Item Response Theory and Computer-based Testing

1 March 2018

Seminar/Proseminar: Intelligente Lehr-/Lernsysteme
Rainier Robles, WS 17/18

Hintergrundinformationen

Name: Rainier Raymond Robles
Matrikelnummer: 5002400

(a) Studienordnung: Bachelor Informatik (150 LP, Studienordnung 2014)
(b) Modul: Wissenschaftliches Arbeiten in der Informatik 0086cA6.1
(c) Modulprüfung: Vortrag (ca. 30 Minuten) mit anschließender Diskussion (ca. 10 Minuten)
Formen aktiver Teilnahme: Schriftliche Ausarbeitung, Teilnahme an den Diskussionen zum Vortrag

Structure

I. Motivation
II. Computer Adaptive Tests
III. Item Response Theory
IV. Case Study: SIETTE
V. Issues and Challenges
VI. Conclusions
VII. References

Motivation

Motivation

"...to go back to a time when teachers could afford to assess each student orally by asking him/her a few well-selected questions..." (Conejo et al. 2004)

Computer Adaptive Tests

Computer Adaptive Tests (CATs)

"The basic notion of an adaptive test is to mimic automatically what a wise examiner would do." (Wainer and Mislevy, qtd. in Conejo et al.)

Definition: CATs are tests administered by a computer, where the selection of the items as well as the decision to terminate the test are based on the examinee's responses.

Advantages of CATs

- Significantly shorter tests, leading to significantly shorter testing times
- More accurate estimation of test takers' proficiency levels
- Test takers' motivation improved
- Large item pools can be stored
- Multimedia content is supported

CATs: Algorithm

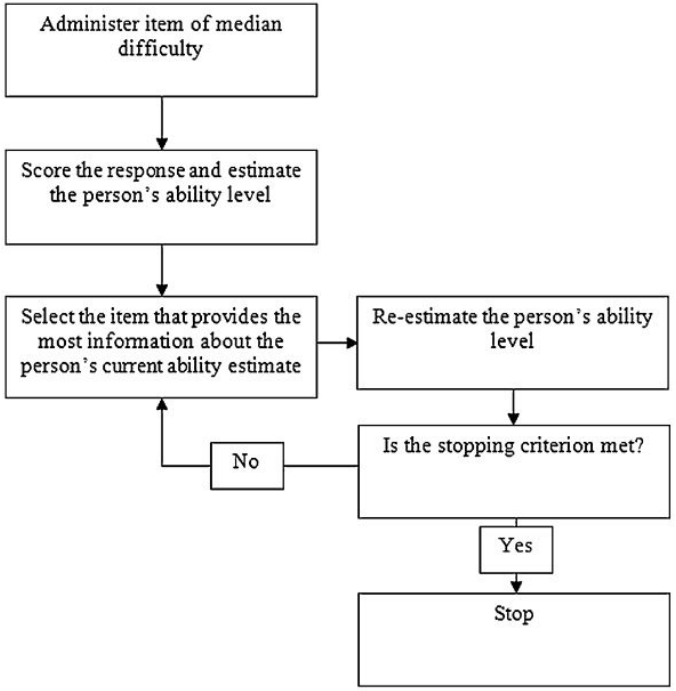


Fig. 1. CAT algorithm flowchart. (Wauters et al. 2010)

Basic Elements in the Development of a CAT

- Response model
- Item pool
- Input proficiency level
- Item selection method
- Termination criterion

Item Response Theory

Classical Test Theory vs Item Response Theory

Freie Universität Berlin

Classical Test Theory

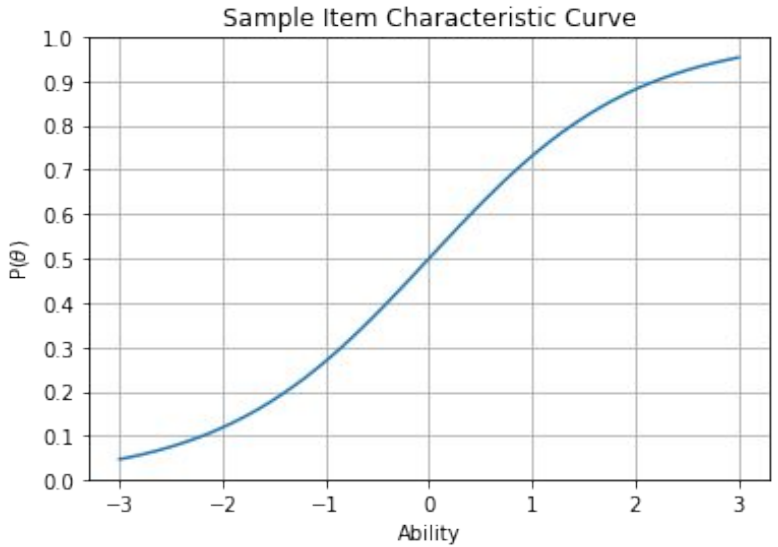
- Focus on aggregation of test responses as a total test score
- Generally produces a standardized achievement level
- Greater possibility to use free-response items in a test

