Gotwals, 2012; Duschl, Maeng, & Sezen, 2011). Learning progressions are attractive in both educational theory and practice because, if valid, they can be used to report student understanding and guide subsequent instruction. However, in the absence of evaluations of instructional efficacy, their usefulness remains to be justified. Until recently most efforts around learning progressions have focused on development, although the work of Wilson (2009) on construct maps has focused on empirical validation of a developmental framework for learning progressions. In this paper, we investigate methods for studying learning progressions in English language arts (ELA). Specifically, we address the empirical recovery of argumentation learning progressions for middle school students and scrutinize the performance of three parallel computer-based assessment forms. The development process of the argumentation learning progressions that we study is described in further detail by Song et al. (2013) and in the companion paper by Deane and Song in this issue.

Our focus is on argumentation because it is a highly important skill in the language arts. For instance, it forms a key element in the US' Common Core State Standards (CCSS) for reading and writing (CCSS Initiative, 2010)[1]. In addition, it is one of the six text types in the assessment framework for reading in the Programme for International Student Assessment (PISA) since the first edition (OECD, 1999, 2013a). Argumentation is not only important in the language arts, but also in mathematics and science. To wit, argumentation and critique are considered essential in evaluating scientific claims based on data in the PISA 2015 science framework (OECD, 2013b).

Our aim is to study the empirical validation aspect of argumentation learning progressions, and we discuss an approach based on item response theory (IRT) methodology to the empirical recovery of levels in the learning progressions. A distinguishing feature of our study is that we have both data on parallel assessment forms and students taking more than one form. Therefore, our results will have considerably more impact on the validation of the learning progression than other studies in which students take only one assessment form. There are a number of studies that use similar psychometric modeling tools. Examples that have been applied in the context of learning progressions include latent class analyses (Steedle & Shavelson, 2009), Bayes nets (West et al., 2012), and Rasch and partial credit models (Black, Wilson, & Yao, 2011; Wilmot, Schoenfeld, Wilson, Champney, & Zahner, 2011). In our approach, we will discuss the relation between learning progression levels, task properties, and response time. Our task is not easy because the relation between learning progression level and task difficulty can be affected by task properties such as response format, and the assessment forms consist of a mix of selected-response (SR) and constructed-response (CR) tasks. Similarly, the relation between learning progression level and response time is expected to be different for different item types. That is, it is expected that students in higher levels of the learning progressions will likely respond faster than lower-level students to lower-level SR tasks, but probably spend more time on providing an answer to higher-level CR tasks. We have two major goals with our present research, which is conducted as part of the Cognitively Based Assessment *of*, *for*, and *as* Learning (CBAL) research initiative at Educational Testing Service (Bennett, 2010). Our first goal is to find a psychometric model that fits the data from the three parallel assessment forms and can be used to make inferences about students. Secondly, we aim to provide a method that is consistent in classifying students in the levels of the argumentation learning progression when different forms are used. To achieve our goals, we need to recover the order of the levels in the learning progression as they are assigned to tasks of the assessments by the theoretical framework (see the paper by Paul Deane & Yi Song in this issue and the discussed literature). In addition, we want to determine if the assessments forms that were designed to be parallel are in fact parallel to legitimize comparisons on the basis of tasks that are linked to particular levels in the argumentation learning progression.

Method

Three parallel CBAL assessment forms for argumentative writing were developed employing principles of evidence-centered design (Deane, Fowles, Baldwin, & Persky, 2011; Mislevy, Almond, & Lukas, 2003; Mislevy & Haertel, 2006). Each assessment form contains a different unifying scenario that includes multiple source texts and a mix of selected-response and constructed-response items. There are six tasks in each form: the first two tasks in the assessment deal mostly with summarization skills, while the remaining four tasks deal with argumentation skills. Each assessment form culminates in an extended writing task. Table 1 shows the design of the assessment, which is exactly the same for each of the three forms, with the number of items, the maximum score, the item type (selected response [SR] or constructed response [CR]), the learning progression (LP), and the LP level of each task. A more elaborate assessment design with more details on the links between tasks and argumentation learning progression levels is shown in Tables 3 and 4 in the paper by Deane and Song (this issue). The three scenarios are titled Ban Ads, Cash for Grades, and Social Networking.

Table 1
Task Mapping to two ELA Learning Progressions

Task	Description	Items	Max. score	type	Learning progression	Level
1	Evaluate summaries	9	10	SR[1]	Summarization	-
2	Write summaries	2	6	CR[2]	Summarization	-
3	Evaluate an argument	1	4	CR	Argumentation	4
4	Classify arguments	1	2	SR	Argumentation	1
5	Classify evidence	6	6	SR	Argumentation	2
6	Write an argument essay	1	10	CR	Argumentation	3

Note.[1]Selected-response; [2]constructed-response

In the first half of 2013, the three parallel argumentative writing assessments were administered to a sample of 1,840 seventh-, eighth-, and ninth-grade students from 18 schools in six different states in the US. Schools from one state volunteered to participate in the study, whereas the remainder of the schools were paid a stipend. The amount of the stipend depended on the number of students. Students randomly took two of the three parallel forms within a month and the order of administration was counterbalanced.

