Creative Education, 2021, 12, 1912-1925
https://www.scirp.org/journal/ce
ISSN Online: 2151-4771
ISSN Print: 2151-4755

Proposing a Framework of Validity Evidence for a Score Report

Weilie Lu[1*], Yongqiang Zeng[2], Jin Chen[1]

[1]Center for Linguistics & Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China
[2]Guangdong Teachers College of Foreign Language and Arts, Guangzhou, China
Email: *891285101@qq.com

How to cite this paper: Lu, W. L., Zeng, Y. Q., & Chen, J. (2021). Proposing a Framework of Validity Evidence for a Score Report. *Creative Education, 12,* 1912-1925.
https://doi.org/10.4236/ce.2021.128146

Received: July 27, 2021
Accepted: August 14, 2021
Published: August 17, 2021

Abstract

Score reports are critical to the valid use of test scores and the interpretability of score reports has been recommended to be included as one type of validity evidence in validation research. In practice, how to support the validity argument for a score report has become the focus of practitioners' concern. In this paper, the authors first emphasize the significance of score reporting in test development before going on to review what a score report is, what it may possibly contain, validity and validation. Based on previous literature on validity of reports, a four-source framework of validity evidence for a score report is thus proposed. The four sources are 1) content alignment, 2) users' correct interpretation, 3) users' appropriate actions and 4) users' perception. Possible methods of collecting these different kinds of evidence are also suggested.

Keywords

Score Report, Validity and Validation, Validity Evidence for a Score Report

1. Introduction

A psychological or educational test is a procedure designed to elicit certain behavior from which one can make inferences about certain characteristics of an individual (Carroll, 1968: p. 46). To guarantee justified inferences, test developers have to follow a set of procedures. Test development is conceptually organized into three stages including design, operationalization and administration (Bachman & Palmer, 1996). In each of these three stages, there are some activities leading to different products. For example, the test administration stage involves giving the test to a group of individuals, collecting information, and analyzing this information with two purposes in mind: 1) assessing the usefulness of

the test, and 2) making the inferences or decisions for which the test is intended. Before making inferences or decisions, some other activities are also involved including: 1) describing test scores; 2) reporting test scores; 3) item analysis; 4) estimating reliability of test scores and investigating the validity of test use (Bachman & Palmer, 1996: p. 92), among which, reporting scores is an important step of test administration. Bachman & Palmer (2010) further described the activities involved in using a language assessment as including: 1) obtaining samples of individuals' language performance, 2) recording their performance in some way, 3) interpreting these records as indicators of some aspect of the test takers' language ability, 4) and making decisions on the basis of these interpretations. In Bennett's (2011) understanding, the cyclic process of educational measurement consists of four activities: "…designing opportunities to gather evidence, collecting evidence, interpreting it, and acting on interpretations (p. 16)", from the perspectives of the above researchers' and others' (Downing & Haladyna, 2006; Zenisky & Hambleton, 2012; Hambleton & Zenisky, 2013; Van der Kleij et al., 2014; Gandara & Rick, 2017; O'Leary, 2018), score reporting is an integral part in the test development cycle, and "the responsibility of developing reports that are accessible, useful, and understandable is increasingly a priority for testing agencies and program managers" (Hambleton, & Zenisky, 2013: p. 479).

The criticality of a score report is also recognized by other researchers. According to Klesch (2010), score reporting is a crucial aspect of educational testing, without which, test results would not be conveyed in a standardized way and may be left open to interpretation, most probably leading to misinterpretation. Score reports are critical to the valid use of test scores, as they are the only medium that connects test developers with stakeholders. If stakeholders are not able to understand or use the information in a score report, all the efforts and resources put into test development and data collection are simply wasted (Gandara & Rick, 2017). This viewpoint is also shared by Slater et al. (2018) "The most carefully developed, research-tested procedures for assessment design, item development, and psychometric analysis will be wasted if the score report does not communicate the test results in a way that encourages proper interpretation and use (p. 91)".