Item Response Theory

- Focus on individual responses to each test item
- Produces a probability of a test taker's proficiency level
- Mostly composed of multiple-choice items

Item Characteristic Curve (ICC)

- "(T)he basic building block of item response theory." (Baker 2001)
- Plots probability that a test taker on a particular ability level can answer the question correctly
- Each item has its own ICC



Technical Properties of ICC

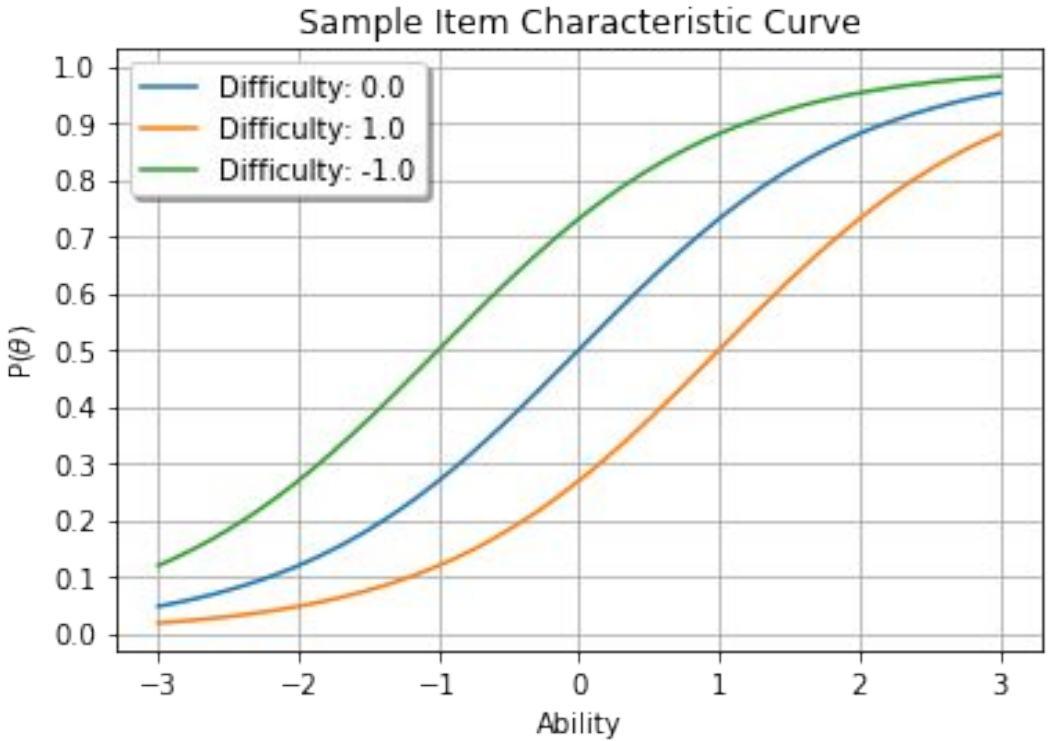
1. Item difficulty

   Describes where an item "functions"; a location index.

