TWO DECADES OF TRANSLATION VERIFICATION:
RELEVANCE AND LIMITATIONS OF METRICS, EVALUATION REPORTS AND STANDARDIZATION

Steve Dept
steve.dept@capstan.be

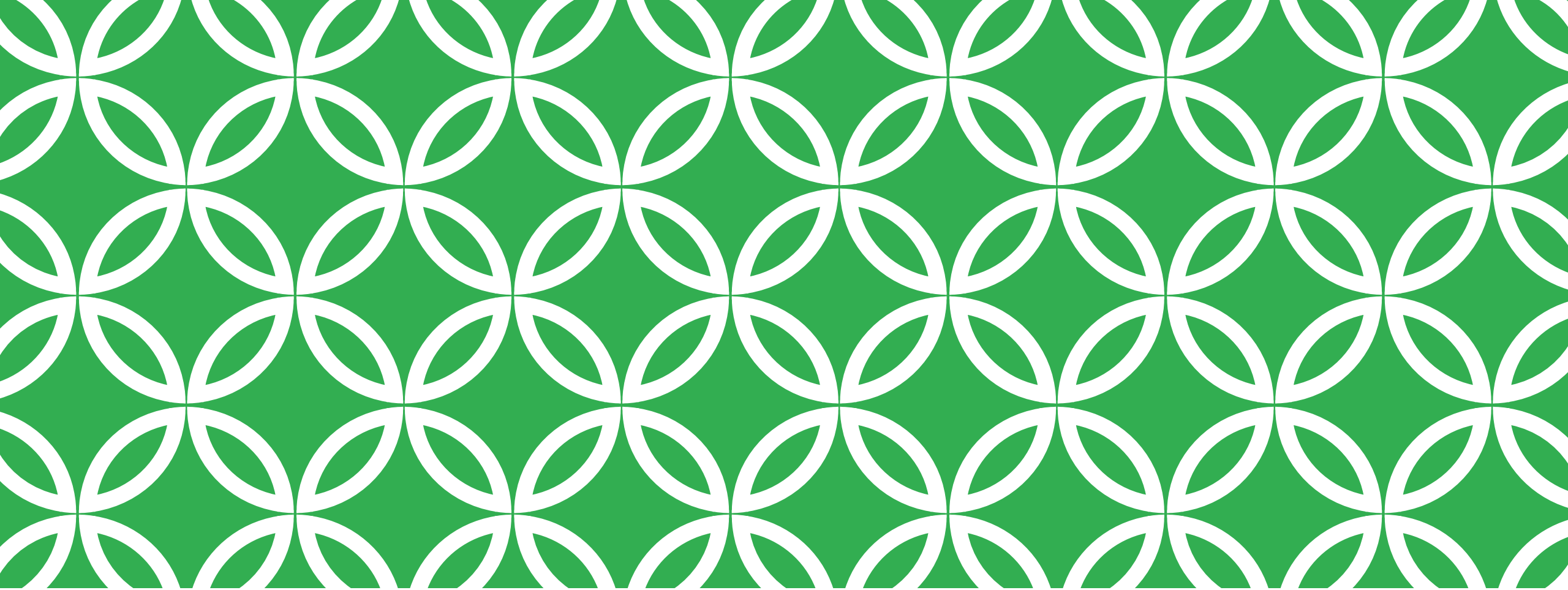
CONTENTS

SETTING THE STAGE – brief history of translation quality management in testing and in comparative research

REQUIREMENTS for a robust **translation verification design**

METHODOLOGY – diagnostics, metrics, corrective action, reporting practices

CONCLUSIONS – where do we stand? – where do we go from here?



SETTING THE STAGE

LANGUAGE QUALITY MANAGEMENT IN ILSAs

HISTORICALLY,

participation in high-stakes tests presupposes

- proficiency in the dominant language (Imperial examination)
- or in a scholarly language (Latin in European universities)

Austro-Hungarian Empire fostered equality of languages, but with different contents in different languages

Proponents of similar approaches in 20th and 21st century:

Triandis, H.C. (1964; 1972, 1976); Bonnet, G. & De Glopper, C. (2003); Boenhke, K. et al – emic approach (2014)

EARLY COMPARATIVE SURVEYS

IEA Pilot Project (1959-1961)

Political Attitudes and Democracy in Five Nations (G. Almond and S. Verba, 1963)

IEA Cross-national Study of Mathematics (1964)

IEA Six Subject Study (1970-1971)

MILESTONES

In the late 60s: “test translation changes test difficulty to the extent that comparisons across language groups may have limited validity”

In the 70s: linguistic quality control methods are introduced, e.g. back translation (Brislin,1970, 1976, 1986)

new insights are gained in how and why different forms of adaptation (to local context and usage) affect measurement

A good summary of breakthroughs: Adapting Achievement Tests into Multiple Languages … (Hambleton, 2002)

FIRST COMPARATIVE SURVEYS WITH LQC

IALS (1994-1998) – adults 16-65 from 22 countries

"data that were comparable across cultures and languages"

TIMSS (1995) – 500,000 students from 3 grades and 45 countries

"rigorous procedures have been developed for the direct and inverse translation of the items into the different languages of the participating countries, in order to ensure the levels of difficulty are maintained, over and above the specific language used for the test"

PISA 2000

Aletta GRISAY: experience in surveys, in linguistics, [illegible]
with data : a bridge between IRT and language ex

cApStAn: selection of language professionals with

- Teaching experience (school setting)
- Translation experience (from ENG/FRA)

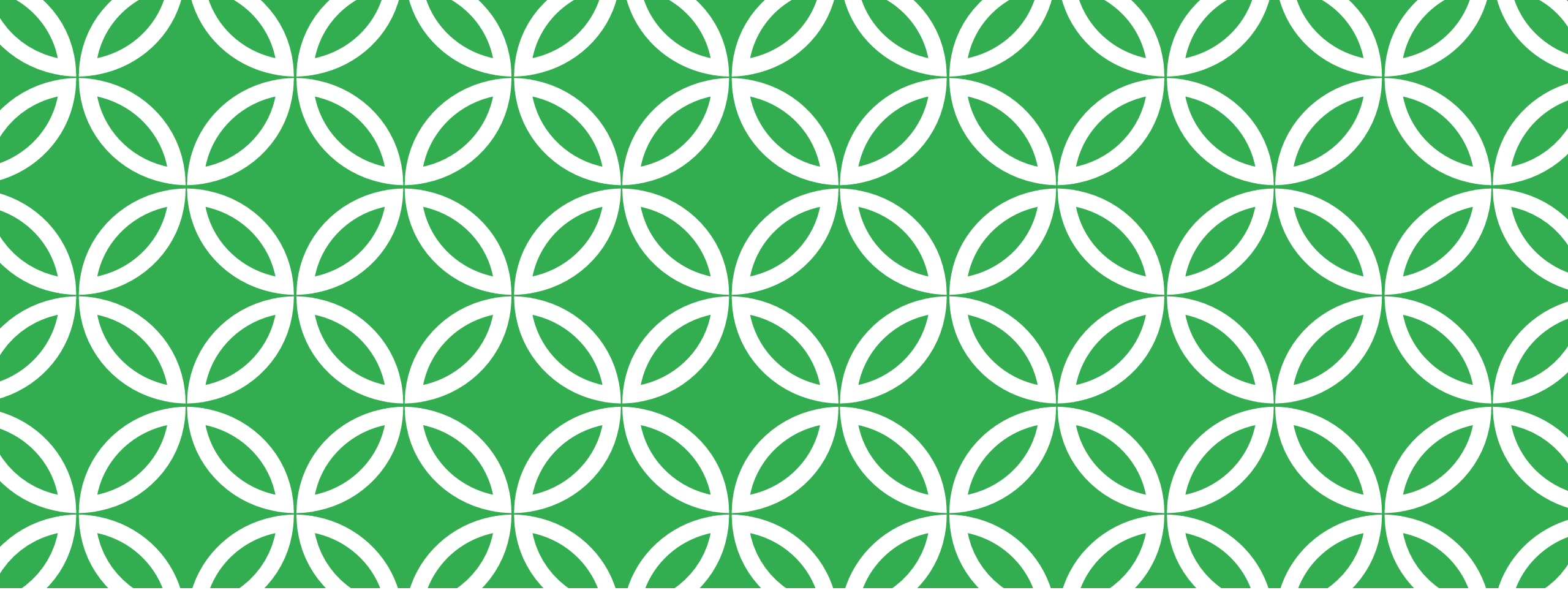
TAV Guidelines: collaborative effort, external validation; reference to ITC guidelines (1997)