In order to score the CR tasks, 40 scorers, mostly teachers, were recruited and paid. Scorers received a full day of training and scored the responses online at home. Because not all responses were double-scored, we use single-scored data in our IRT analysis. Overall rater agreement statistics in the form of percentage exact agreement, percentage adjacent agreement, and quadratically weighted kappa for double-scored data will be reported.

In order to reach our first research goal, we compare several psychometric models based on IRT. To prevent local dependencies due to items being linked to a particular source text and other similarities, the unit of analysis is the task score and not the item score (Yen, 1993). However, in order to acknowledge the discrete nature of the data, we make use of IRT methods rather than factor analysis. That is, we fit several unidimensional and multidimensional IRT models, which are either simplifications or extensions of the generalized partial credit model (GPCM; Muraki, 1992; Reckase, 2009). ETS software for estimating multidimensional IRT models developed by Haberman (2013) is used. In order to compare model fit, we use Akaike's information criterion (AIC) and the Bayesian information criterion (BIC). The multidimensional IRT models that we fit all have so-called between-item multidimensionality (Adams, Wilson, & Wang, 1997). The IRT models are:

1. A unidimensional PCM
2. A unidimensional GPCM
3. A two-dimensional GPCM with dimensions for SR and CR tasks
4. A two-dimensional GPCM with dimensions for summarization and argumentation tasks.
5. A three-dimensional GPCM with separate dimensions for each scenario

With the unidimensional GPCM, we can make use of what we refer to as task progression maps to link the latent variable to levels in the learning progression. Our focus in this analysis is on the argumentation learning progression. These progression maps are based on segments of the ability scale that link task difficulty with ability (van der Schoot, 2001), and are often used in standard setting and reporting results (see e.g., Zwick, Senturk, Wang, & Loomis, 2001). The task progression maps are defined by the ability interval that runs from a 50% score on the task to an 80% score on the task under the IRT model. These values can be found by using the item response function. For example, the lower bound of the task progression map for the culminating essay task, which has a maximum possible score of ten, is the ability for which $\Pr(X = 5|\Theta)$ and the upper bound is the ability for $\Pr(X = 8|\Theta)$. This is illustrated in Figure 1. We shall refer to these ability values as P50 and P80 (note that the values can be different for different tasks). A task progression map then shows the ability range that corresponds to medium to good understanding of each task, and can be used to set cut points for assigning levels in the overall argumentation learning progression (see Deane & Song, this issue, Table 4). These progression maps are a generalization of so-called Wright maps, which have been used in the context of a learning progression in mathematical functions for college readiness (Wilmot et al., 2011). The main difference between the two is that Wright maps make use of the Rasch model (for dichotomous items) or the partial credit model (for polytomous items).

In order to reach our second research goal, we investigated if the progression maps for the same tasks in different parallel forms are similar. In addition, we checked if the order of the task progression maps is the same as the order specified by the theoretical argumentation learning progression. If the ordering was recovered and the tasks were in fact parallel in terms of the progression maps, we used the mid points of the task progression maps to classify students into levels of the argumentation learning progression. Since students take two parallel forms, we can obtain two LP classifications for each student, one for each assessment they took. We then made use of standard statistics for computing agreement among these classifications with the three parallel forms: percentage agreement (exact and adjacent) and quadratically weighted kappa. Finally, we explored the relations among learning progression levels, task properties and response time.

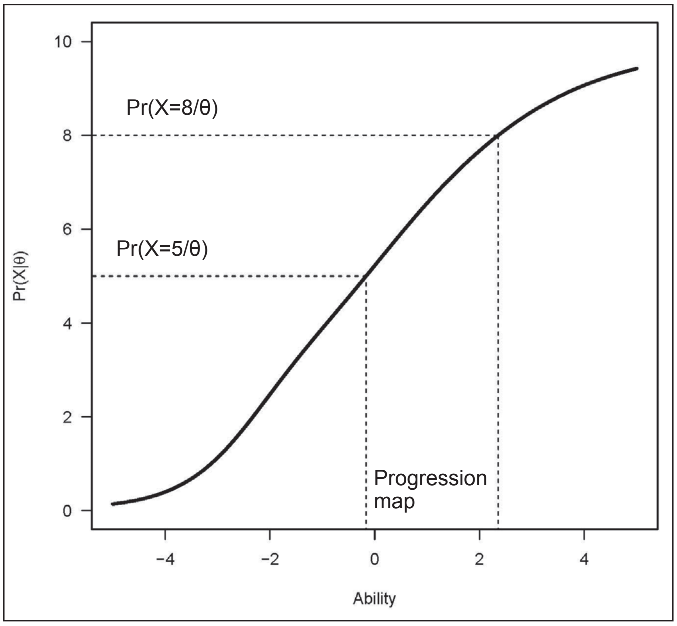


Figure 1. Example of how progression map can be derived from item response function for essay task.

Results

Students who finished at least one of the assessments were included in the sample. This resulted in a sample of 1,840 students with 47% girls, 43% boys, and 10% unreported, and with 21% seventh-, 50% eighth-, and 29% ninth-grade students.