2. What Is a Score Report?

As the primary interface between test developers and multiple educational stakeholders like teachers, students, parents, score reports are critical in determining the success (or failure) of any assessment program. They are used to convey the performance of the examinee and serve as a catalyst for action in response to the performance. So, what is a score report? Different researchers provide different versions of viewpoints. Bachman & Palmer (2010) considered it to be a kind of feedback including the assessment record plus an interpretation. In some cases, more information like the decision that is made will also be included. According to Bachman & Palmer (2010), the specific contents of this assessment report will vary, depending upon the situation in which the

assessment is used. Table 1 lists the definitions proposed by different researchers.

Researchers have different focuses when talking about what a score report is. Some (Ryan, 2006; Zapata-Rivera & Katz, 2014; Rankin, 2016) emphasized its function as a communication tool, while others (Hambleton & Zenisky, 2013; Zenisky & Hambleton, 2015) focused on its layout and content. Concerning what should be included in a score report, there has been no fixed standard in literature. It all depends on the context. Just like what Hambleton & Zenisky (2013) commented, "Across testing contexts, the usability and quality of score-reporting materials has historically varied considerably from a simple listing of a total score to extensive breakdowns of test performance that are instructive for future improvement (p. 479)". A score report should contain all information necessary for interpretation, as intended by the designer. They further provided detailed components, including a description of the test purpose, intended uses for the test and test scores, and confidence intervals with accompanying explanations. Including subscore is what Van der Kleij et al. (2014) suggested because in their view reporting subscores can help users know test-takers' strengths and weaknesses, and improve the formative potential of the score report (Van der Kleij et al., 2014). Hambleton & Zenisky (2013) considered the following to be possibly included: 1) basic administration data such as test date, examinee name and contact information, 2) a performance-level classification such as pass-fail, 3) a description of a psychological state into which a respondent was classified. In Slater et al.'s (2018) view,

Table 1. Definitions of a score report by different researchers.

Researcher/year	what it is
Ryan (2006)	A score report was defined as a form of **communication**, with a sender, message, medium, intent, and audience. The message of a score report is ultimately interconnected to the other report aspects of sender, medium, intent, and audience, and is a culmination of decisions about what intended users of test scores need to know and how it can be presented to them in user-friendly ways.
Hambleton & Zenisky (2013)	A score report is a page **containing** a test score printed on it for a test taker, along with basic administration data such as test date, examinee name and contact information, and perhaps a performance-level classification such as pass-fail, or a description of a psychological state into which a respondent has been classified.
Zapata-Rivera & Katz (2014)	A score report is the **bridge** between the information captured by the test and the decisions or actions of the information-users.
Zenisky & Hambleton (2015)	A score report has most commonly been a physical piece of paper sent home with children or mailed to examinees' addresses from a testing agency. As a general rule, such reports are conceptualized as stand-alone and complete, so the narrative structure of the document's contents has had to reflect that orientation.
Rankin (2016)	Score reports have the purpose of **communicating data**, through tables, graphs and words to achieve a purpose, typically helping to turn data into actionable information for the intended audience.

the information on a score report "will nearly always include some kind of overall test score. It may also include classification levels, subscores, graphs, photographs or illustrations, and written text intended to help score users to understand the results or to help test takers interpret their scores or improve their performance (p. 93)". Up till now, the most detailed version of what a score report should include was proposed by Zenisky & Hambleton (2015), who suggested that reports should necessarily include both descriptive elements and data elements, and the detailed elements under these two categories are listed in **Table 2**.

A score report is the bridge between the information captured by the test and the decisions or actions of the information-users (Zapata-Rivera & Katz, 2014). It is a vehicle to let users know how scores can be understood appropriately in the context of the assessment and what are the supported actions that can be taken based on the results. It also shoulders the responsibility for supporting accurate user interpretation and use of test scores (O'Leary, 2018). In the test development cycle, score reporting appears to be the last step, but for those intended uses of tests, the creation and publication of a score report is only just the beginning. Indeed, it is only after scores are published (via a report), interpreted and acted upon, that the intended outcomes have a chance of actually taking place (O'Leary, 2018). Test users, or decision makers, interpret assessment records as indicators of the particular aspect of language ability and then use these

Table 2. Report content proposed by Zenisky & Hambleton (2015: p. 589).