No tools yet

PROOFREADERS' MARKS USED FOR THE VERIFICATION PROCEDURE

H f	delete; take these words out
⌢	close up; print as one word
f ⌢	delete and close up
#	insert a space
∧ or ⌊	insert here ⌊something
	space evenly ∧ where ∧ indicated
stet	let marked text stand as set
⌒ tr	transpose; change order the
/	used to separate two or more marks; often used as a concluding stroke at the end of an insertion
⊏	⊏set farther to the left
⊐	set ⊐ farther to the right
⌢	set æ or fl as ligature: æ
=	straighten align—nment—
‖	‖ straighten or align
¶	begin a new paragraph
(sp)	spell out (set 5 lbs. as five pounds)

bB	set in capitals (CAPITALS)
Bb	set in lowercase (lowercase)
ital	set in italic (*italic*)
Rom	set in roman (roman)
bf	set in boldface (**boldface**)
=	hyphen
en	en dash (1965-72)
em	em — or long — dash
∨	superscript ($\sqrt[3]{}$ as in πr^2)
∧	subscript ($\frac{1}{2}$ as in H_2O)
⌢	comma
⌣	apostrophe
⊙	period
;/	semicolon
:/	colon
« »	quotation marks
(/)	parentheses
[/]	brackets



PREREQUISITES

For standardised feedback on translations and adaptations

PREREQUISITES

A set of criteria

A common understanding of these criteria

A method to report on the extent to which criteria are met

A framework that is usable for linguists/reviewers

A link between criteria and corrective action

Feedback meaningful for test developers/psychometricians

MUST HAVES IN ILSA-SPECIFIC TQM FRAMEWORK

Clear, easily accessible TAV notes (item-by-item)

Train the trainer approach for translator training

Verifier training

Ownership >< Traceability

} workflows & procedures

Documentation

- FT Verification statistics per version
- FT Verification statistics per item

Process for FT to MS revisions to master + changes in nat'l versions

DIRECT ASSESSMENT VERSUS QQ

Cognitive Assessments

measure knowledge and skills

focus on level of difficulty

Maintain same quantity and quality of information, same clues

Adaptation: maintain register, matches & patterns, distractors

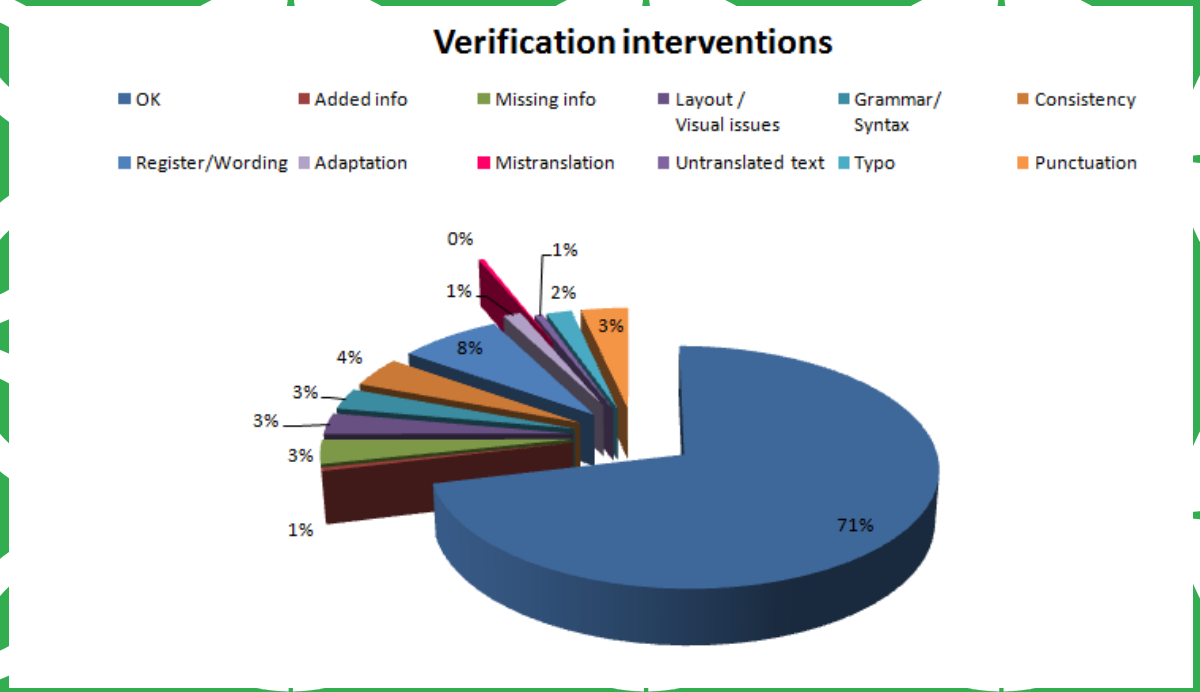
Background Questionnaires

collect data on background variables

focus on unambiguous formulation and clarity

Ask the same question

Adaptations to local context (ISCED, LANG, country-specific wealth indicators)



