УДК 378:372'811.131.1

SOME CRITICAL ISSUES IN TESTS TAKEN FROM CELI AND CILS CERTIFICATIONS (Italian as a Foreign Language)

Paolo Torresan (Catania University)

This paper aims to sum up 3 years of research into the CELI and CILS certifications in Italian as a Foreign Language, the results of which have been reported in several articles and essays. We have discovered some sources of bias affecting the quality of the exams. Here, these areas of bias are briefly described and classified into groups.

Key-words: Italian as a Foreign Language, Item Analysis, Language Testing, Validity.

Паоло Торресан. Проблеми відхилень в результатах тестування CELI і CILS (італійська мова як іноземна). Ця стаття підводить підсумки трьох років досліджень в сертифікації CELI і CILS (італійська мова як іноземна), результати яких були описані у низці статей і есе. Ми виявили деякі джерела відхилень, що впливають на якість іспитів. Дані області відхилень коротко описані та класифіковані в групи.

Ключові слова: італійська мова як іноземна, аналіз товару, мова тестування, дії.

Паоло Торресан. Проблемы отклонений в результатах тестирования CELI и CILS (итальянский язык как иностранный). Эта статья подводит итоги трех лет исследований в сертификации CELI и CILS (итальянский язык как иностранный), результаты которых были описаны в ряде статей и эссе. Мы обнаружили некоторые источники отклонений, влияющие на качество экзаменов. Данные области отклонений кратко описаны и классифицированы в группы.

Ключевые слова: итальянский язык как иностранный, анализ товара, язык тестирования, действия.

1. CELI and CILS

CELI and CILS are certifications in Italian as a Foreign Language designed by teams of item writers at the CVCL Center of the University for Foreigners in Perugia and the CILS Center at the University for Foreigners in Siena, respectively. These certifications are well known in the context of Italian as a Foreign Language. Various institutions make use of them, such as Italian Cultural Institutes, universities and schools in general. Alongside the PLIDA certification, offered by the Dante Alighieri Society and the .IT certification, offered by Roma Tre University, the CELI and CILS certifications are included in the CLIQ Project (*Certificazione Lingua Italiana di Qualità*, Italian Language Quality Certification), recognised by the Italian Ministry of Foreign Affairs in 2013 and intended to guarantee shared quality standards.

2. Critical issues

In the following paragraphs, we will consider some critical issuesidentified through the analysis of certain tests used in both certifications. Most of these tests are available online (<www.cils.unistrasi.it>; <www.cvcl.it>).

These issues refer to:

- tasks (§ 2.1.)
- texts (§ 2.2.)
- items (§ 2.3.)
- calibration (§ 2.4.)
- layout (§ 2.5.)
- score (§ 2.6.)
- keys (§ 2.7.)
- instructions (§ 2.8.)
- prompts (§ 2.9.)

We will deal with each of these issues in the following sections in detail.

We will adopt the following acronyms throughout the article:

TT: test taker

EN: educated native (cfr. Hulstijn, 2011; Mulder, Hulstijn, 2011)

TF: true and false exercise

MCQ: multiple choice test

OT: original text

MV: modified version (of the source text).

2.1. Tasks

First, let us consider two tasks used in CILS certification.Both of them raise critical validity issues.

In CILS A1 for teenagers, May 2012 session, there is a **gap-filling exercisedesigned to assess listening skills**.

They encounter certain issues:

- the students are provided with a written text, which they are supposedto fill in by listening to the complete recorded audio version of the same text. In a survey we carried outon 94teenage Swiss students (A1 proficiency level),we discovered that many pupils were able to fill in the cloze text without referring to the audio text, by simply reading the incomplete script. There is thus evidence of a construct-irrelevant variance. Students proficient at reading are better equipped to complete the test (Torresan 2014a).
- By virtue of this,for any dictation-type text, it is unclear whether,in completing the task, students are relying on a simple recognition and decoding strategy (low-order skill) rather than carrying out a broader interpretation of the whole text (high-order skill). It is also true that anyone understanding the meaning of the passage may miswrite the target word (Buck 2001). In any dictation-style test, such a construct under-representation factor may impact upon the student's performance.