Category	Contents
Descriptive elements:	Test name and/or test logo Test date Report title Report purpose Test purpose Introductory statement from testing agency or governing body personnel Headers with identifying details, such as name, address/school, groups membership or status (IEP, Language, etc.) Details for external links to additional resources, such as curriculum materials, and interpretive guides Information about the location of frequently-asked-questions documents or other resources for score inquiries Guidance on test score use Glossaries of terms Next step
Data elements	Summative results (scores and/or performance levels, as relevant to testing program) Performance-level descriptions Subdomain performance breakdowns Item-level results Norm-referenced results (to facilitate comparisons between groups or between individuals and relevant groups) Formative or diagnostic information Growth projections Item maps

interpretations to make decisions, which will, they believe, lead to their intended consequences (Bachman & Palmer, 2010). To decision makers, clear and easily-understood score reports are of outmost importance, which can ensure their correct interpretation and justified decision. To test-takers, clear and easily-understood scores are also important, but not enough, what is further needed is clues as to what plans should be made for next-step action. That is to say, accurate interpretations made and appropriate actions planned based on a score report are of utmost importance, which are what validity argument strive for.

3. Validity Evidence for a Score Report

3.1. Validity and Validation

Validity is the most critical consideration in test development (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 2014). It is a proposed abstract argument while validation is a practical process in which difference sources of evidence are collected to support such an argument. Traditionally, validity in testing and assessment has been understood to mean discovering whether a test measures accurately what it is intended to measure, or uncovering the "appropriateness of a given test or any of its component parts as a measure of what it is purposed to measure" (Henning, 1987: p. 170). When validity standards were first codified in the 1954 Standards (American Psychological Association, 1954), validity information was considered to indicate the degree to which the test is capable of achieving certain aims. And four types of validity, namely predictive validity, concurrent validity, content validity and construct validity, were identified corresponding to different aims of a test. In the 1966 version, predictive validity and concurrent validity, which were mainly concerned about external criterion, were combined and renamed as criterion-related validity, thus leading to the "holy trinity" (Guion, 1980). The three-type division concerning the concept of validity could also be seen in the 1974 version, but this view of validity is fragmented and incomplete, especially in failing to take into account evidence of the value implications of score meaning as a basis for action and of the social consequences of test use. Views on validity changed in the subsequent versions. According to the 1985 Standards (AERA, APA, & NCME, 1985), validity refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. In Guion's (1980) words, validity refers to an evaluation of the quality of the inferences drawn from test scores and he further emphasized that validity is a property of inferences from scores, not of the measuring instrument or test itself. Going a step forward, Messick (1989) described validity as "an integrated evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores" (p. 13), which has been widely accepted ever since.

The changing views on the concept of validity lead to different approaches to or frameworks of collecting evidence, namely, validation. Since the 1950s, four stages of approaches to validation can be generally classified: the stage of criterion-based approach, the stage of tripartite approach, the stage of unified approach and the stage of argument-based approach. In the stage of criterion-based approach, evidence has to be shown that the test is highly correlated with a criterion or selected standard. In the stage of tripartite approach, evidence of content validity, criterion-related validity and construct validity has to be accumulated. In the stage of unified approach, although validity is considered to be a unitary concept, which means that fundamentally there is only one kind of validity, namely, construct validity, several complementary forms of evidence need to be integrated in construct validation (Messick, 1989), including evidence bearing on test content, score structure, substantive processes, generalizability, external relationships, and testing consequences. In the stage of argument-based approach, what kind of evidence needs to be collected depends on what kind of claims have been proposed, in Kane's (2013) words, "it is the proposed score interpretations and uses that are validated ... the validity of a proposed interpretation or use depends on how well the evidence supports the claims being made... More ambitious claims are harder to validate than less-ambitious ones in that they require more support than less-ambitious claims (p. 1)".