METHODOLOGY

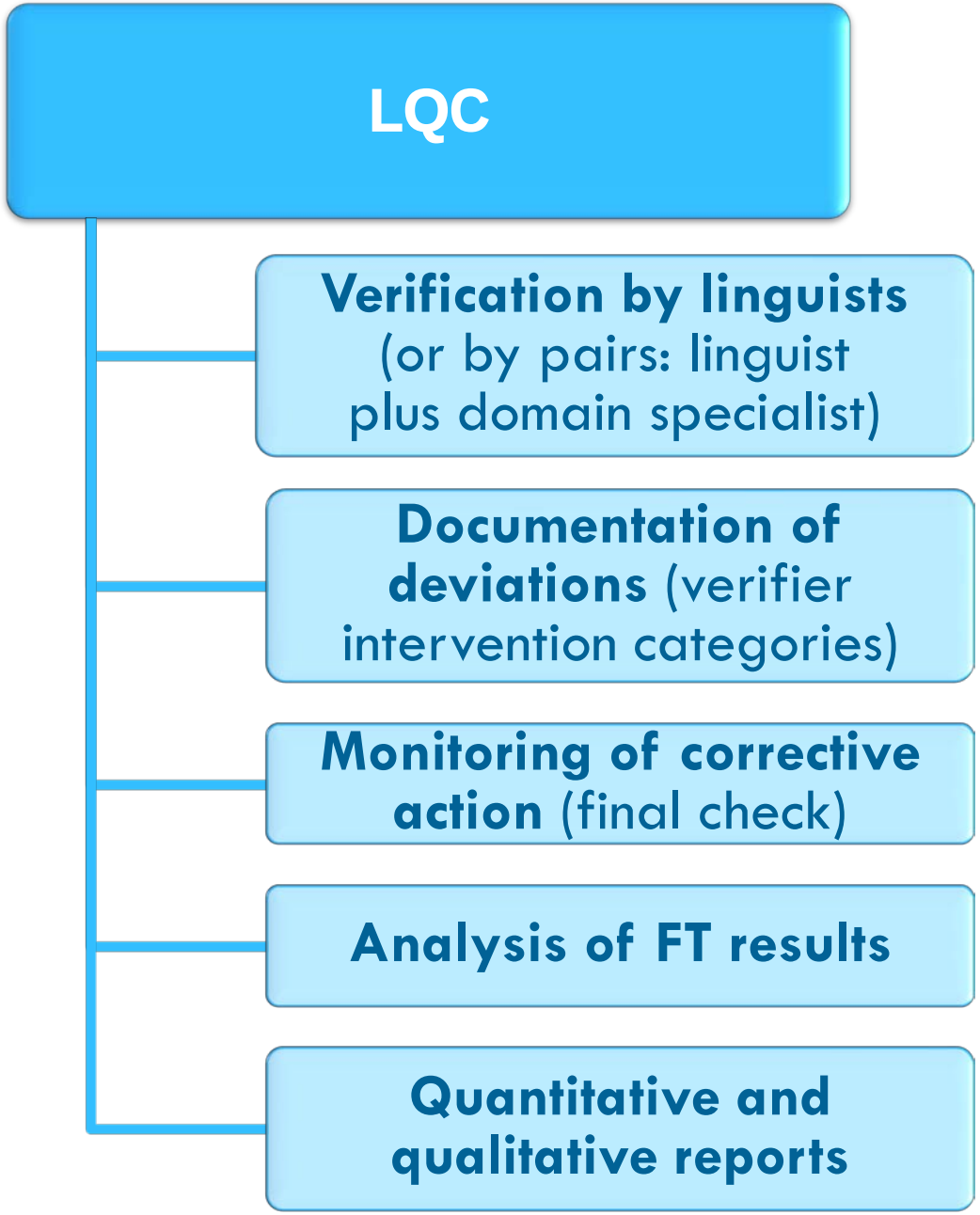
What metrics are most suitable to report on Translation Quality in ILSAs

Defining Linguistic Quality Control in the ILSA setting:

Check whether translated/adapted versions of data collections comply with general TA guidelines and item-specific TA notes

Report deviations as well as risks (potential equivalence issues)

Propose, implement and follow up corrective action



SEVERITY CODES (IEA)

Simple set of codes

In practice, difficult to standardise perception of the level of deviation

1. **Major Change or Error:**
   e.g. incorrect order of choices in MCQ; omission of a question;
   incorrect translation which changes the meaning or difficulty of the passage or question
2. **Minor Change or Error:**
   e.g. spelling errors that do not affect comprehension.
3. **Suggestion for Alternative:**
   translation may be adequate, but you suggest a different wording.
4. **Acceptable Change:**
   change is acceptable and appropriate.
   E.g. a reference to winter is changed from Jan to Jul for SH

1? **In case of Doubt:** not sure what code to apply
=> use "1?", so that no serious issue is left unaddressed.

SEVERITY CODES

Difficult to escape verifier effect

Digits may give impression of objectivity

Experiments w/ double verification show variance between verifiers having received the same training & instructions

Several Translation Evaluation frameworks combine severity and taxonomy (e.g. PIAAC Cycle 1 or Lionbridge)

VERIFIER INTERVENTION CATEGORIES (LIONBRIDGE)

Evaluation Summary	
Evaluation	80.8
Percentage of Correctness	81%

Evaluation Schema	
Evaluation	**Evaluation in Words**
95-100%	Excellent
85-94%	Good
75-84%	Satisfactory
<75%	Unsatisfactory

In this model, weighted severity is calculated automatically:

critical error = 5 minor errors
major error = 2 minor errors
minor error = 1 minor error

	Severity			
Error Category	**Minor**	**Major**	**Critical**	**Total per Category**
Accuracy	1	0	0	1
Wrong translation	0	1	1	2
Language	0	1	0	1
Style	0	0	0	0
Terminology	0	1	2	3
Country standard	1	0	0	1
Formatting	1	0	0	1
Instructions	0	0	0	0
Total per Severity	3	3	3	
Weighted Severity	3	6	15	

Definition of Severity	
Critical Weighted Severity 5	Critical error leads to extreme consequences and has been pointed out by the client as particularly severe for a particular
	Examples
	Errors in a highly visible part of the documentation or software, e.g. cover page, menu command.
	Errors that may carry legal, safety or health consequences.
Major Weighted Severity 2	Very serious errors that jeopardizes the meaning of a translation.
	Examples
	Errors in a visible part of documentation, e.g. header, TOC, dialog window.
	Accuracy errors that change the meaning of the translations, e.g. omitted text.
	Significant grammar or language usage errors.
	Previous corrections have not been taken into account.
	Error that provokes abusive or deragotory statement.
	Minor error that repeats throughout the translation.
Minor Weighted Severity 1	Minor error does not mislead a reader, does not change the meaning of the text.
	Examples
	Accuracy errors that result in a slight change in meaning.
	Small errors that would not confuse or mislead a user but could be noticed.
	Formatting errors not resulting in a loss of meaning, e.g. wrong use of bold or italics.
	Typos and misspellings that do not result in a loss of meaning.
	Style errors that does not change meaning of the text.
Weighted severity is calculated automatically!	
Critical	Weighted severity of a critical error is as 5 minor errors.
Major	Weighted severity of major error is as 2 minor errors.
Minor	Weighted severity of minor error is as 1 minor error.