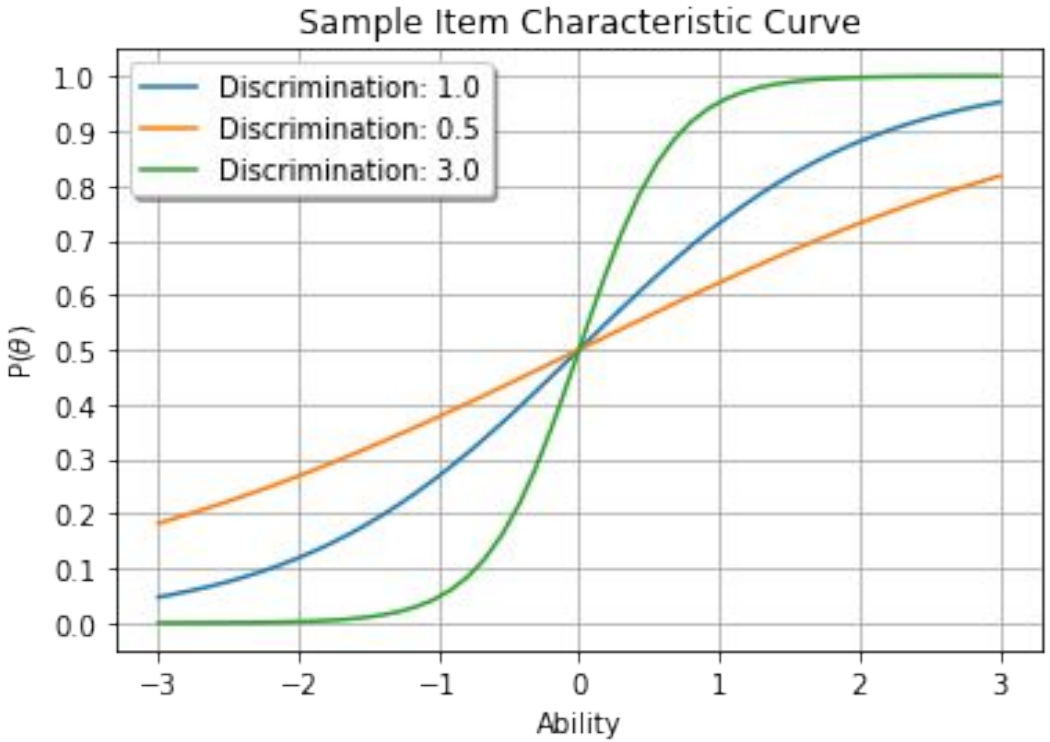
2. Discrimination

   Describes how well an item discriminates among test takers with abilities above and below the item's difficulty.

Item Difficulty



Discrimination



ICC Models

1. Two-Parameter Logistic Model

2. Rasch / One-Parameter Logistic Model

3. Three-Parameter (Logistic) Model

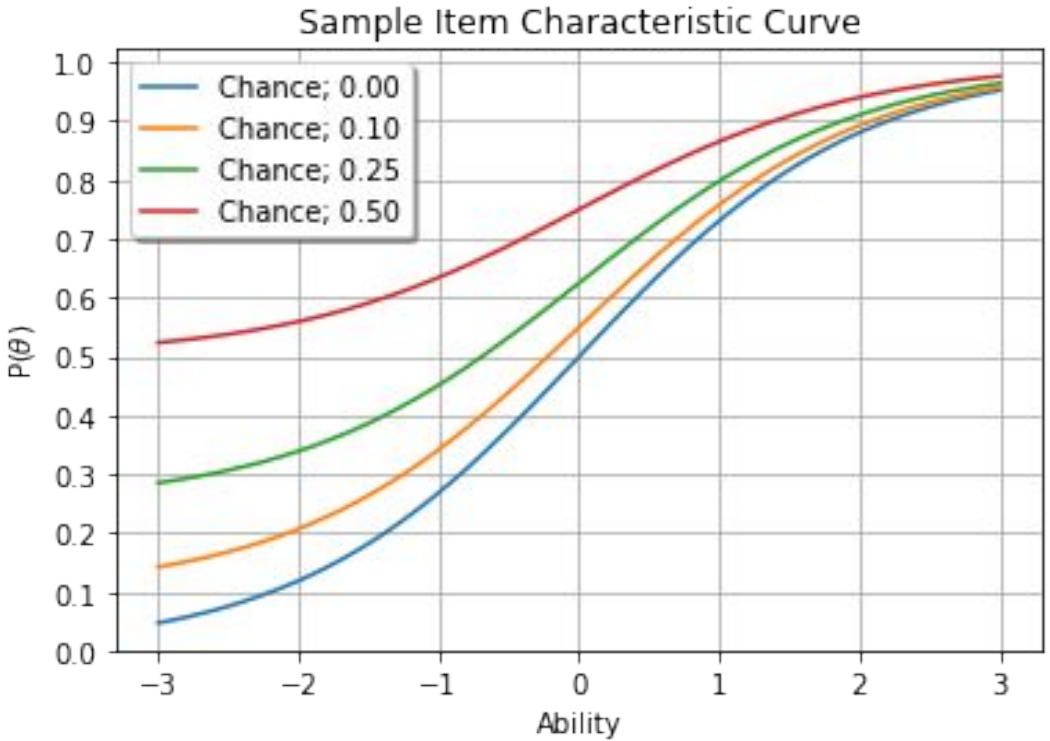
Three-Parameter Logistic Model

$$P(\theta) = c + (1 - c)\frac{1}{1 + e^{-a(\theta - b)}}$$

Where:

- b = difficulty
- a = discrimination
- c = guessing parameter
- θ = ability level

Three-Parameter Logistic Model



Case Study:
SIETTE

SIETTE

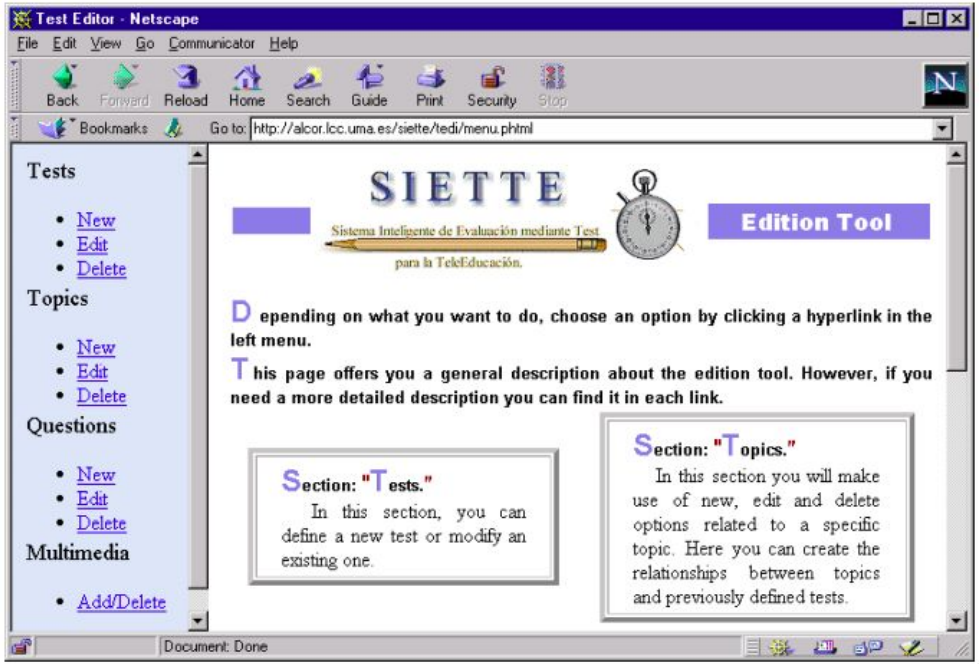


Fig. 2. SIETTE test editor homepage. (Conejo, et al. 2004)

A Brief Overview of SIETTE

Freie Universität Berlin

- **SI**steme de **E**valuación de **T**ests para la **T**ele**E**ducación *(Intelligent Evaluation System using Tests for TeleEducation)*
- Initially developed in 1998 by Antonia Rios as part of a master thesis
  - Department of Computer Science, University of Malaga, Spain
- Started as an Intelligent Tutoring System (ITS) module as part of the TREE (**TR**aining of **E**uropean **E**nvironmental Trainers and Technicians) project

A Brief Overview of SIETTE

- Designed to be a reusable, domain-independent module
- Teachers can design and develop the parameters of the tests and items
- Students can take automatically generated adaptive tests
- Uses discrete ability levels

Student Model

STUDENTS					
StudentID	TestID	Date	Level of Proficiency	Lower Confidence Level	Upper Confidence Level
John Smith	TREE	04/08/98	Level 1	0.9	1.1

KNOWLEDGE DISTRIBUTION					
StudentID	TestID	Level 0	Level 1	...	Level 10
John Smith	TREE	0.001	0.9	...	0.001

TOPIC DISTRIBUTION			
StudentID	TestID	TopicID	% Questions
John Smith	TREE	PINUS	40%
John Smith	TREE	ABIES	40%
John Smith	TREE	CEDRUS	20%

QUESTIONS POSED		
StudentID	QuestionID	AnswerID
John Smith	Q_1	$A_{1,1}$
John Smith	Q_3	$A_{3,2}$
John Smith	Q_5	$A_{5,1}$
John Smith	Q_6	$A_{6,3}$
John Smith	Q_8	$A_{8,4}$
John Smith	Q_{10}	$A_{10,5}$
John Smith	Q_{12}	$A_{12,3}$
John Smith	Q_{14}	$A_{14,1}$
John Smith	Q_{17}	$A_{17,2}$
John Smith	Q_{20}	$A_{20,1}$