The **CILS jigsaw text** comprises many sentences which the TT has to connect together, paying attention to the chronological order of the events and the linking devices.

We argue that this task (Torresan 2015a):

- involvesa cognitive overload due to the over-fragmentation of the text
- is unfamiliar to the TTs (the super-jigsaw version)
- is biased in favor of analytic/cognitive learning styles (and the logical-intelligence student type) rather thanholistic/impulsive learning styles

The **CELI summary test**, aimed at assessing writing skills foradvanced candidates, presents these problems:

- theinstructions are not clear (cfr. § 2.8.)
- it is a complex and integrated task, so how can we be sure to assess only writing and not reading as well?[1] A poor performance in writing a summary may in fact also depend on poor reading skills. In the CELI exam, the item writer tries to avoid producing any construct-irrelevant variance factor by providing a list of points that the TT has to follow. We believe that this may even complicate the situation: what if the reader does not agree with the selection made by the item writer? After all, even NEs may not be in agreement regarding the main points (Alderson *et al.* 1995: 61).Moreover,there a pedagogical issue to be considered:is this conception of writing, as a reproduction of what is expected to be written by an external 'authority', an authentic task? Or, is the reality that, when writing a report/summary,most writers re-organize the meaning conveyed by a text according to their own criteria? Could an over-scaffolded task, such assummarizing a text based on a list to be followed, biased againstholistic-learning style students?

Both certifications feature**cloze exercises** aimed at testing morphosyntactic, textual or lexical competencies among intermediate (CELI) and intermediate/advanced students (CILS).

Cloze tests have an intrinsic limitation: theyinvite localized readings confined to specific sentences or clauses (Alderson 1979; Porter 1983; cfr. fig. 1)[2].

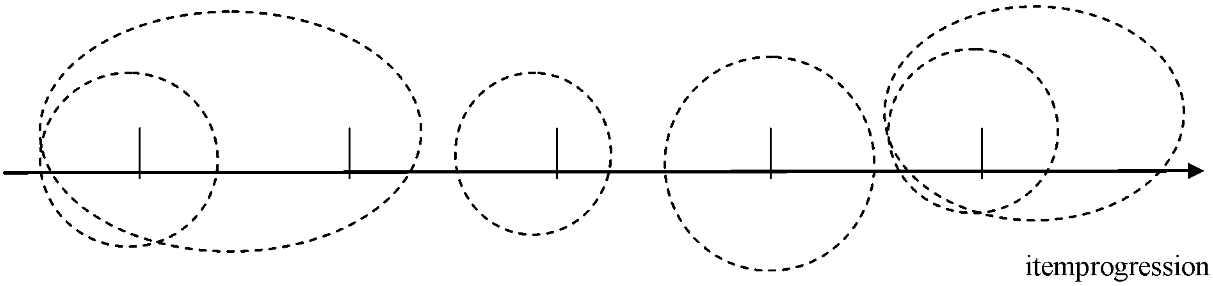

Fig. 1. Localized reading processing while completing a cloze

This means that when we have a clause forcing the reader to return to what was said a few lines before or at the beginning of the text, even the EN reader may (Torresan 2015b):

- get stuck and fail to respond
- force the interpretation of the clause to fit the hypotheses made thus far, adapting it in line with the meaning developed up until that point

Moreover, many types of closes from both certifications are two-dimensional, testing both morphology and vocabulary at the same time, and have a single score for each item: therefore, if a student fails to respond correctly, is he/she lacking morphological competency or vocabulary? (Purpura 2004; Grabowsky, Dakin 2014).