3.2. Validity Evidence for a Score Report

3.2.1. Previous Research

To date, there is concensus that the most important concern for any test is the validity of its score interpretations or uses (Bachman & Palmer, 2010; Kane, 2013; O'Leary et al., 2017). Such a view is also echoed by Van der Kleij et al. (2014: p. 25): "a correct interpretation of test results is a necessary precondition for adequate use" and "a correct interpretation of reports is especially relevant when the test results are meant to inform important or irreversible decisions (p. 25)". However, much of the discussion pertaining to validity and validation has been focused on theoretical interpretations and use of scores, particularly on those uses that were (are) intended by test developers and designers (O'Leary et al., 2017; O'Leary, 2018), which can be clearly reflected in Kane's (2013) statement "it is the proposed score interpretations and uses that are validated… (p. 1)". According to O'Leary et al. (2017) and O'Leary (2018), if validity is to be truly concerned with the appropriateness of interpretations and use, then evidence of the quality, appropriateness, and effectiveness of the actual interpretations that test score users make and the actions they plan based on how scores are reported must be central to both the validity and validation processes. To emphasize the importance of user's actual interpretation of score reports, MacIver et al. (2014) proposed the concept of user validity to captures the "overall accuracy and effectiveness of interpretation resulting from test output" which focuses on "the validity of the interpretations in use and the decisions that form part of these interpretations (p. 155)". Being aware of the fact that supporting

users in interpreting assessment results is an important but underexposed aspect of validity, Van der Kleij et al. (2014) suggested that the interpretability of score reports should be included as an aspect of validity. Combining the views of MacIver et al. (2014) and Van der Kleij et al. (2014) and based on the then current conception of validity and validity evidence, O'Leary et al. (2017) and O'Leary (2018) went one step forward to incorporate evidence of user interpretation (interpretability) as a new kind of evidence when conducting validation research, which is shown in **Figure 1**.

In the above figure, validity judgment was based on the latest version of Standards which is mainly about "the degree to which evidence and theory support the interpretations of test scores" (AERA et al., 2014: p. 11). The process of validation involves synthesis of relevant, appropriate evidence in the formulation of "a sound scientific basis" in support of interpretations of scores for proposed uses. Five kinds of validity evidence listed include test content, score structure, substantive processes, external relationships and testing consequences. To emphasize the significance of user's actual interpretation based on score reports, O'Leary et al. (2017) and O'Leary (2018) proposed the expansion of the range of validity evidence to include the interpretability of score reports, which is represented in **Figure 2**.

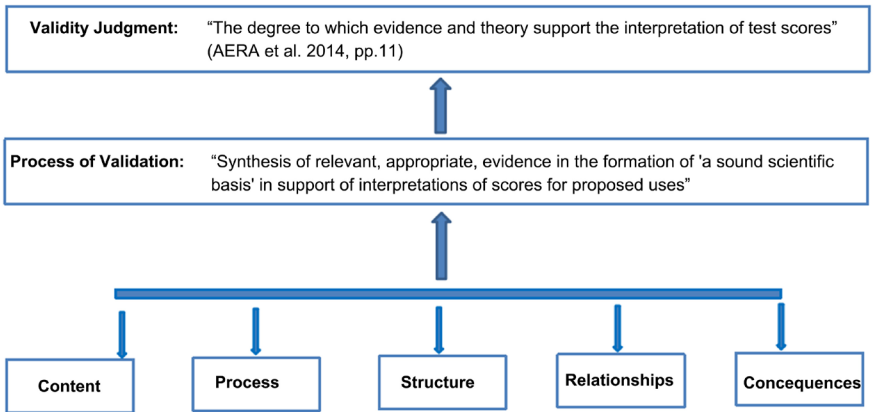

Figure 1. Conception of validity and validity evidence in Standards (AERA, APA, & NCME, 2014), adapted from O'Leary et al. (2017: p. 20).

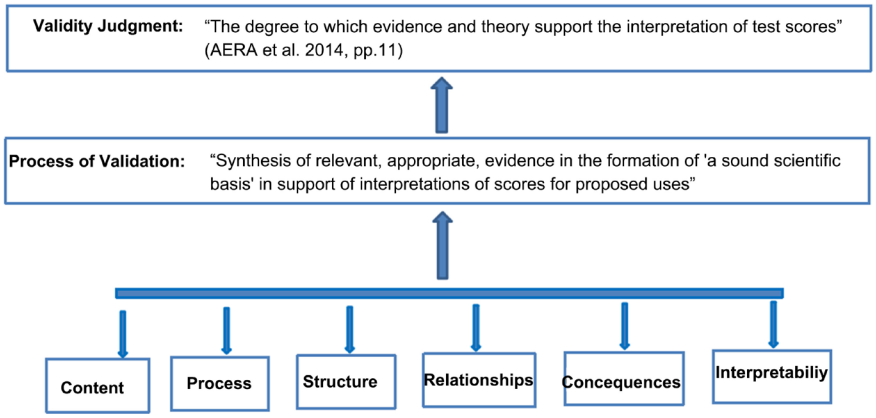

Figure 2. Expansion of validity evidence proposed by O'Leary et al. (2017: p. 20).