Fig. 3. Example of a temporary student model in SIETTE. (Conejo, et al. 2004)

Student State Before Answering

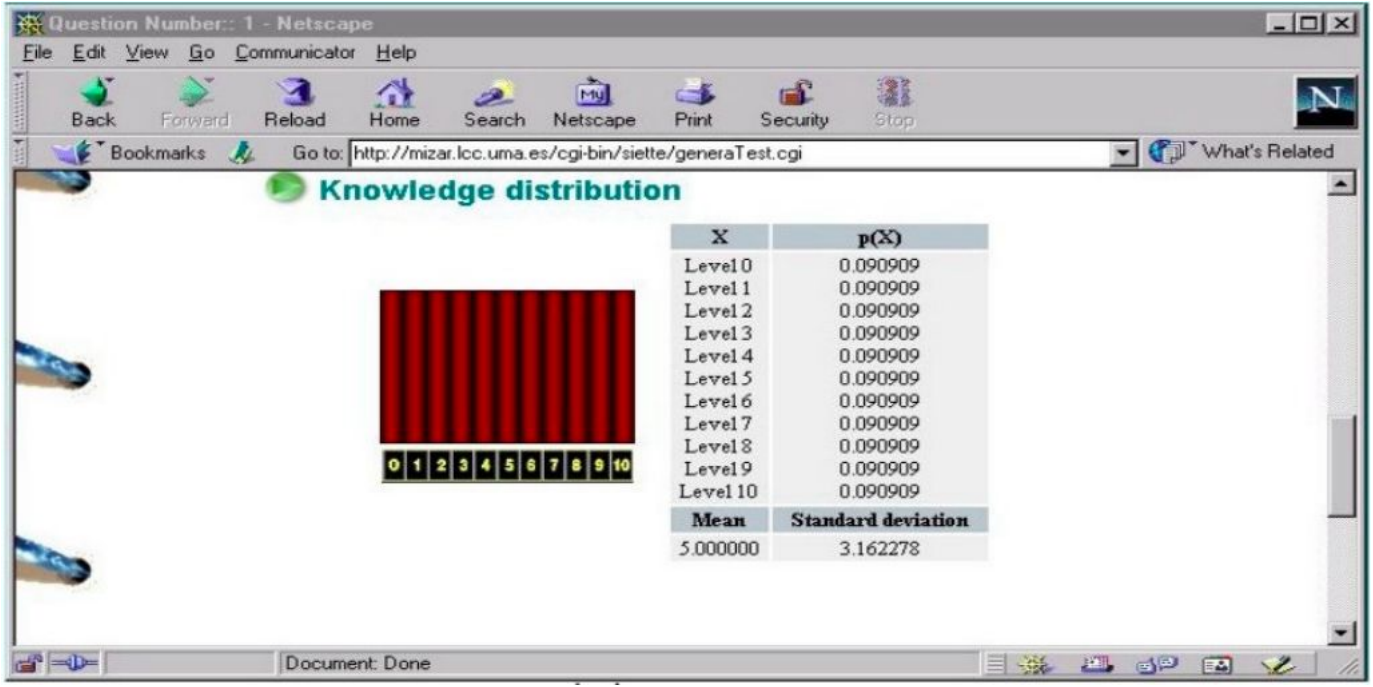


Fig. 4. Page showing estimated student levels before the first question. (Conejo, et al. 2004)