Error Categories	
Accuracy	Omission, redundancy, incorrect cross-references (interface options, chapter titles, book titles etc.), mistakes caused by negligence, untranslated text, spaces, shallow check of matches, etc.
Wrong Translation	Improper perception of a source text, literal translation.
Language	Grammar, punctuation, syntax errors.
Style	Wrong register, inappropriate level of formality, style conventions not followed.
Terminology	The terminology does not follow generally accepted industry terminology; inconsistent use of terms, ignorance of a glossary.
Country Standard	Any regional or country standards not followed. This includes date format, units of measurement, currency, delimiters, addresses, etc., rendition of country names, person and proper names.
Formatting	Formatting errors, such as incorrect styles, fonts, bulleted and numbered lists; inadequate usage of italics, bold; hiddent text is translated, tag errors, links does not work properly.
Instructions	Guidelines not adhered to (style guide, language guidelines, technical instructions, etc.), previous references not observed.

Widely used to evaluate translators rather than to evaluate linguistic equivalence of target with source

VERIFIER INTERVENTION CATEGORIES (CAPSTAN)

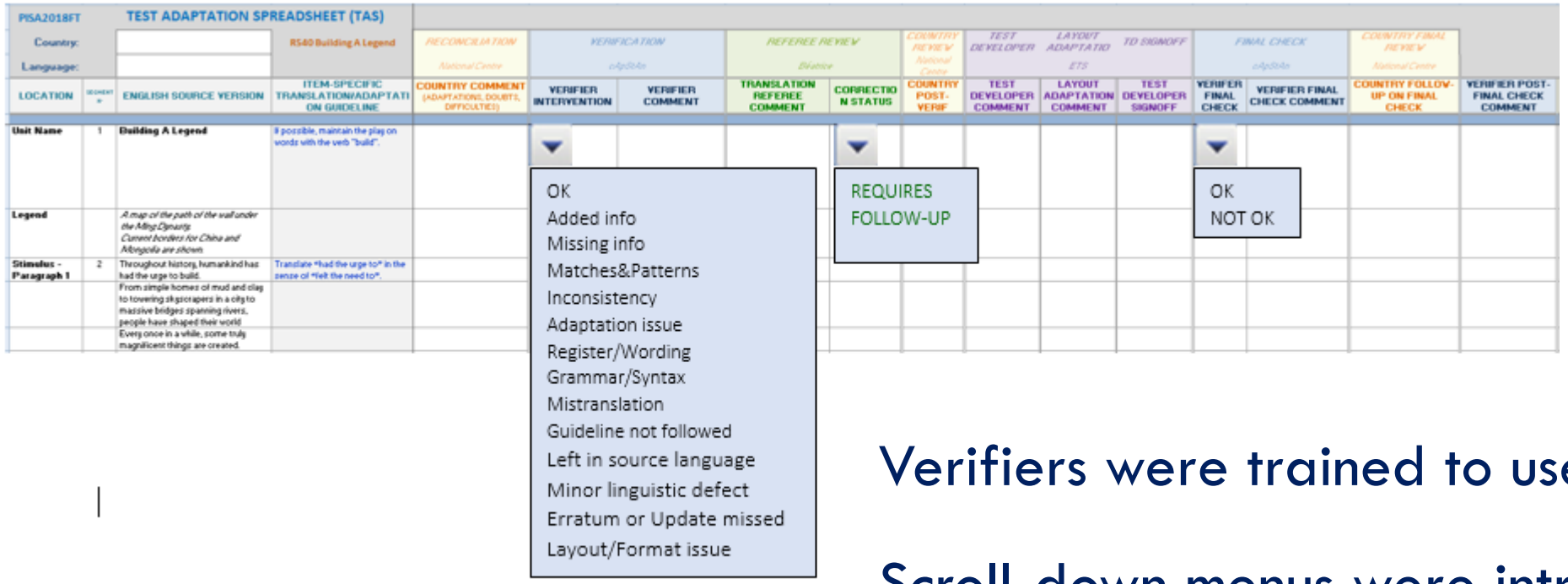
In PISA 2006 FT, verifiers commented on issues identified, with special focus on potential equivalence issues

5,380 verifier comments, covering 42 national versions in 36 languages for 38 countries, were analysed and described with key words:

A taxonomy of verifier intervention categories was developed:

OK	**No intervention** is needed. The verifier has checked and confirms that the text element or segment is equivalent to source, linguistically correct, and – if applicable – that it conforms to an explicit translation/adaptation guideline. This category may also be used to report an appropriate but undocumented adaptation.
ADDED INFORMATION	An information is present in the target version but not in the source version, e.g. an explanation between brackets of a preceding word.
MISSING INFORMATION	An information is present in the source version but omitted in the target version.
MATCHES AND PATTERNS	1) A **literal match** (repetition of the same word or phrase) or a **synonymous match** (use of a synonym or paraphrase) in the source version is not reflected in the target version. Most important: literal or synonymous matches **between stimulus and item and between a question stem and response categories**. 2) A **pattern** in multiple choice items is not reflected in the target version (e.g. all but one option start with the same word, proportional length of responses options.)
INCONSISTENCY	A recurring element across units (e.g. an instruction or prompt) is inconsistently translated, and this appears to be unintentional.
ADAPTATION ISSUE	An adaptation is an intentional deviation from the source version made for cultural reasons or to conform to local usage. An adaptation issue occurs when an adaptation would be needed but was not made, or when an inappropriate or unnecessary adaptation was made.
REGISTER / WORDING ISSUE	1) *Register*: difference in level of terminology (scientific term >< familiar term) or level of language (formal >< casual, standard >< idiomatic) in target versus source. 2) *Wording*: inappropriate or less than optimal choice of vocabulary or wording in target to fluently convey the same information as in the source. This category is used typically for vague or inaccurate or not quite fluent translations.
GRAMMAR / SYNTAX ISSUE	1. *Grammar*: grammar mistake that could affect comprehension or equivalence, e.g. wrong subject-verb agreement, wrong case (inflected languages), wrong verb form. 2. *Syntax*: syntax-related deviation from the source, e.g. a long (source) sentence is split into two (target) sentences or two (source) sentences are merged into a single (target) one; or another syntactic problem due e.g. to overly literal translation of the source.
MISTRANSLATION	A wrong translation, which seriously alters the meaning. A mistranslation should always be reported with a back-translation. Note: a vague or inaccurate translation should rather be classified as a Register/Wording issue (or sometimes a Grammar/Syntax issue). This category covers cases where the source has been misunderstood, but also copy/paste errors that unintentionally result in a wrong text element or segment.
GUIDELINE NOT FOLLOWED	An explicit translation/adaptation guideline for a given text element or segment was overlooked or was not addressed in a satisfactory way.
LEFT IN SOURCE LANGUAGE	A text element or segment that should have been translated was left in source language.
MINOR LINGUISTIC DEFECT	Typo or other linguistic defect (spelling, grammar, capitalization, punctuation, etc.) that does not significantly affect comprehension or equivalence. Correcting such errors is usually not controversial and can be made in track changes without documenting them.
ERRATUM/UPDATE MISSED	An erratum or update notice has been overlooked.
LAYOUT / FORMAT ISSUE	A deviation or defect in layout or formatting: disposition of text and graphics, item labels, question numbering, styles (**boldface**, underlining, *italics*, UPPERCASE), legibility of captions, tables, number formatting (decimal separators, "five" versus "5"), etc. In computer-based materials, this includes truncated words in the preview, undesired scrolling, etc.