In the above figure, the researchers consider interpretability of score reports to be as important as evidence based on content, process, structure, relationship and consequences. To O'Leary et al. (2017) and O'Leary (2018), the above amendment is appealing but not sufficiently ambitious, they further recategorized the above six kinds of validity evidence into evidence of interpretation/use and technical evidence to show their different focuses, as is shown in **Table 3**.

From the above recategorization, we can see that the broad definition of validity as being "the degree to which evidence and theory support the interpretations of test scores" (AERA et al., 2014: p. 11) is retained. What is slightly changed is the identification of validity evidence as being one of two meta-forms: evidence of interpretation/uses and technical evidence. What is added is a clearer focus on the actual understanding of test users and the uses they plan. Such type of evidence focuses on the interpretation, uses, and consequences of how scores are actually interpreted by users. The changed model better aligned with their contention that "if validity is truly conceptualized with interpretations and use as being central, then evidence of the actual interpretations made and the use planned by score users must be included" (O'Leary et al., 2017: p. 19).

MacIver et al. (2014), Van der Kleij et al. (2014) and (O'Leary et al., 2017) contributed greatly to the theory and practice of validity and validation process. All of them attached importance to actual interpretations made and uses planned users (teachers, students, parents etc.) based on a score report, which is clearly reflected in O'Leary et al.'s (2017) statement "Central to the modern conception of validity are the interpretations made, and uses planned, on the basis of test scores (p. 16)". In a similar vein, when talking about the validity of score reports, Hattie (2009) claimed that "the validity of reports is a function of the users' correct and appropriate inferences and/or actions about the test taker' performance based on the scores from the test (p. 1)". In Hattie's (2009) argument, to address this claim, readers, based on a score report, should at least correctly answer two major questions: *What do you see*? *What would you do next*? In Hattie's (2009) words, "These two questions focus on the two most critical aspects of validity: the appropriate interpretions and actions from reports (p. 3)".

What is clear from the above discussion is that when collecting validity evidence for a score report, users's correct interpretation and appropriate actions planned are essential. However, what more sources of evidence and how evidence should be collected haven't been mentioned. Starting from a different

Table 3. Recategorization of validity evidence (Adapted from O'Leary et al., 2017).

Types of Evidence	Focuses
Interpretation an Uses	Interpretability Consequences
Technical Evidence	Content Process Structure Relationship

perspective and based on five sources of validity evidence included in the 2014 Standards, Tannenbaum (2018) proposed five sources of validity evidence for score reports, which are presented in the following **Table 4**.

Of the five sources of evidence above, the first one is concerned with content alignment—the content of a score report should truly reflect the content of the test, without which, correct interpretation is impossible. In Tannenbaum's (2018) words, "A score report that is not well-aligned with the test is of little value (p. 9)". The other four sources of evidence are all about users' correct interpretation. Evidence should show that users attend to the more salient information in the report and interpret it as intended (Evidence based on processes). Evidence should show that different subgroups of users interpret the same reported information in the way intended (Evidence based on structure). Evidence should show that users' interpretation of the students' competency is consistent with the teachers' evaluation of those students (Evidence based on relationship). Evidence should show that users do not make inaccurate interpretations leading to inappropriate decisions (Evidence based on consequences). Tannenbaum (2018) greatly emphasized the interpretation of a score reports in that "stakeholders cannot make reasonable decisions or take reasonable actions from information that they do not satisfactorily understand, no matter how accurate that information may be in reality (Tannenbaum, 2018: p. 9)".

It has been shown that evidence of users' correct interpretation and uses are critical in validation. Only when users use the interpreted information for some purpose is a score report meaningful. Different users use the information in the

Table 4. Validity evidence for score reports (Tannenbaum, 2018).