Item Characteristic Curve in SIETTE

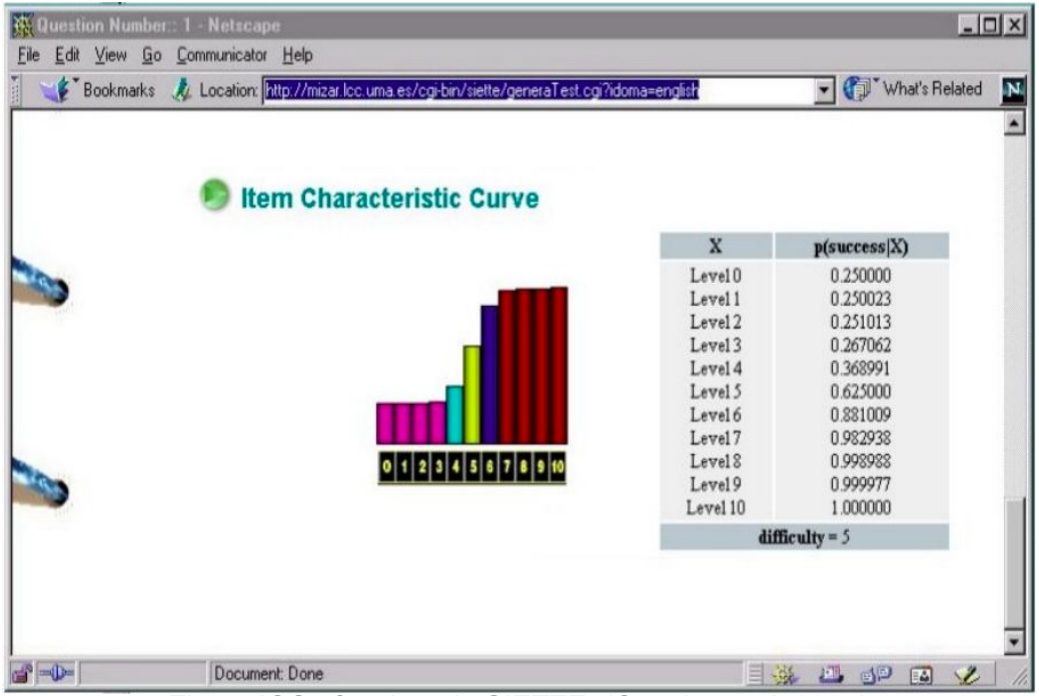
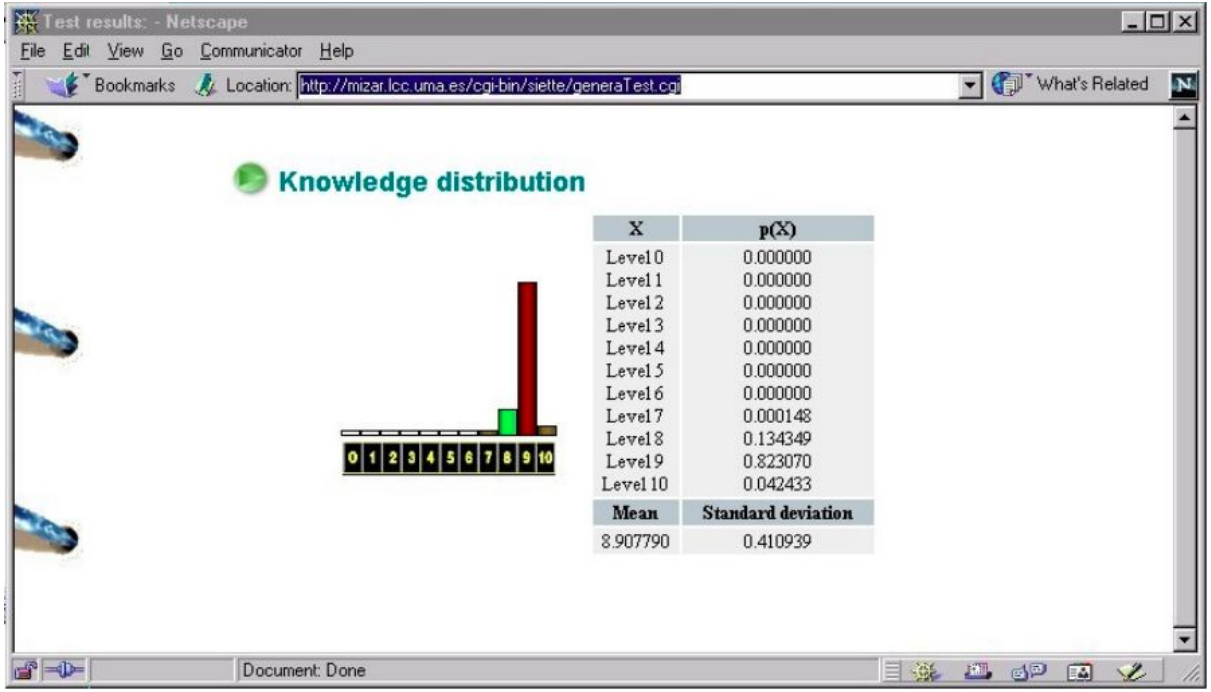


Fig. 5. ICC of an item in SIETTE. (Conejo, et al. 2004)