As noted, not all data for CR items was double-scored. The average rater agreement statistics for 13,736 constucted-responses to 15 different items containing either 4, 5, or 6 categories are: 53% exact agreement, 92% adjacent agreement. The average quadratically weighted kappa is .68, indicating good agreement overall (see e.g., Altman, 1991; Fleiss, Levin, & Paik, 2003).

Descriptive statistics of the total test scores using only the data from the first rater are shown in Table 2. Test reliability (Cronbach's alpha) is computed from both item and task scores in order to assess the impact of local dependence on measurement precision (Wainer & Thissen, 1996). The reliabilities of the task scores are somewhat lower than the reliabilities of the item scores, which is an indication of local dependence. That is, if we would use the item scores as the unit of analysis instead of the task scores, the measurement precision would be inflated. For this reason, we choose the somewhat conservative, yet the most straightforward solution, which is to use the task scores as the unit of analysis in our subsequent IRT analysis. The administration design allows the computation of the correlations between the total test scores of the three assessment forms. These (Pearson) correlations are .75, .73, and .80 for Ban Ads-Cash for Grades, Ban Ads-Social Networking, and Cash for Grades-Social Networking, respectively. When corrected for attenuation using

Table 2
Descriptive Statistics of Total Test Scores

							Cronbach's α	
Assessment	Items	Tasks	Score range	*n*	Mean	*SD*	Items	Tasks
Ban ads	21	6	0-38	769	21.2	6.6	.81	.77
Cash for grades	21	6	0-38	1111	19.9	6.6	.82	.78
Social networking	21	6	0-38	1060	21.0	6.8	.83	.80

Cronbach's α of the task scores, we find correlations of .96, .92, and 1.00 for the same pairs.

Table 3 shows the relative fit measures for the five different IRT models that were fitted to the data from the three assessments forms. The two-dimensional GPCM with dimensions for item types SR and CR shows the best relative fit. The estimated correlation between these two dimensions is .89. The estimated correlation in the two-dimensional GPCM with dimensions for summarization and argumentation is .99. Although the two-dimensional GPCM SR-CR shows the best relative fit, we selected the unidimensional GPCM for further analysis because of the very high correlation in the two-dimensional model.

Table 3
Comparative Fit for Different IRT Models

Model	Dimensions	Parameters	Log-likelihood	AIC	BIC
1. PCM	1	115	-29733.5	59697	60331
2. GPCM	1	132	-29608.7	59481	60210
3. GPCM SR-CR	2	133	-29546.2	59358	60092
4. GPCM summ.-arg.	2	133	-29607.8	59482	60215
5. GPCM scenario	3	135	-29589.9	59450	60195

Figure 2 shows the task progression maps for each of the 18 tasks under the unidimensional GPCM. Each rectangle indicates the ability interval that is linked to a 50-80% task score. Although there are differences among the task progression maps for the three parallel forms, there is considerable overlap. The only task that shows no overlap among the forms is task 4 in Cash for Grades and Social Networking. The overlap of the culminating essay task (task 6) is quite large. In general, there seem to be larger differences between the three assessment forms for the SR tasks than for the CR tasks. We highlight the differences for task four. This task consists of classifying ten statements drawn from the source articles as for or against a position (see Table 1, and Deane & Song, this issue). The scoring rule is that students get zero points if they classify five or less correct, one point for six, seven, or eight correct, and two points for nine or ten correct. An inspection of each of the statements separately reveals that the ten statements in Cash for Grades are uniformly easier than the statements in the other two forms: all statements have more than 80% correct classification. In Ban Ads and Social Networking, respectively, seven and five statements have more than 80% correct classification. One statement in Ban Ads has only 54% correct classification, barely exceeding chance level. A full discussion and interpretation of results of such detailed item analysis would reveal a lot of interesting information, but surpasses the present purposes.

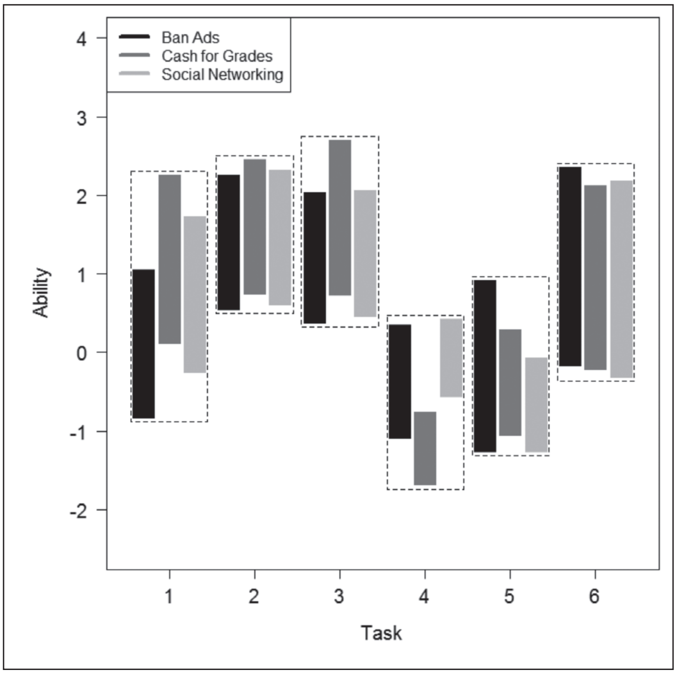


Figure 2. Progression maps using six tasks in three parallel ELA assessment forms.