2.2. Text

Many texts used for reading, vocabulary and grammar assessment in CILS certification are produced by manipulating authentic texts[3]; few are written by the item writer (e.g. the **jigsaw text**). Sometimes OTs are not chosen with care; they may feature in accuracies and some of these are retained in the MVs (Torresan 2015b). Beyond this, the manipulation of the texts often leads to new problems, primarily regarding coherence and cohesion, due to the erasure of some structural information. Some of these problems affect the comprehensibility of the text (Torresan 2015b, 2015d). In some MVs, we are also confronted with implausible information (Torresan 2015b). In some CILS MVs composed of **MCQ and TF reading tests**, we may also encounter syntactical, lexical and morphological problems as well as typos (Torresan 2015b, 2015d).

In some CELI MVs, instead of being manipulated, the text is re-written, using only simple fragments of the OT (Torresan 2016a). If we use the analogy of building construction, the CILS MVs are like a kind of restoration work: changes are minor, whilst the main structure is kept. By contrast, The CELI MV is like erecting a new building, just as, in Antiquity, many churches would incorporate bricks, columns, and other architectural items from Roman Empire monuments into a new building. In both cases, the examinations fail to comply with the widely recognized *Guidelines for Writers of Reading Tests*, established by the Hungarian *British Council* as part of the *Into Europe* Project (cfr. Alderson, Cseresznyés 2003 and webliography):

> "*Do not make any changes to the original* [i.e. *authentic*] *text. Do not delete words, sentences or paragraphs from the selected body of the text. If text contains any offensive words that you think should be replaced, only change these with great care and always seek the advice of a fellow teacher or a native speaker as to the acceptability of the changes you have made*" (# 3.1.10.)

2.3. *Items*

In the **CELI TF reading exercise** for A1 students, we have short sentences (statements, prohibitions, utterances) paired with images: the TT has to choose the picture corresponding to each linguistic input content. Three problems arise (Torresan2016b):

- pictures are not always clear and instantly recognizable
- some sentences are negative, hence the TT has to exclude the picture which illustrates the positive content of the given sentence (a logical conundrum!). With regard to these items, even intermediate TTsrespond erratically
- some utterances are out of context (e.g. it isunclear who the speaker is), making it difficult to ascertain the corresponding image

- some statements and prohibitions are expressed usingdifficult vocabulary

In the **CILS TF exercise** (“*information detection test*”), a format applied for testing reading and listening comprehension, the guess factor in volved in the test(any student has a 50% chance of guessing the correct response) is balanced out by having a long list of items. Some problems arise:

- some itemstarget exactly the same information as others(consequently, if the TT does not choose the right information, he/she will be penalized many times) (Torresan 2014b)
- items do not follow theorder in which the information progresses (Torresan 2015c)
- some items are ambiguous and would appear puzzling even to native speakers (Torresan 2014, 2015c)

The second problemis easily detectablein some**MCQ CILS listening tests**also (Torresan 2015e).

2.4. Calibration

Some comprehension tasks (text + items) are pitched too high/low for the target level.

For example, the **CILS TF reading test** for B1 students of the 2009summer session is over-set (even natives have difficulty doing it! cfr. Torresan 2015c).By contrast, the CILS achievement test (session August 2012) aimed at assessingcommunicative competency among Chinese students coming to Italy within the *Marco Polo* Project presentsan under-set 7-item**TF reading test**: it is an A2task,[4] opposed to the B1/B2 target proficiency level[5].

2.5. Layout

In some cases the manipulation of OT involves some changes in the original layout.

For example, in the text the **TF reading comprehension test** for B1 students of 2009 CILS summer session is based on, the original paragraphs and subtitles are replaced by a compact text, with no interruptions and no subtitles (Torresan 2015c). Some devices intended to scaffold comprehension are thus erased.

In the CELI certification, every single part of the test (instructions, text, items) is in bold. We believe this is uncomfortable on the eye, as everything is highlighted.

2.6. Score

Some CELI and CILS tests (e.g. **TF reading and listening CILS tests**) involve negative marking for wrong answers.