Knowledge Distribution at the End of a Test



X	p(X)
Level 0	0.000000
Level 1	0.000000
Level 2	0.000000
Level 3	0.000000
Level 4	0.000000
Level 5	0.000000
Level 6	0.000000
Level 7	0.000148
Level 8	0.134349
Level 9	0.823070
Level 10	0.042433
Mean	**Standard deviation**
8.907790	0.410939

Fig. 6. Student state after finishing a test. (Conejo, et al. 2004)

Evaluations Performed on SIETTE

- Evaluation with simulated students
  - Number of classes K and confidence factor ρ
  - Discrimination a and guessing factor c
  - Item selection method (Bayesian vs. difficulty-based vs. random)

- Evaluation with real students

Evaluation with Simulated Students

- Setup: N randomly generated students for each discrete ability level in [0,K-1]
- Assumption: Items are correctly calibrated
- Test Termination: Probability of student belonging to a particular ability level reaches ρ

Experiment 1

- Evaluated: Number of classes K and confidence factor ρ

- Results:
  - Impossible to classify all test takers correctly
  - Percentage of correctly classified test takers dependent on ρ
  - Number of questions in the test strongly correlated to K

Experiment 1

	Confidence factor $\rho = 0.75$		*Confidence factor* $\rho = 0.90$		*Confidence factor* $\rho = 0.99$	
Number of classes K	*% of correctly classified students*	*Average number of questions posed T*	*% of correctly classified students*	*Average number of questions posed T*	*% of correctly classified students*	*Average number of questions posed T*
3	84.05	2.00	95.82	3.58	99.46	5.65
5	81.61	6.23	92.76	10.38	99.37	19.27
7	80.96	11.11	92.85	18.16	99.38	33.12
9	80.86	16.15	92.93	26.39	99.42	47.27
11	80.52	21.19	92.92	34.54	99.26	60.85

Table 1. Experiment 1 results. (Conejo, et al. 2004)

Experiment 2

- Evaluated: Discrimination a and guessing factor c

- Results:
  - Best results from items with higher a and lower c
  - a does not influence number of questions past a certain point
  - Extremely high number of questions needed for lower a

Experiment 2

Guessing factor c	*% of correctly classified students*	*Average number of questions posed T*
0.00	92.85	18.16
0.10	92.37	25.34
0.25	92.11	36.05
0.33	91.73	43.37
0.50	91.49	63.37

Discrimination index a	*% of correctly classified students*	*Average number of questions posed*
0.20	90.4	174.9
0.50	91.5	35.2
0.70	91.9	26.3
1.20	92.8	18.1
1.70	93.8	15.3
2.20	95.4	14.8

Table 2. Experiment 2 results. (Conejo, et al. 2004)

Experiment 3

- Evaluated: Item selection method (Bayesian vs. difficulty-based vs. random)

- Results:
  - Adaptive methods (Bayesian and difficulty-based) performed better than random selection
  - Higher value of K showed even better performance of adaptive methods
  - Ultimately difficulty-based selection was chosen because of lower computational cost

Experiment 3

	Bayesian		*Difficulty-based*		*Random*	
Number of classes K	*% of correctly classified students*	*Average number of questions posed T*	*% of correctly classified students*	*Average number of questions posed T*	*% of correctly classified students*	*Average number of questions posed T*
3	96.06	3.58	95.62	3.58	95.82	3.58
5	93.31	6.87	94.67	7.37	92.76	10.38
7	92.75	8.70	94.43	9.03	92.85	18.16
9	92.53	9.85	94.23	10.14	92.93	26.39
11	92.10	10.71	94.14	11.02	92.92	34.54

Table 3. Experiment 3 results. (Conejo, et al. 2004)

Evaluation with Real Students

- Informal evaluation study in June 2000
- 24 Computer Science students of the subject Artificial Intelligence at the University of Malaga, Spain
- Students had an exam in about four weeks
- Students took an online test about LISP then filled out a questionnaire
- Participation was voluntary and anonymous