Sources	validity evidence
1) Evidence based on content	There should be evidence that the reported information is aligned with the test content, and presented in a way that is understandable to the stakeholders. The score report should be a faithful reflection of what the test measures and how the test taker(s) performed on the test, which may include areas to improve upon and the identification of resources to assist in that regard.
2) Evidence based on process	Evidence should support that the score-report users are attending to the more relevant or salient features of the report, and interpreting that information as intended.
3) Evidence based on structure	Evidence should confirm that stakeholders recognize the intended relationship among the information reported. Evidence should support that subgroups of stakeholders understand the same reported information in the way intended.
4) Evidence based on relationship	Evidence should take the form of comparing how closely the level of students' competency expressed on the score report is to teachers' evaluation of those students' competencies.
5) Evidence based on consequences	Evidence should indicate that stakeholders are acting on the reported information in ways consistent with reasonable expectations, and not making inaccurate interpretations leading to inappropriate decisions.

score report in different ways. For decision-makers, after interpreting the score report, they make decisions which will affect other stakeholders especially test takers. For school principals or programme providers, they use the information to make adjustment to the program so that it can effectively help learners to achieve certain goals. For teachers, they use the information to know what areas students need to improve and what they should focus on teaching to help students improve. For parents, especially parents of younger learners like primary pupils and students in junior high schools, who are still not mature enough in both mentality and behavior and who are still in great need of parental guidance in daily life and schoolwork, they use the information to better help their children to improve. For test-takers themselves, especially for independent learners like college students, they use the information to make justified study plans so that they can improve themselves in the near future. If there is no action following users' correct interpretation of a score report, which means that the information is not used at all, such interpretation is meaningless. What could cause the interpreted information to be used more deeply? It's found out and emphasised that perceptions of the users is one important factor that determines to what degree feedback will be used. A negative attitude towards performance feedback can be an obstacle for feedback use (Bosker, Branderhorst, & Visscher, 2007). Vanhoof et al. (2011) suggest that the degree in which feedback is actually used is affected by the level of confidence users have in their own knowledge and ability to use data, as well as by their attitude towards feedback. From these researchers' statement, it can be concluded that if users have positive perceptions of a score report, it is more likely for them to use the information contained. Appropriate actions are more likely to follow accurate interpretation. In this sense, users' perception of a score report has to be collected as one source of validity evidence.

3.2.2. A Four-Source Framework of Validity Evidence

Although researchers are increasingly interested in the topic of score report validity, there is no easy-to-follow framework for collecting validity evidence. Both Van der Kleij et al. (2014) and O'Leary et al. (2017) focused on interpretation of score reports and proposed that it should be included as one kind of evidence in validation. However, they made no further statements on validity of score reports. Hattie (2009) directly talked about validity of score reports, and proposed two critical aspects of validity: 1) the appropriate interpretation and 2) actions from reports, which greatly narrowed the scope of score report validity. Tannenbaum (2018) proposed five sources of validity evidence for score reports based on those included in the Standards (AERA, APA, & NCME, 2014). However, as can be seen from previous discussions, the five sources are partly overlapped and mainly focused on interpretation. Equipped with this framework, researchers are not clear as to what evidence should be collected to support the argument of a valid score report. Taking all these aspects into consideration, we deem a clearer framework to be badly needed.

Based on previous literature and enlightened by the above researchers' findings and arguments (Hattie, 2009; Bosker et al., 2007; Vanhoof et al., 2011; Kane, 2013; MacIver et al., 2014; van der Kleij et al., 2014; Tannenbaum, 2018), we propose a four-source framework of validity evidence to support the validity argument for a score report. First of all, evidence should be collected showing that the content in the score report faithfully reflect the content covered in the test, which is content alignment. This kind of evidence can be collected through expert judgment based their review of the score report, the test paper, test specifications or teaching materials. Questionnaire or interview asking report users questions concerning what is covered in the score report can also reveal the degree of alignment between the two. Second, evidence should be collected showing that users can interpret the information contain a score report correctly, which is users' correct interpretation. This kind of evidence can be collected through comprehension test, questionnaire and interview. For example, to investigate whether teachers could accurately interpret a standardized score report, Impara et al. (1991) used a 17-item score report comprehension test. O'Leary (2018) used a 5-question comprehension test to evaluate the interpretability of a hypothetical score report, both of which proved to be effective research methods. Third, following correct interpretation, evidence should be collected showing that users could take or plan justified actions. The way to collect such evidence can be like this: 1) Ask users to plan actions, 2) experts are invited to judge whether the actions planned are justified or appropriate or not. For example, if we want to know whether students can make appropriate plans based on the score report, we can ask the student to write a study plan for next-step improvement, then their teachers are invited to judge these plans are justified or not. Fourth, evidence should be collected showing that on the whole, users positively perceive the score report to be useful, which explores users' perception on the score report. This kind of evidence can be collected through questionnaire and interview. For example, by using questionnaire, Gorney (1999) investigated participants' perception of the adequacy of information and their presentation mode preferences. The four sources of evidence and some possible methods of collecting evidence are listed in **Table 5**.