Our focus is on argumentation, so only tasks 3 to 6 are used for classifying students into the levels of the associated learning progression. For each task in each assessment form, we compute the 65% score (so the P65). Then, for each task, we took the average over the three assessment forms as the transition point from one level in the argumentation learning progression to the next. Note that this is possible because each task is linked with a single level in the learning progression (see Table 1). A small correction for guessing is applied to task 5 because this task consists of six three-choice items (1/3 + 0.65 x 2/3 ≈ 0.77). Guessing is already accounted for in the scoring rule of task 4. The cut-offs for the learning progressions with this approach are -0.61, 0.14, 0.88, and 1.20 for levels one to four, respectively. Table 4 shows the agreement of the classifications among different assessment form pairs. The statistics for the three pairs are highly similar: the exact agreement is approximately 50%, the adjacent agreement is approximately 90%, and kappa is approximately .71, indicating good agreement.

Table 4
Argumentation Learning Progression Classification Agreement

		% Agreement		
Combination	*n*	Exact	Adjacent	Weighted K
Ban - Cash	398	50	90	0.70
Ban - Social	403	49	90	0.69
Cash - Social	795	50	92	0.73

We wanted to plot the argumentation learning progression level classification by grade level to inspect if students in higher grades would reach higher levels more often relative to students in lower grades. However, since most schools in the sample provided a single grade and the differences between the schools were substantial, the results are confounded (e.g., a very good school could provide only seventh grade students). Therefore, we only show the plot for the two schools that supplied students from multiple grades (see Figure 3). For these 109 seventh-grade and 102 eighth-grade students, we find that a higher percentage of students is classified in higher LP levels (two, three, and four) in eighth grade than in seventh grade. Also, a lower percentage of students is classified in lower LP levels (zero and one) in eighth grade than in seventh grade.

As a final exploratory analysis, we show box plots of response time per learning progression level for two tasks of the three assessment forms. Figure 4 shows these box plots for the classifying evidence task (5), which is a SR task. For all three forms, the mean response times on this task for the different LP levels are significantly different, $F(4, 811) = 11.02$, $p < .001$; $F(4, 1197) = 8.39$, $p < .001$; and $F(4, 1176) = 6.65$, $p < .001$, for Ban Ads, Cash for Grades, and Social Networking, respectively. It can be noted that even though there are some differences in the difficulty of task between the three assessment forms (see Figure 2), the response time pattern is the same across the three forms: students below level one respond relatively quickly. This may be an indication that these students do

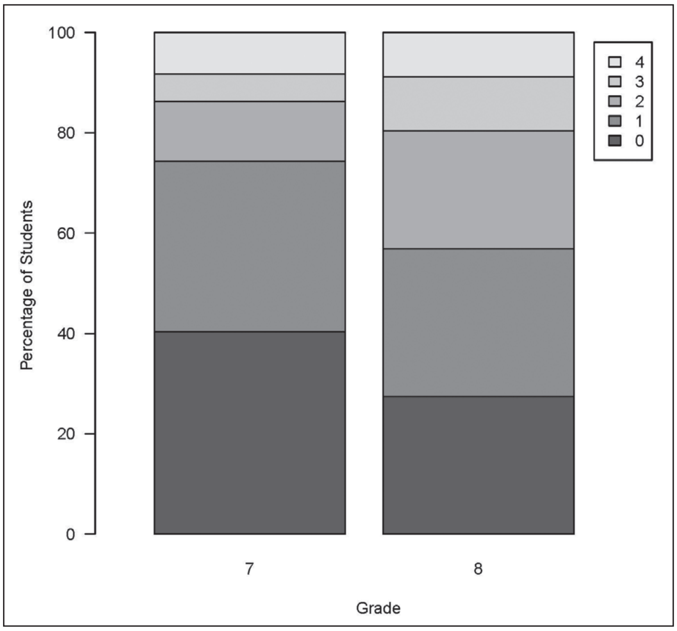


Figure 3. Argumentation learning progression classification by grade for two schools (n = 211).

not take the task very seriously. Another possibility is that these students were guessing, perhaps due to limited understanding. For levels one to four, there is a slight decrease in response time with increasing LP level. For these levels, the decreasing response times would be consistent with greater automaticity in carrying out the required cognitive operations to solve the task. A similar pattern was found for the classifying arguments task (4), although the pattern was less clear due to the larger differences in difficulty between the forms (see Figure 2).

Figure 5 shows the same box plots for the essay task (6), which is an extended CR task. For all three forms, mean response times are significantly different for different LP levels, $F(4, 811) = 53.57$, $p < .001$; $F(4, 1196) = 49.26$, $p < .001$; and $F(4, 1176) = 47.74$, $p < .001$, for Ban Ads, Cash for Grades, and Social Networking, respectively. Again, the pattern is highly similar for the three assessment forms with students at higher levels in the argumentation learning progression spending much more time on the essay than students at lower levels. For example, students at level four spent on average 35% more time on the essay than students at level two in the argumentation learning progression. Finally, the difference in patterns between Figures 4 and 5 is quite interesting. That is, the response time for the SR task goes up after level one and then slightly decreases, while for the essay the rate of increase in response time as a function of learning progression level is relatively constant. This increase is consistent with the supposition that students at higher LP levels are investing more time in planning, text generation, and editing. Almost exactly the same pattern was found in the third task ("Evaluate an argument").