VERIFIER INTERVENTION CATEGORIES (CAPSTAN)



PISA2018FT — TEST ADAPTATION SPREADSHEET (TAS)

Country:
Language:

RS40 Building A Legend

LOCATION	SEGMENT #	ENGLISH SOURCE VERSION	ITEM-SPECIFIC TRANSLATION/ADAPTATION GUIDELINE	COUNTRY COMMENT (ADAPTATIONS, DOUBTS, DIFFICULTIES)	VERIFIER INTERVENTION	VERIFIER COMMENT	TRANSLATION REFEREE COMMENT	CORRECTION STATUS	COUNTRY POST-VERIF	TEST DEVELOPER COMMENT	LAYOUT ADAPTATION COMMENT	TEST DEVELOPER SIGNOFF	VERIFIER FINAL CHECK	VERIFIER FINAL CHECK COMMENT	COUNTRY FOLLOW-UP ON FINAL CHECK	VERIFIER POST-FINAL CHECK COMMENT
Unit Name	1	Building A Legend	If possible, maintain the play on words with the verb "build".													
Legend		*A map of the path of the wall under the Ming Dynasty. Current borders for China and Mongolia are shown.*														
Stimulus - Paragraph 1	2	Throughout history, humankind has had the urge to build.	Translate "had the urge to" in the sense of "felt the need to".													
		From simple homes of mud and clay to towering skyscrapers in a city to massive bridges spanning rivers, people have shaped their world.														
		Every once in a while, some truly magnificent things are created.														

OK
Added info
Missing info
Matches&Patterns
Inconsistency
Adaptation issue
Register/Wording
Grammar/Syntax
Mistranslation
Guideline not followed
Left in source language
Minor linguistic defect
Erratum or Update missed
Layout/Format issue

REQUIRES FOLLOW-UP

OK
NOT OK

Verifiers were trained to use these categories

Scroll-down menus were introduced

formulas embedded in the worksheets

VERIFIER INTERVENTION CATEGORIES PER TEST UNIT

OVERVIEW OF VERIFICATION INTERVENTIONS PER CPS UNIT															
Unit Code	Added info	Missing info	Matches & Patterns	Inconsistency	Adaptation issue	Register/Wording	Grammar/Syntax	Mistranslation	Guideline not followed	Left in source language	Minor linguistic defect	Erratum or update missed	Layout/Format issue	TOT CORRECTIONS	REQUIRES FOLLOW-UP	WORD COUNT
C100-Xandar	4	25	8	79	7	205	97	30	15	8	67	4	44	593	71	1988
C101-The Visit	44	60	59	267	120	690	237	100	31	7	204	28	60	1907	262	8216
C102-Field Trip	12	37	40	161	19	231	95	26	16	5	75	45	38	800	122	3540
C103-Presentation	14	43	36	148	54	298	180	53	23	15	121	57	20	1062	181	4941
C104-Meeting in the park	9	31	33	124	12	232	123	30	9	9	82	15	64	773	104	4191
C105-The Garden	30	70	99	266	39	645	263	88	34	10	177	48	27	1796	288	8194
C106-Making a film	13	31	75	217	22	403	152	48	13	4	108	1	42	1129	151	5829
TOTAL	**126**	**297**	**350**	**1262**	**273**	**2704**	**1147**	**375**	**141**	**58**	**834**	**198**	**295**	**8060**	**1179**	**36899**

MQM ERROR TYPOLOGY

Multidimensional Quality Metrics (MQM)

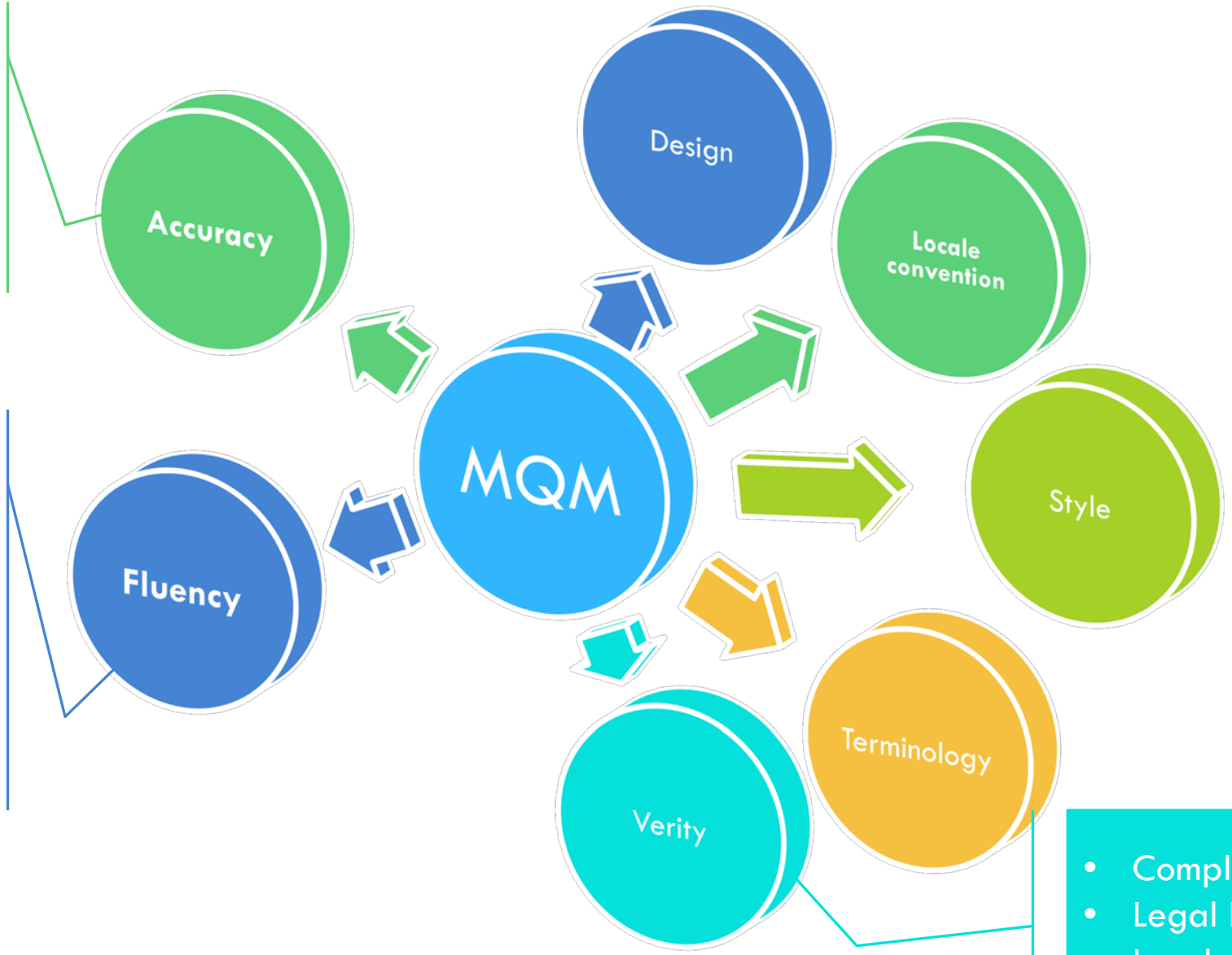
Flexible framework for the definition of custom metrics for the assessment of TQ.