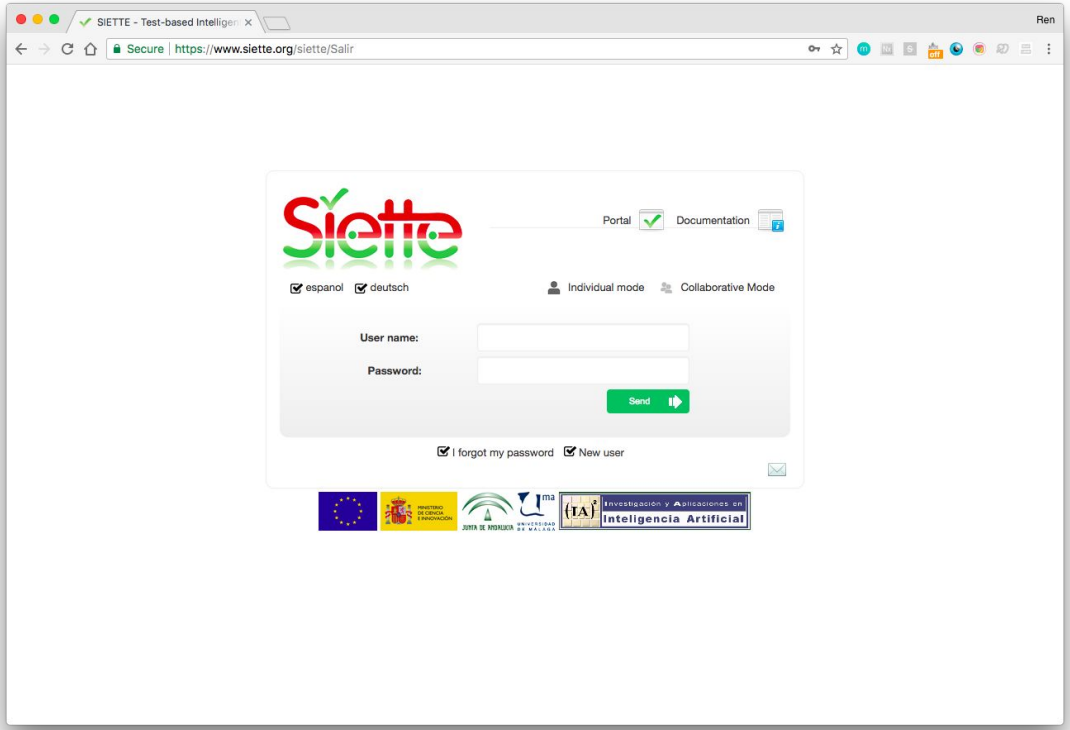
Results of Evaluation 2

- Majority needed less than five minutes to use the system
- Majority considered SIETTE useful
- Half said that they would use the tool again, the rest were either unsure or said that they would not
- Majority would recommend it to other students

Results of Evaluation 2

- Majority considered the grade they received right, fair, or totally fair
- Majority considered difficulty of the test normal
- Tie between pen-and-paper test and SIETTE test, mainly because they would rather wait for SIETTE to be fully developed and tested

Updates on SIETTE



Updates on SIETTE

- SIETTE can be integrated with Moodle LMS
- No longer just a test-based system but an automatic assessment environment
- Includes more question forms such as short answer questions
- Includes other assessment methods including traditional percentage scoring
- Only 18 tests used IRT, compared to 642 which used percentage scoring and 581 which used item scoring

Issues and Challenges

Issues and Challenges

- General resistance to change
- Difficulty in understanding underlying concepts
- Item calibration

Conclusions

Conclusions

- Item Response Theory and Computer Adaptive Tests are a valid alternative to Classical Test Theory and traditional pen-and-paper testing.
- The use of IRT and CAT can reduce test length and therefore test time.
- Members of every level of the education system have to be educated about IRT and CAT before its use is more widely accepted.
- More effective and efficient ways of item difficulty estimation and item calibration should be studied and implemented.

References

References

Baker, Frank B. *The Basics of Item Response Theory*. 2nd ed., ERIC Clearinghouse on Assessment and Evaluation, 2001, 1-57.

Barla, Michal, et al. "On the impact of adaptive test question selection for learning efficiency." *Computers & Education* 55.2 (2010): 846-857.

Conejo, Ricardo, et al. "SIETTE: A web-based tool for adaptive testing." *International Journal of Artificial Intelligence in Education* 14.1 (2004): 29-61.

Conejo, Ricardo, Eduardo Guzmán, and Monica Trella. "The SIETTE automatic assessment environment." *International Journal of Artificial Intelligence in Education* 26.1 (2016): 270-292.

Thissen, David. "Bad questions: An essay involving item response theory." *Journal of Educational and Behavioral Statistics* 41.1 (2016): 81-89.

Wauters, Kelly, Piet Desmet, and Wim Van Den Noortgate. "Adaptive item-based learning environments based on the item response theory: possibilities and challenges." *Journal of Computer Assisted Learning* 26.6 (2010): 549-562.

Thank You!