Discussion

The goals of our research were to find a psychometric model for the data from the three ELA assessments focusing on argumentation and to provide a method for classifying students into levels of an argumentation learning progression. To this end, we fitted several unidimensional and multidimensional IRT models. The differences in fit between a unidimensional GPCM and a two-dimensional GPCM with separate dimensions for each item type, which showed the best fit, were relatively small. A method was proposed to use the unidimensional GPCM for classifying students into levels of the argumentation learning progression. Overall, the order of the learning

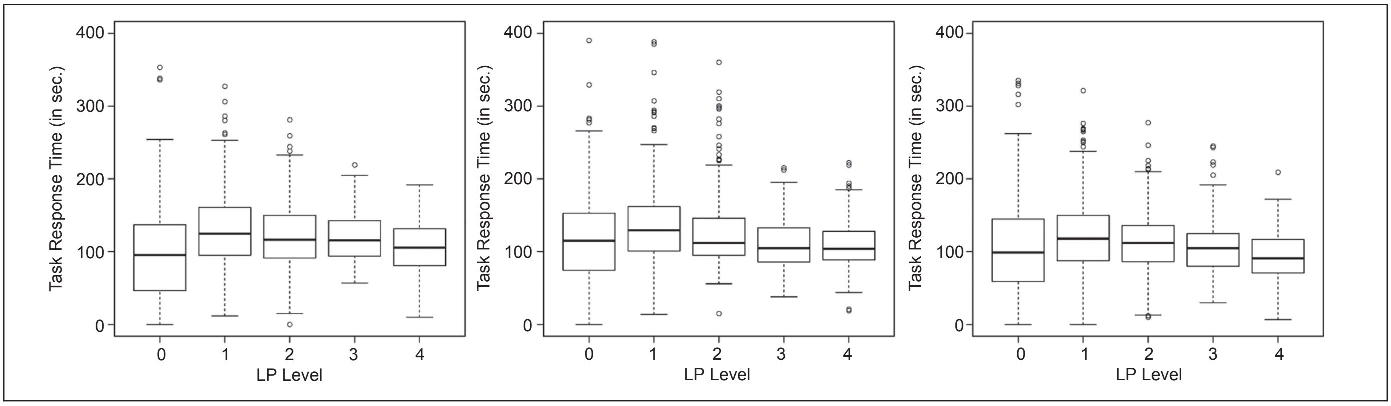


Figure 4. Box plots of response time per argumentation learning progression level for classifying evidence (task 5) of three assessment forms (left = ban ads, center = cash for grades, right = social networking).

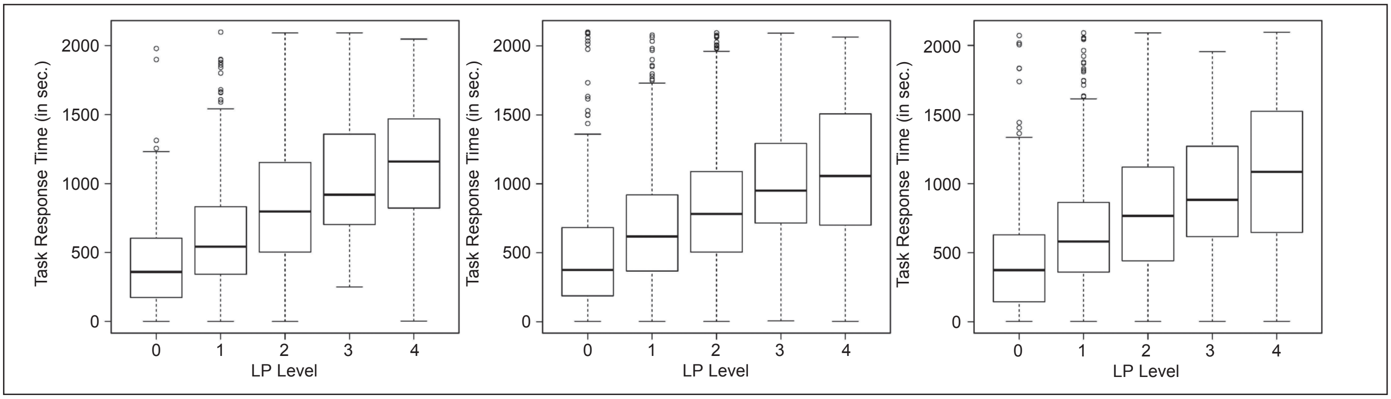


Figure 5. Box plots of response time per argumentation learning progression level for essay (task 6) of three assessment forms (left = ban ads, center = cash for grades, right = social networking).

progression levels was recovered, although there were some differences in difficulty between parallel tasks that addressed the same level in the learning progression. Since students took two of the three assessments, we were able to check the consistency of the classifications (akin to a test-retest approach). An average quadratically weighted kappa of .71 indicated a good agreement between the classifications with the different forms. Such classifications can be used by the teacher as a starting point for formative follow-up, which can include confirming the level placement.