Multiple levels of granularity

Provides a way to describe LQA systems, exchange information between them, and embed that information in XML or HTML5 documents

- Addition
- Mistranslation
- Omission
- Untranslated

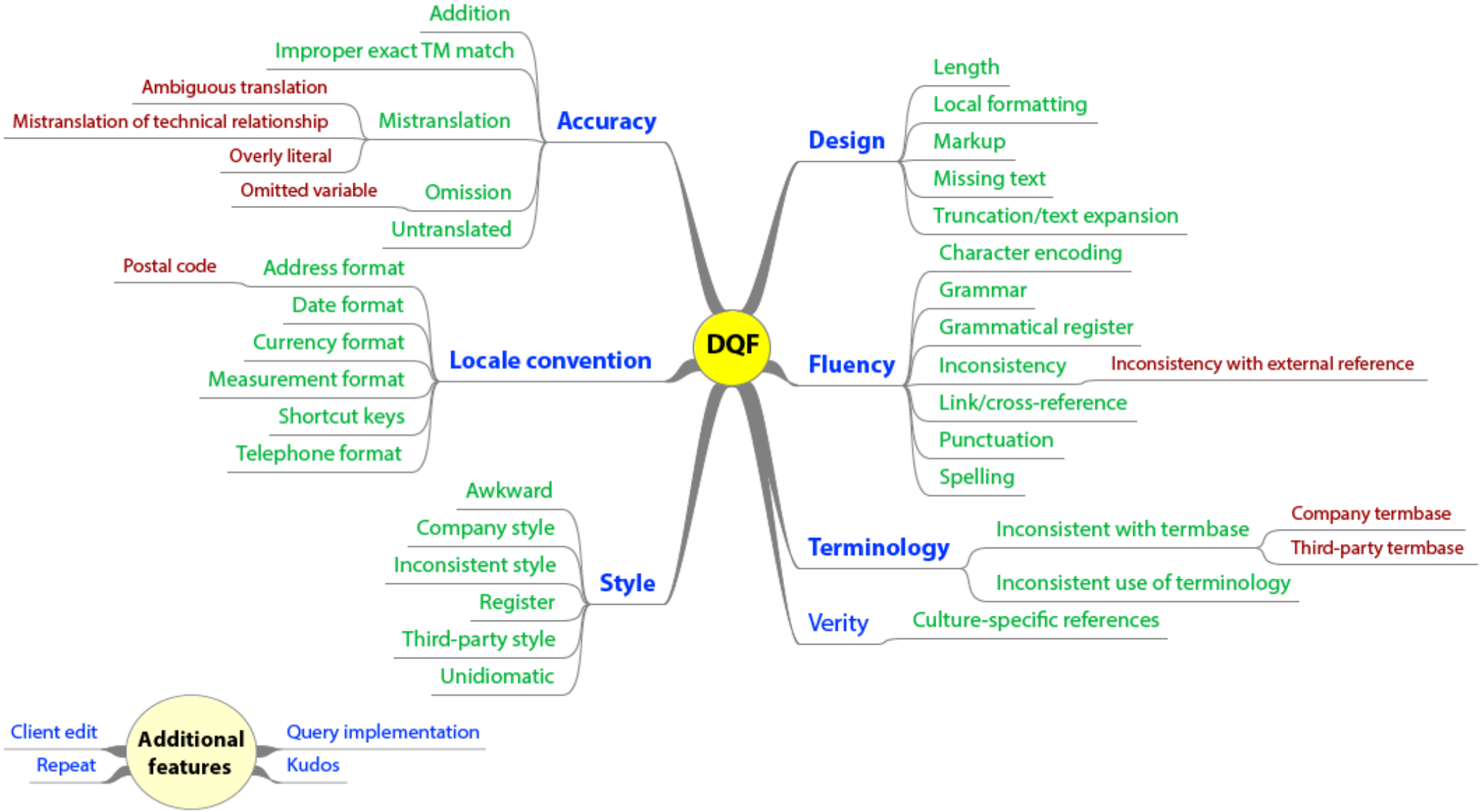
- Grammar
- Grammatical register
- Inconsistency
- Spelling
- Typography
- Unintelligible



- Completeness
- Legal Requirements
- Locale-specific content

www.taus.net

The TAUS Dynamic Quality Framework (DQF) Error Typology is a recognized subset of MQM, developed and maintained by the Translation Automation User Society based on input from its members.



THEORY OF TEST TRANSLATION ERROR (SOLANO-FLORES ET AL)

Solano-Flores, G., Contreras-Niño, L. A., & Backhoff, E. (2005). The Mexican translation of TIMSS-95: Test translation lessons from a post-mortem study. Paper presented at the meeting of the National Council on Measurement in Education. Montreal, Quebec, Canada

Conceptually very satisfying

Resource-intensive and time-consuming

Works best for post hoc verification

CHALLENGES AND TENSIONS

Tension between budget and quality requirements (e.g. PIAAC, PISA)

Insufficient attention given to existing resources on known issues (new teams want to implement new revisions)

Without thorough project management, multiple revision layers add more time and cost than value

automated checks save time and increase accuracy – but project-specific rules, glossaries, style guides etc. need to be prepared in advance

Deliverables are the result of a collaborative effort:
verifiers and reviewers of verification feedback focus on different areas

CONCLUSIONS

WHERE DO WE GO FROM HERE?

EXPORT, ANALYZE, EDIT, IMPORT BACK

Test delivery platforms are not TQM systems

ILSAs: a mix of professional and non-professional players

Proposed solution:

A) interaction between platform developer and translation technologist

B) export in a standard format (XLIFF)