Negative marking may lead to a number of problems:

- *Pedagogical problems*. The message conveyed is that errors have to be punished. Is this coherent with the principles of the communicative approach?
- *Validity issues*. Imagine we have two students undertaking the CILS T Freading test: student A and student B. They are required to identify only those items referred to within the text. Both of them have identified out three correct items, so, ideally, they get the same score (3 pts). Nevertheless, the second one thought that a fourth option was also correct, but it is not a key. She will be thus be penalized with a negative score (-.5 pt). At the end, the total scores will differ: the first student gets 3 pts, while the second gets 2.5 pts. Does this difference truly reflect a real reading comprehension gap between student A and student B?
- *Reliability issues*. A negative score can lead to a situation such as that depicted below, where some TTs received a negative figure as their total score (TT # 1, 2).

Tab. 1

Scores received by Santa Monica College students' of a TF CILS reading comprehension test (B1 level, summer session 2009) in the 2012/13 year (Torresan 2015c)

	I1	I2	I3	I4	I5	I6	I7	I8	I9	I10	I11	I12	I13	I14	I15	
S1	-.5	0	1	0	-.5	0	1	-.5	0	-.5	-.5	0	0	0	1	**.5**
S2	0	0	1	0	-.5	-.5	0	-.5	0	-.5	-.5	0	-.5	0	1	**-1**
S3	0	0	0	1	-.5	0	0	0	-.5	-.5	-.5	0	-.5	1	0	**-.5**
S4	-.5	0	1	0	-.5	0	1	0	0	0	0	1	0	1	1	**4**
S5	-.5	0	0	0	-.5	0	1	0	0	0	0	1	-.5	1	1	**2.5**
S6	0	0	0	1	-.5	0	1	0	0	0	-.5	1	0	0	1	**3**
S7	-.5	0	1	1	-.5	0	1	0	0	-.5	0	1	0	0	1	**3.5**
S8	-.5	-.5	0	1	-.5	0	1	0	0	0	0	0	0	0	1	**2.5**
S9	0	0	1	1	-.5	0	1	-.5	0	-.5	0	1	0	1	1	**4.5**
S10	0	-.5	0	1	-.5	0	0	-.5	0	0	0	1	0	0	1	**1.5**

How can we approach a situation such as this, where reliability and discrimination values are to be calculated statistically?

2.7. Keys

In some CILS tests, there is evidence of:

- incorrect keys (Torresan2015b, 2015c)[6]
- duplicate keys (Torresan2014c,2016c)

With a specific reference to the **CILS and CELI cloze tests**, we also have:

- empty keys (the sentence can remain incomplete and the meaning of the passage does not change; Torresan2015c, 2016a)
- correct keys that do not spring even to the native expert's mind. In a **cloze test**, a good key is not only linguistically correct but also psycholinguistically plausible. That is, it is likely that a student of the target proficiency level (or even a native) will understand the semantic intention and not be 'pushed' by his/her mental lexicon towards a word or even a chunk of language different from the target word. In this instance, we would incorrectly evaluate the TT's morphosyntactic competency: he/she may know a specific tense very well but does not complete the task correctly due to the stimulated mental 'urge'. Hence, due the item's semantics, the co-text seems to 'attract' certain words or chunks, which are neither predicted nor allowed for by the item writer (Torresan 2016a)

2.8. Instructions

The instructions of the **CELI summary test** for advanced students are somewhat complex:

> *Riassumere il testo tenendo conto delle indicazioni fornite, senza riutilizzare integralmente frasi, espressioni o costrutti usati nel testo.*
>
> Summarise the text in line with the indications provided, without re-using whole sentences, expressions or constructs within the text.

Even for an EN the difference between "*expressions*" and "*constructs*" is not obvious.

In the **CILS textual closes** (Torresan 2015b), students are invited to complete the text but, as far as we know, they are not told how may words they can use (conversely, in CELI cloze tests they are).