A limitation of the method that we used for classifying students into learning progressions is that the 65-th% task score is subjective. However, it is a criterion-referenced method and our results lead to relatively consistent classifications with the parallel assessment forms given that five different levels are used. In real applications of our assessments, such cut scores would need to be validated, e.g., by means of a standard setting procedure (see e.g., Cizek & Bunch, 2007). Other limitations are that our sample is not representative and that we used single-scored data for the CR tasks. With respect to the latter, a generalizability study for these parallel assessment forms is currently underway in which person, scenario, and rater effects, as well as their interactions will be scrutinized.

The differences in difficulty between tasks that were designed to be parallel need further inspection, and revisions to tasks, items, and distractors are planned. The largest differences were found for SR tasks. It could be argued that we should expect to see more variation across performance tasks (essays) because of specific task effects such as a different topic, but we found larger differences for the SR tasks than for the CR tasks. This is not that surprising, because it is well known that specific item and distractor features can have substantial impact on item difficulty (Avalon, Meyers, Davis, & Smits, 2007; Dudycha & Carpenter, 1973; Rupp, Ferne, & Choi, 2006). For example, distractor features can strongly influence difficulty in object assembly items (Embretson & Gorin, 2001) and analytical reasoning items (Newstead, Bradon, Handley, Dennis, & Evans, 2006). As noted, a fully detailed item analysis would reveal a lot of interesting information and is planned so that the quality of the tasks can be improved upon. In this analysis, response times can also be informative in order to develop further hypotheses regarding underlying cognitive processes.

Since a learning progression ultimately represents individual development, validation of a learning progression needs to be explored in a longitudinal context. In our study, such data were available in principle, but the one month time period between the two assessments was in our opinion too short to see consistent within-student change from one level to the next.

In our future research, we aim to extend the method to classify students into multiple learning progressions with multidimensional IRT models. Although the correlation between the two dimensions in the learning progression MIRT model (model 4 in Table 3) was found to be very high (.99) in the present study, it would nevertheless be interesting to explore both compensatory and noncompensatory models (Way, Ansley, & Forsyth, 1988) to see if certain conditional relationships between different learning progressions or tasks can be detected (e.g., students need to reach level *X* in learning progression *A* before they can reach level *Y* in learning progression *B*). Such evidence would be particularly useful to guide subsequent instruction.

Resumen ampliado[2]

La progresiones de aprendizaje han recibido bastante atención durante la última década por el gran atractivo que tienen tanto para la teoría como para la práctica educativa ya que si se logra disponer de progresiones válidas se pueden utilizar no solo para informar acerca del progreso del estudiante sino como guía en el proceso de instrucción. Sin embargo, hasta la fecha la inmensa mayoría de los trabajos realizados se han centrado en el desarrollo y no en la validación de estas progresiones.

Esta investigación se ocupa justamente de la validación de las progresiones de aprendizaje formuladas para la capacidad de argumentar en el trabajo de Deane y Song en este mismo número, donde se describe un marco para diseñar evaluaciones basadas en escenarios que combina las fases de la argumentación con las progresiones de aprendizaje formuladas para dicha habilidad. En la presente investigación se trabaja con tres formas paralelas de una prueba, construida con arreglo al diseño anterior y administrada a una muestra de 1.840 estudiantes de enseñanza secundaria obligatoria.

El objetivo del trabajo es doble. Por un lado, encontrar un modelo psicométrico que se ajuste a los datos obtenidos tras administrar esas tres formas y que permita realizar inferencias acerca de los estudiantes. Por otro, proporcionar un método para clasificar a los estudiantes en el nivel correspondiente de la progresión de aprendizaje de manera consistente al trabajar con las distintas formas de la prueba. Esto es, se trata de replicar empíricamente el orden de los niveles de la progresión de aprendizaje de la argumentación, tal y como fueron asignados los niveles a las tareas de la prueba en el marco teórico. Además, queremos ver si las distintas formas de la prueba –diseñadas en principio para ser formas paralelas– lo son de verdad, para así poder garantizar la comparación en base a las tareas que están ligadas a determinados niveles en la progresión de aprendizaje de la argumentación.

Cada una de las tres formas de la prueba opera con un escenario distinto pero consta siempre de seis tareas, que incluyen preguntas con formato de elección y de construcción: las dos primeras tareas movilizan habilidades relacionadas con la capacidad de sintetizar y las otras cuatro tareas con la capacidad de argumentar. Cada estudiante respondió a dos de estas tres formas de la prueba, que fueron asignadas aleatoriamente a cada uno de ellos con un intervalo de un mes entre las dos formas aplicadas y contrabalanceando el orden de administración.

La unidad de análisis es la puntuación en la tarea y no en el ítem, por la dependencia local de las preguntas al trabajar en un contexto de evaluación basada en escenarios. Se ha trabajado en el marco de la Teoría de Respuesta al Ítem (TRI) ajustando a los datos varios modelos de crédito parcial: dos modelos unidimensionales (un modelo de crédito parcial y otro de crédito parcial generalizado), dos modelos bidimensionales (con dimensiones relativas al formato de las preguntas –elección y construcción– y al tipo de tarea demandada –síntesis y argumentación) y un modelo trimensional (con dimensiones separadas para cada escenario de trabajo); todos los modelos multidimensionales considerados fueron de crédito parcial generalizado. Aunque el modelo que presentaba los mejores índices de ajuste fue el modelo con las dimensiones relativas al formato de las preguntas (véase tabla 3), se optó por trabajar con el modelo unidimensional de crédito parcial generalizado, dado el valor tan alto obtenido para la correlación entre estas dos dimensiones (.89).