C) analysis, QA checks, edits & documentation in dedicated TQM environment

D) import back

USER MANAGEMENT IN DEDICATED QA PLATFORM

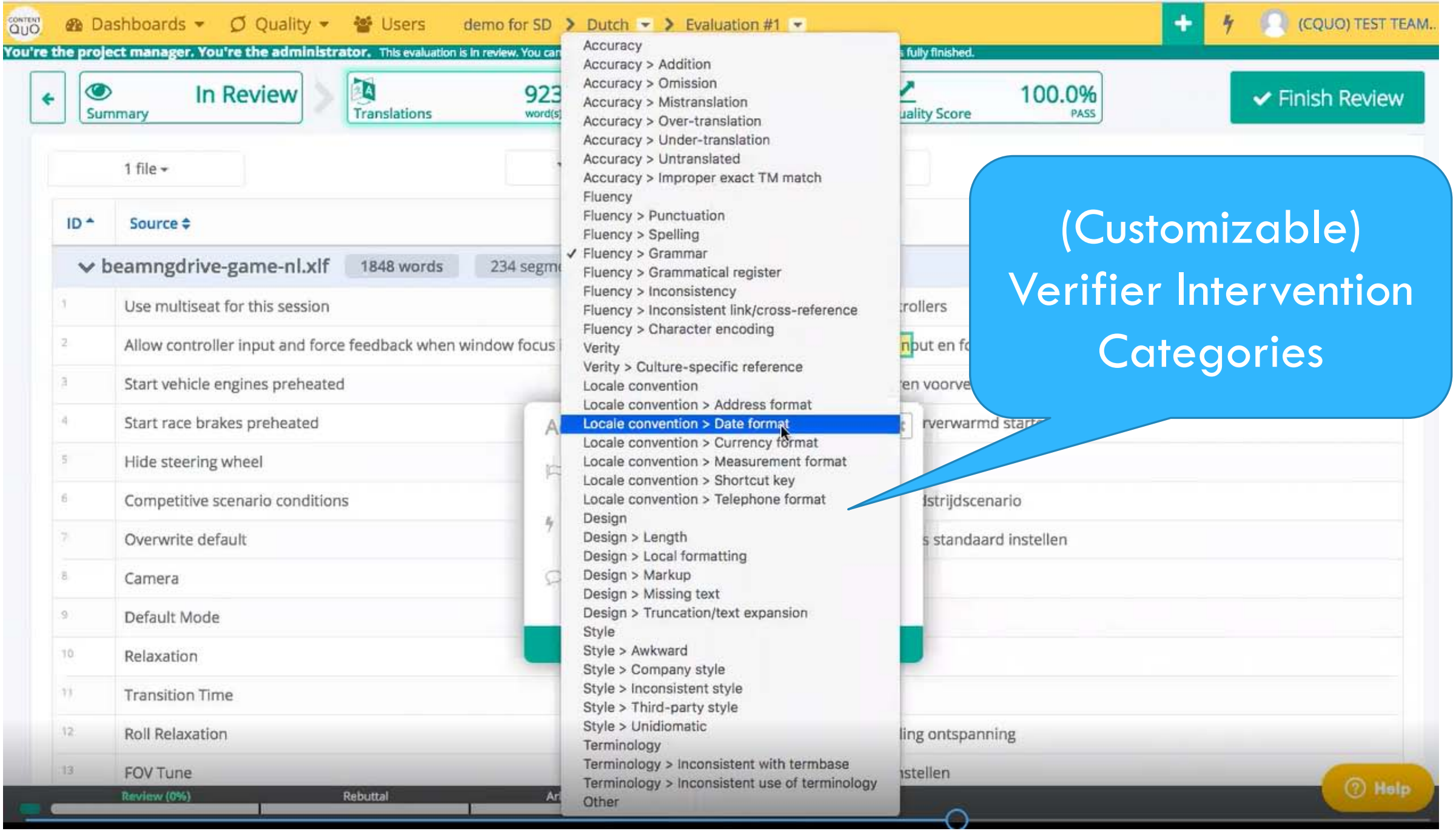
1. Verifier(s)
2. Verification reviewer (LQC organisation)
3. Translation team (National Centre)
4. Appointed Referee