Table 5. Framework of validity evidence for a score report.

Sources	focus	possible methods
Content alignment	whether the content in a score report faithfully reflects the content in a test	expert judgment questionnaire interview
Users' correct interpretation	whether users can accurately interpret the information conveyed by the score report	comprehension test questionnaire interview
Users' appropriate actions	whether users can plan next-step actions based on the interpretation of the score report	expert judgment interview
Users' positive perception	whether users positively perceive the score report to be useful for their purposes	questionnaire interview

4. Concluding Remarks

Theoretically, interpretability of a score report should be included in validity judgment and process of validation (O'Leary et al., 2017; O'Leary, 2018). In practice, researchers have to be equipped with the knowledge and methods as to how and where to collect validity evidence for a score report, which is the purpose of this paper. Hopefully, the four-source framework of validity evidence proposed in the paper can better guide researchers in validation studies concerning the effectiveness or usefulness of a score report. Considering its critical status in test development, validity arguments have to be supported by various sources of evidence. It should be guaranteed that the contents in a score report should well align with those in a test, which is a fundamental step. Evidence should be collected to show that users, when faced with a score report, can accurately interpret it as intended, which is the precondition for appropriate and justified actions planned. For a score report to achieve its intended aims, evidence should be collected to show that based on correct interpretation, users can plan next-step actions. If there are no actions to be taken, then what is the point of devoting so much time and effort into the business of reporting scores? Only when the interpreted information is used can a score report realize its value. Considering that users' attitude towards the usefulness of the score report impacts the degree to which involve themselves in user feedback, evidence should be collected to show that overall users positively perceive the score report to be useful. If there are negative perceptions exist, then the reasons have to be explored, which could be useful information for further revision of the score report.

Funding

This work was supported by the [National Education Science "13th Five-Year Plan" 2018 Ministry of Education Youth Project: Research on computerized adaptive test with individual characteristics] under Grant [number EIA180491].

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (1985). *Standards for Educational and Psychological Tests and Mannuals*. American Educational Research Association.

American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME) (2014). *Standards for Educational and Psychological Tests and Mannuals*. American Educational Research Association.

American Psychological Association (1954). *Technical Recommendations for Psycholog-*

ical Test and Diagnostic Techniques. Author.

Bachman, L. F., & Palmer, A. (2010). *Language Assessment in Practice*. Oxford University Press.

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford University Press.

Bennett, R. E. (2011). Formative Assessment: A Critical Review. *Assessment in Education: Principles, Policy & Practice, 18,* 5-25. https://doi.org/10.1080/0969594X.2010.513678

Bosker, R. J., Branderhorst, E. M., & Visscher, A. J. (2007). Improving the Utilisation of Management Information Systems in Secondary Schools. *School Effectiveness and School Improvement, 18,* 451-467. https://doi.org/10.1080/09243450701712577

Carroll, J. B. (1968). The Psychology of Language Testing. In A. Davies (Ed.), *Language Testing Symposium: A Psycholinguistic Approach* (pp. 46-69). Oxford University Press.

Downing, S. M., & Haladyna, T. M. (2006). *Handbook of Test Development*. Erlbaumb.

Gandara, F., & Rick, F. (2017). Empowering End Users to Evaluate Score Reports: Current Challenges and Proposed Solution. *Pensamiento Educativo, Revista de Investigación Educacional Latinoamericana, 54,* 1-24. https://doi.org/10.7764/PEL.54.2.2017.11

Gorney, B. T. (1999). *The GRE Score Report: Exploring Alternatives for Presentation of Individual and Aggregate Scores to Institutional Recipients*. Unpublished Ph.D. Thesis, University of North California.