Para abordar el segundo objetivo de la investigación es preciso definir mapas de progresión para cada tarea de la prueba administrada (6 x 3 = 18). Estos mapas definen un rango de habilidad que va desde un nivel medio de conocimiento o comprensión de la tarea (la habilidad estimada que corresponde a la puntuación situada en el medio del rango posible de puntuación en dicha tarea: 5 en una tarea cuyo rango de puntuación va de 0 a 10) hasta un nivel alto (correspondiente a la puntuación que ocupa la posición 80 en el rango posible de puntuación en la tarea: 8 en el ejemplo de la tarea anterior). En particular, en este estudio se ha examinado si (1) son similares los mapas de progresión para las mismas tareas en las tres formas paralelas de la prueba y (2) el orden especificado para las tareas en el marco teórico con el que se construyó la prueba es el mismo que el orden en el que éstas aparecen con los mapas de progresión. De ser así, los intervalos de habilidad definidos en estos mapas se pueden utilizar para establecer puntos de corte que sirvan para asignar a los estudiantes a los distintos niveles de la progresión de aprendizaje de la argumentación.

El método propuesto para clasificar a los estudiantes consiste en comparar su capacidad de argumentar (estimada con el modelo unidimensional de crédito parcial generalizado) con el punto medio del intervalo definido en el mapa de progresión (la habilidad correspondiente a la puntuación que ocupa la posición 65 en el rango posible de puntuación en la tarea: 6.5 en el ejemplo de la tarea anterior). De este modo, se asigna a cada estudiante al nivel de la progresión de aprendizaje que señala la media de la habilidad correspondiente al punto anterior (posición 65) obtenida en la tarea en cuestión en las tres formas de la prueba: este valor medio constituye el punto de transición de un nivel a otro en la progresión de aprendizaje. Dado que cada estudiante responde a dos formas de la prueba, se obtienen para cada uno de ellos dos clasificaciones diferentes en la progresión de aprendizaje.

Los resultados obtenidos muestran que hay un solapamiento importante en los intervalos definidos para las mismas tareas en las 3 formas de la prueba, si bien hay algunas diferencias (véase figura 2); en particular, en la tarea 4 hay diferencias importantes entre dos formas de la prueba y se observan más diferencias entre las tres formas para las tareas con formato de elección (1, 4 y 5) que de construcción (2, 3 y 6). Por otro lado, el orden especificado por el marco teórico para las tareas (véase tabla 1) es replicado empíricamente en los mapas de progresión y se ha encontrado también un notable grado de acuerdo (véase tabla 4) entre las clasificaciones de los estudiantes en los niveles de la progresión de aprendizaje obtenidas en base a las distintas formas de la prueba administradas.

En resumen, en el presente trabajo se analiza una estrategia basada en la TRI para replicar empíricamente los niveles de una progresión de aprendizaje y se propone un método que sirve para clasificar de forma consistente a los estudiantes en los niveles de la progresión de aprendizaje de la capacidad para argumentar, estimando dicha capacidad con el modelo de crédito parcial generalizado. El profesor puede utilizar estas clasificaciones para confirmar el nivel o punto de partida de los estudiantes y para su posterior seguimiento formativo.

Conflict of Interest

The authors of this article declare no conflict of interest.

Notes

[1] As an example, in eighth grade writing, standard CCSS.ELA-LITERACY.W.8.1 states: "Write arguments to support claims with clear reasons and relevant evidence" (Common Core State Standards Initiative, 2010).

[2] Este resumen ha sido realizado por la editora del número, María José Navas.

References

Adams, R., Wilson, M., & Wang, W.-C. (1997). The multdimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*, 1-23.

Alonzo, A., & Gotwals, A. (Eds.) (2012). *Learning progressions in science: Current challenges and future directions*. Rotterdam, The Netherlands: Sense.

Altman, D. G. (1991). *Practical statistics for medical research*. London, England: Chapman and Hall.

Arieli-Attali, M., Wylie, E. C., & Bauer, M. I. (2012, April). *The use of three learning progressions in supporting formative assessment in middle school mathematics.* Paper presented at the annual meeting of the American Educational Research Association (AERA), Vancouver, CA.

Avalon, M. E., Meyers, L. S., Davis, B. W., & Smits, N. (2007). Distractor similarity and item-stem structure: Effects on item difficulty. *Applied Measurement in Education*, *20*, 153-170.

Bennett, R. (2010). Cognitively based assessment of, for, and as learning (CBAL): A preliminary theory of action for summative and formative assessment. *Measurement*, *8*, 70-91.

Black, P., Wilson, M., & Yao, S.-Y. (2011). Road maps for learning: A guide to navigation of learning progressions. *Measurement: Interdisciplinary Research and Perspectives*, *9*, 71-123.