In the **CILS TF comprehension tests**, students are given general instructions without being told *exactly* what they are being asked to do ("*scegli*", "choose" is a vague instruction), nor are they given an example, nor are they told how many items are supposed to be correct. In our view, such vagueness may generate confusion even for a native (Torresan 2014b, 2015c).

> *Leggi le seguenti informazioni. Scegli le informazioni presenti nel testo.*
>
> Read the following information. Choose the information presented in the text.

2.9. Prompt

In the**CELI speaking test**, there are visual prompts which are difficult to decipher. The doughnut chart, for example, that C1 candidate students had to comment upon during the 2013 summer session was in black and white, and so we would argue it was quite difficult for them to discriminate between the different sectors. Moreover, there was a duplicate item (two sectors referring to the same element) with different percentages in each case, which could prove puzzling for some candidates.

3. Conclusions

In addition to summarising analyses illustrated in a variety of articles, this article is intended to:

- reflect on language test validation practices, including trial runs on *educated natives*
- offer useful indications for teachers and administrators
- provide a quick overview and recognising the complexity that a well-designed test entails
- promote the as yet scarce debate on language assessment in Italy

[1] On the assessment of integrated tasks, cfr.Lee, Kantor 2005.

[2] Bachman (1982) argues that textual clozes are exceptions to this pattern (1982).

[3] The OT is only mentioned in the CELI certification.

[4] It is taken from "*CILS integrazione in Italia*" May 2012 session.

[5] Chinese students taking part in the *Marco Polo/Turandot* Project are supposed to reach the B1/B2 level so as to be allowed to attend University courses after a 6-month intensive course in Italy (cfr. Rastelli, Bonvino 2011). The target level is under-set compared to other foreigners who are required to achieve the B2 level to get into Italian Universities. So, having a A2 task has a huge impact on the consequent validity of the test as a whole (consider, too, that huge percentages – around 70% – of Chinese students taking part in the *Marco Polo* Project leave Italy and return to China by the end of the first Academic Year! Cfr. Rastelli 2010).

[6] The passage below is taken from the CILS **textual cloze** for C2 students (2007 summer session).

The topic is a digital data system. Logically, students have to insert data *into* the system (as it is in the OT). The change made by the item writer results in an incoherent sentence and an incorrect key!

OT	MV
Inserendo ***nel*** *sistema, poi, auto, pedoni e mezzi di trasporto pubblico, si opera la vera e propria interazione tra singolo individuo e città.* If we then integrate cars, pedestrians and public transport ***into*** the system, we see the true interaction between the individual and the city at work.	*Inserendo* ****questo*** *sistema, poi, auto, pedoni e mezzi di trasporto pubblico, si opera la vera e propria interazione tra singolo individuo e città.* If we then ****this*** system, cars, pedestrians and public transport, we see the true interaction between the individual and the city at work.

REFERENCES

Alderson, C., 1979, "The Cloze Procedure and Proficiency in EFL", *TESOL Quarterly*, 13, 219-227.

Alderson, C.; Clapham, C.; Wall, D., 1995, *Language Test Construction and Evaluation*, Cambridge: CUP.

Alderson, J. C., Csereszпyés, M., 2003, *Reading and Use of English*, Budapest: Teleki László Foundation.

Bachman, L., 1982, "The Trait Structure of Cloze Test Scores", *TESOL Quarterly*, 16, 1, 61-70.

Buck, G., 2001,*Assessing Listening*. Oxford: OUP.

Douglas, D., 2010, *Language Testing*, London: Hodder.

Grabowski, K. C.; Dakin J. W., 2014,"Test Development Literacy". In: A. J. Kunnan (ed.), *The Companion to Language Assessment*, Hoboken, NJ: John Wiley and Sons, 751-768.

Hulstijn, J. H., 2011, "Language Proficiency in Native and Non-native Speakers: An Agenda for Research and Suggestions for Second-language Assessment", *Language Assessment Quarterly*, 8, 3, 229-249.