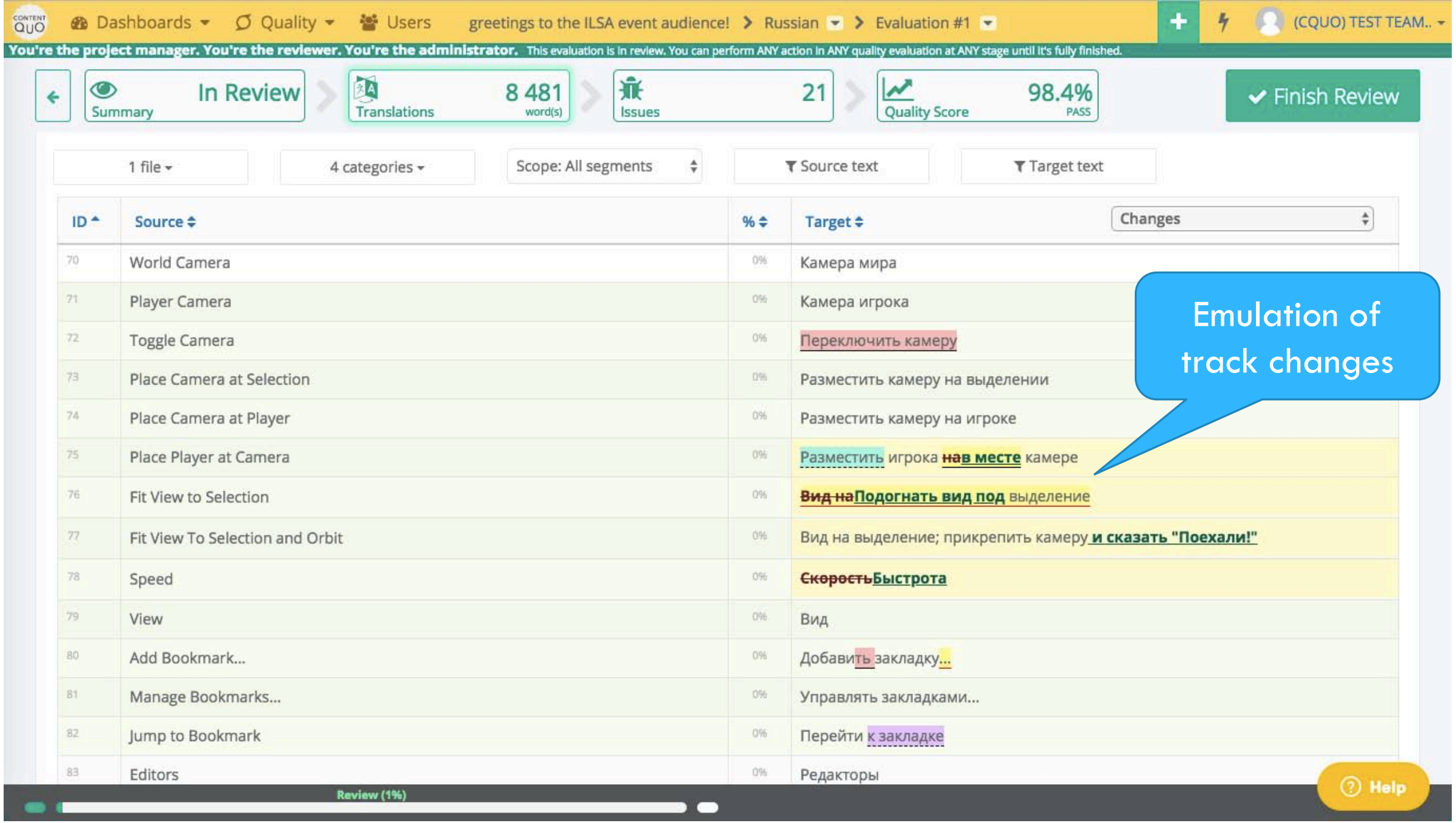
0. Project manager

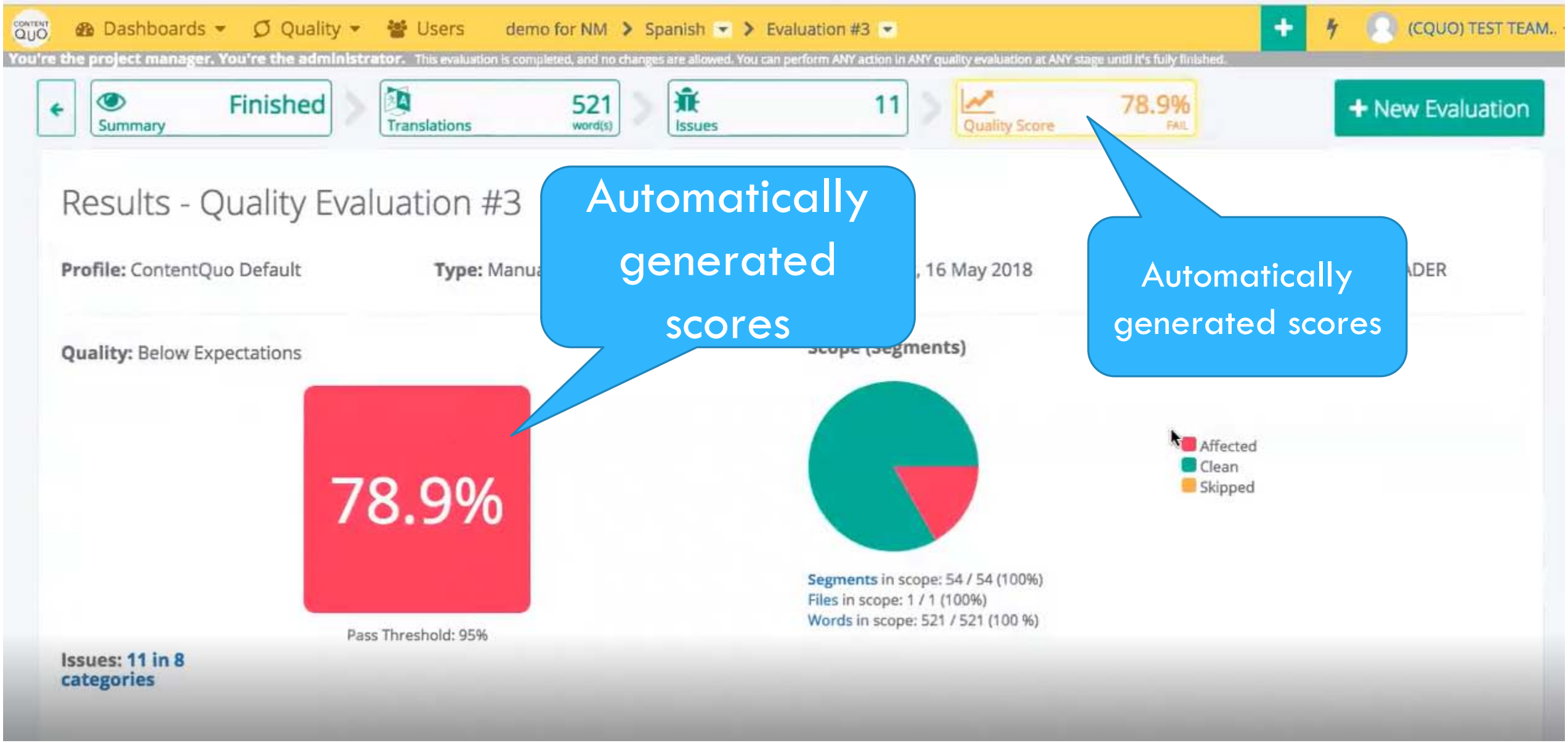
All successive stages saved and locked

ALL THE METRICS IN ONE PLACE

- A value is assigned to each intervention category
- Can be combined with severity (critical, major, minor)
- Quality evaluation in real time
- Dashboard
- Statistics and documentation can be consulted by all **& exported**
- A range of proxy indicators can be construed







OUTLOOK

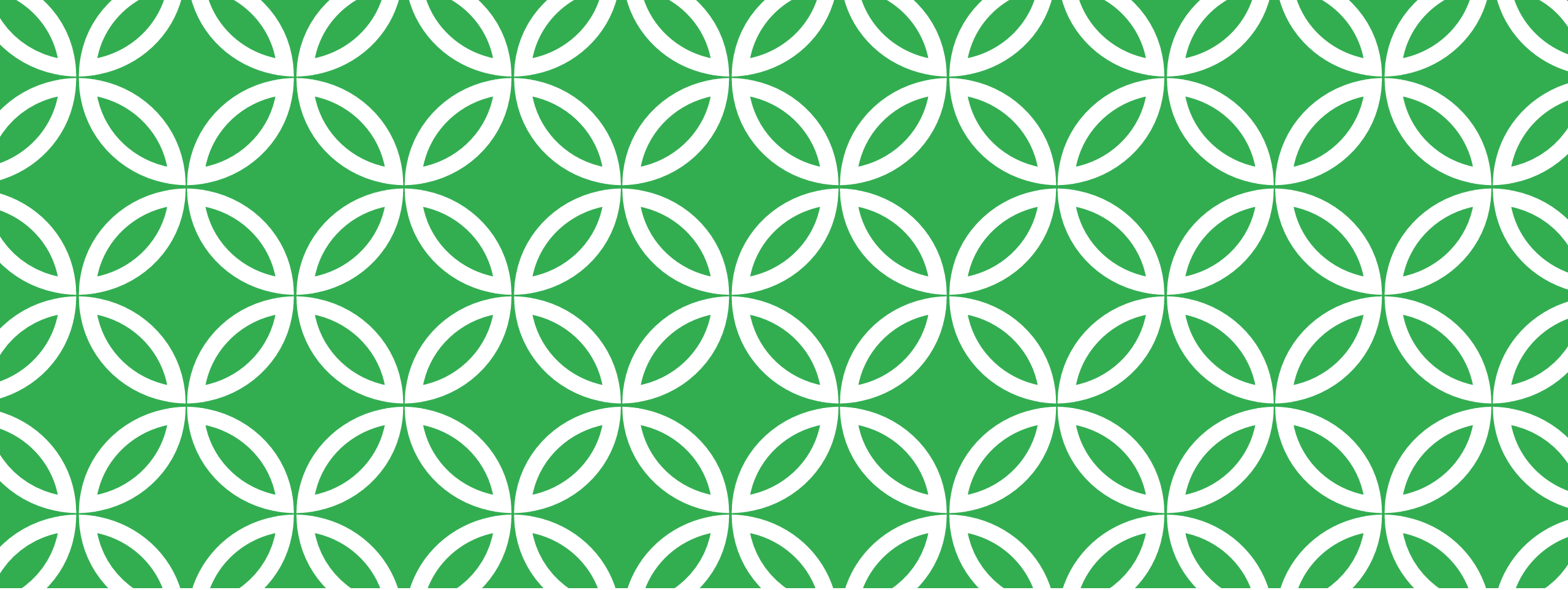
More automated QA for segments not linked to measurement

(combined with in-depth verification of representative sample?)

Focussed verification of pre-defined key segments
Linked to a shift towards more upstream preparation work

More work to be done on relation between important segments

Various combined TE scores, frequency tables



THANK YOU VERY MUCH

steve.dept@capstan.be