Carr, M., & Alexeev, N. (2011). Fluency, accuracy, and gender predict developmental trajectories of arithmetic strategies. *Journal of Educational Psychology*, *103*, 617-631.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks, CA: Sage.

Clements, D. H., & Sarama, J. (2004). Learning trajectories in mathematics education. *Mathematical Thinking and Learning*, *6*, 81-89.

Common Core State Standards Initiative (2010). *Common core state standards for Englishlanguage arts and literacy in history/social studies, science, and technical subjects.* Retrieved from http://www.corestandards.org/ELA-Literacy/

Deane, P., Fowles, M., Baldwin, D., & Persky, H. (2011). *The CBAL summative writing assessment: A draft eighth-grade design* (Research Memorandum 11-01). Princeton, NJ: Educational Testing Service.

Deane, P., & Song, Y. (2014). A case study in principled assessment design: Designing assessments to measure and support the development of argumentative reading and writing skills. *Psicología Educativa, 20*, 99-108.

Dudycha, A. L., & Carpenter, J. B. (1973). Effects of item format on item discrimination and difficulty. *Journal of Applied Psychology*, *58*, 116-121.

Duschl, R., Maeng, S., & Sezen, A. (2011). Learning progressions and teaching sequences: A review and analysis. *Studies in Science Education*, *47*, 123-182.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement*, *38*, 343-368.

Fleiss, J. L., Levin, B., & Paik, M. (2003). *Statistical methods for rates and proportions* (3rd ed.). New York: Wiley.

Haberman, S. (2013). *A general program for item-response analysis that employs the stabilized Newton-Raphson algorithm* (ETS Research Report 13-32). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., Almond, R. G., & Lukas, J. F. (2003). *A brief introduction to evidence centered design* (Research Report 03-16). Princeton, NJ: Educational Testing Service.

Mislevy, R. J., & Haertel, G. D. (2006). Implications of evidence-centered design for educational testing. *Educational Measurement: Issues and Practice*, *25*(4), 6-20.

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

Newstead, S. E., Bradon, P., Handley, S. J., Dennis, I., & Evans, J. S. B. T. (2006). Predicting the difficulty of complex logical reasoning problems. *Thinking & Reasoning*, *12*, 62-90.

OECD (1999). *Measuring student knowledge and skill. a new framework for assessment.* Retrieved from http://www.oecd.org/edu/school/programmeforinternationalstudentassessmentpisa/33693997.pdf

OECD (2013a, March). *PISA 2015: Draft reading literacy framework.* Retrieved from http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Reading%20Framework%20.pdf

OECD (2013b, March). *PISA 2015: Draft science framework.* Retrieved from http://www.oecd.org/pisa/pisaproducts/Draft%20PISA%202015%20Science%20Framework%20.pdf

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer.

Rupp, A. A., Ferne, T., & Choi, H. (2006). How assessing reading comprehension with multiple-choice questions shapes the construct: A cognitive processing perspective. *Language Testing*, *23*, 441-474.

Smith, C., Wiser, M., Anderson, C., & Krajcik, J. (2006). Implications of research on children's learning for standards and assessment: A proposed learning progression for matter and the atomic-molecular theory. *Measurement: Interdisciplinary Research and Perspectives*, *4*, 1-98.

Song, Y., Deane, P., Graf, E. A., & van Rijn, P. W. (2013). *Using argumentation learning progressions to support teaching and assessments of English language arts* (R & D Connections No. 22). Princeton, NJ: Educational Testing Service.

Steedle, J., & Shavelson, R. (2009). Supporting valid interpretations of learning progression level diagnoses. *Journal of Research in Science Teaching*, *46*, 669-715.

Van der Schoot, F. C. J. A. (2001, April). *The application of an IRT-based method for standard setting in a three-stage procedure.* Paper presented at the annual meeting of the National Council of Measurement in Education (NCME), New Orleans, LA.

Wainer, H., & Thissen, D. (1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability? *Educational Measurement: Issues and Practice*, *15*(1), 22-29.

Way, W., Ansley, T. N., & Forsyth, R. A. (1988). The comparative effects of compensatory and noncompensatory two-dimensional data on unidimensional IRT estimates. *Applied Psychological Measurement*, *12*, 239-252.

West, P., Rutstein, D., Mislevy, R., Liu, J., Levy, R., DiCerbo, K., … Behrens, J. (2012). A Bayesian network approach to modeling learning progressions. In A. Alonzo & A. Gotwals (Eds.), *Learning progressions in science: Current challenges and future directions* (p. 257-292). Rotterdam, The Netherlands: Sense.

Wilmot, D., Schoenfeld, A., Wilson, M., Champney, D., & Zahner, W. (2011). Validating a learning progression in mathematical functions for college readiness. *Mathematical Thinking and Learning*, *1*, 259-291.

Wilson, M. (2009). Measuring progressions: Assessment structures underlying a learning progression. *Journal of Research in Science Teaching*, *46*, 716-730.

Yen, W. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, *30*, 187-213.

Zwick, R., Senturk, D., Wang, J., & Loomis, S. C. (2001). An investigation of alternative methods for item mapping in the National Assessment of Educational Progress. *Educational Measurement: Issues and Practice*, *20*(2), 15-25.