Lee, Y.-W.;Kantor, R., 2005, "Dependability of New ESL Writing Test Scores: Evaluating Prototype Tasks and Alternative Rating Schemes", *TOEFL Monograph Series*, Report MS-31, Educational Testing Service: Princeton, NJ. Available online at: https://www.ets.org/research/policy_research_reports/publications/report/2005/ibao

Mulder, K.; Hulstijn, J. H., 2011, "Linguistic Skills of Adult Native Speakers, as A Function of Age and Level of Educatiom", *Applied Linguistics*, 32, 5, 475-494.

Porter, D., 1983, "The Effects of Quantity of Context on the Ability to Make Linguistic Predictions: A Flow in a Measure of General Proficiency". In: A. Hughes, D. Porter (eds.),*Current Developments in Language Testing*, London, Academic Press, 63-74.

Purpura, J. E., 2004, *Assessing Grammar*. Cambridge: CUP

Rastelli, S. (ed.), 2010, *Italiano di cinesi, italiano per cinesi: dalla prospettiva della didattica acquisizionale*, Guerra, Perugia.

Rastelli, S.; Bonvino, E., 2011, *La didattica dell'italiano a studenti cinesi e il progetto Marco Polo*, Pavia: Pavia UP.

Torresan, P., 2014a, "Il dettato e il dictocloze sono prove valide per valutare la comprensione orale?", *Romanica Cracoviensia,* 14, 138-150.

Torresan, P., 2014b, "Test a individuazione di informazioni della certificazione di italiano per stranieri CILS: aspetti critici",*Rassegna Italiana di Valutazione*, 18, 59, 104-123

Torresan, P., 2014c, "*Item Analysis* di una prova di lettura a scelta multipla della certificazione di italiano per stranieri CILS (livello B1; sessione estiva 2012)", *Caligrama: Revista de Estudos Românicos*, 19, 2, 17-33.

Torresan, P., 2015a, "Insidie nella confezione di un *sequencing task* quale test di lettura: uno studio a partire da una prova della certificazione CILS", *EL.LE Educazione Linguistica* [forthcoming].

Torresan, P., 2015b, *Studio su cloze mirati della certificazione CILS, livello avanzato*, Rio de Janeiro: Dialogarts.

Torresan, P., 2015c,"Individuazione di informazioni nella certificazione CILS: una nuova indagine",*Romanica Cracoviensia* [forthcoming].

Torresan, P., 2015d, "Analisi (classica, Rasch, dei distrattori) di una prova di lettura a scelta multipla della certificazione di italiano per stranieri CILS (livello B1; sessione estiva 2009)", *Euro-American Journal of AppliedLinguistics and Languages* [forthcoming].

Torresan, P., 2015e, "*Item Analysis* di prove di ascolto a scelta multipla della certificazione di italiano per stranieri CILS", *Signum. Estudos da Linguagem* [forthcoming].

Torresan, P., 2016a, "Insidie nella confezione di un cloze. Appunti a partire dall'osservazione di cloze morfolessicali della certificazione di italiano per stranieri CELI (livello B2, sessione 2007)", *Euro-American Journal of AppliedLinguistics and Languages* [forthcoming].

Torresan, P., 2016b, "Quando le immagini non 'mediano': analisi di item di una prova di lettura per apprendenti di italiano LS (CELI Impatto A1)", in Burgio, S.; Fontana, S. (ed.). *Appunti sulla mediazione linguistica*, Lugano: Agorà [forthcoming].

Torresan, P., 2016c, "Distrattori e chiavi in un cloze lessicale a scelta multipla di livello avanzato: l'opportunità di considerare il giudizio di nativi esperti", *Revista de Italianistica* [forthcoming].

WEBLIOGRAPHY

Into Europe Project, British Council, Hungary, 2002/ Available at: http://www.lancs.ac.uk/fass/projects/examreform/Media/GL_Reding&Use.pdf