Guion, R. M. (1980). On Trinitarian Conceptions of Validity. *Professional Psychology, 11,* 385-398. https://doi.org/10.1037/0735-7028.11.3.385

Hambleton, R. K., & Zenisky, A. L. (2013). Reporting Test Scores in More Meaningful Ways: A Research-Based Approach to Score Report Design. In K. F. Geisinger (Ed.), *APA Handbook of Testing and Assessment in Psychology: Vol. 3 Testing and Assessment in School Psychology and Education* (pp. 479-494). American Psychological Association. https://doi.org/10.1037/14049-023

Hattie, J. (2009). Visibly Learning from Reports: The Validity of Score Reports. *Online Educational Research Journal.* http://www.oerj.org/View?action=viewPDF&paper=6

Henning, G. (1987). A Guide to Language Testing. *Newbury House.*

Impara, J. C., Divine, K. P., Bruce, F. A., Liverman, M. R., & Gay, A. A. (1991). Does Interpretive Test Score Information Help Teachers? *Educational Measurement: Issues and Practice, 10,* 16-18. https://doi.org/10.1111/j.1745-3992.1991.tb00212.x

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50,* 1-73. https://doi.org/10.1111/jedm.12000

Klesch, H. S. (2010). *Score Reporting in Teacher Certification Testing: A Review, Design, and Interview/Focus Group Study* (p. 251). Open Access Dissertations.

MacIver, R., Anderson, N. A., & Evers, A. (2014). Validity of Interpretation: A User Validity Perspective beyond the Test Score. *International Journal of Selection, 22,* 149-164. https://doi.org/10.1111/ijsa.12065

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational Measurement* (3rd ed., pp. 13-103). Macmillan.

O'Leary, T. M. (2018). *Effective Score Reporting: Establishing Evidence-Informed Principles for Outcomes Focused Score Reports*. Doctoral Dissertation.

O'Leary, T. M., Hattie, J. A., & Griffin, P. (2017). Actual Interpretations and Use of Scores as Aspects of Validity. *Educational Measurement: Issues and Practice, 36,* 16-23. https://doi.org/10.1111/emip.12141

Rankin, J. (2016). *Standards for Reporting Data to Educators: What Educational Leaders*

Should Know and Demand. Routledge. https://doi.org/10.4324/9781315623283

Ryan, J. M. (2006). Practices, Issues, and Trends in Student Test Score Reporting. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 677-710). Erlbaumb.

Slater, S., Livingston, S. A., & Silver, M. (2018). Score Reports for Large-Scale Testing Programs. In D. Zapata-Rivera (Ed.), *Score Reporting Research and Applications* (pp. 174-200). Routledge. https://doi.org/10.4324/9781351136501-8

Tannenbaum, R. J. (2018). Validity Aspects of Score Reporting. In D. Zapata-Rivera (Ed.), *Score Reporting Research and Applications* (pp. 33-52). Routledge. https://doi.org/10.4324/9781351136501-2

Van der Kleij, F. M., Eggen, T. J. H. M., & Engelen, R. J. H. (2014). Towards Valid Score Reports in the Computer Program LOVS: A Redesign Study. *Studies in Educational Evaluation, 43,* 24-39. https://doi.org/10.1016/j.stueduc.2014.04.004

Vanhoof, J., Verhaeghe, G., Verhaeghe, J. P., Valcke, M., & Van Petegem, P. (2011). The Influence of Competences and Support on School Performance Feedback Use. *Educational Studies, 37,* 141-154. https://doi.org/10.1080/03055698.2010.482771

Zapata-Rivera, D., & Katz, R. I. (2014). Keeping Your Audience in Mind: Applying Audience Analysis to the Design of Interactive Score Reports. *Assessment in Education: Principles, Policy & Practice, 21,* 442-463. https://doi.org/10.1080/0969594X.2014.936357

Zenisky, A. L., & Hambleton, R. K. (2012). Developing Test Score Reports That Work: The Process and Best Practices for Effective Communication. *Educational Measurement: Issues and Practice, 31,* 21-26. https://doi.org/10.1111/j.1745-3992.2012.00231.x

Zenisky, A. L., & Hambleton, R. K. (2015). Good Practices for Score Reporting. In S. Lane, M. R. Raymond, & T. M. Haladyna (Eds.), *Handbook of Test Development* (pp. 585-602). Routledge.