Educación XX1
ISSN: 1139-613X · e-ISSN: 2174-5374

# Formative evaluation and quality of feedback: design and validation of scales for school teachers

## *Evaluación formativa y calidad de la retroalimentación: diseño y validación de escalas para docentes escolares*

Juan Romeo Dávila Ramírez [1 *]
Juan Antonio Huertas Martínez [2]

[1] University of Tarapacá, Chile
[2] Autonomous University of Madrid, Spain
* Corresponding author. E-mail: jdavilar@academicos.uta.cl

**Cómo referenciar este artículo/ How to reference this article:**

Dávila Ramírez, J. R, & Huertas Martínez, J. A. (2024). Formative evaluation and quality of feedback: design and validation of scales for school teachers. *Educación XX1*, *27* (2), 167-194. https://doi.org/10.5944/educxx1.38283

**Date received:** 08/09/2023
**Date accepted:** 23/01/2024
**Published online:** 28/06/2024

## ABSTRACT

Formative evaluation has been described as one of the main pedagogical practices to promote the development of learning since it allows teachers and students to visualize the gaps between the level of mastery achieved and the level of mastery expected and direct their actions towards improvement. Despite this, the operationalization of the strategies that make up said methodology is still confusing and it has not been precisely defined what optimal use implies in terms of quality. The objective of this study is to design and validate a Strategies Scale for Formative Evaluation and a Feedback Quality Scale for Learning. The data of 364 primary and secondary teachers have been analyzed through exploratory factor analysis and confirmatory factor analysis with structural equation models. Then, a cross-validation analysis has been carried out for each scale between two random subsamples. Also, differences have been explored according to the educational level where the participating teachers worked, primary or secondary, and differences according to their gender. The results have indicated an adequate goodness of fit for the Formative Evaluation Strategies Scale: $\chi 2$ /df = 3.2, CFI = .91 and RMSEA = .07 and for the Quality of Feedback for Learning Scale $\chi 2$ /df = 1.8, CFI = .94 and RMSEA = .05. No significant differences were found according to educational level or gender. The discussion presents a heuristic model that illustrates the relationships between how formative evaluation and quality of feedback have been defined with learning and teaching processes, considering the influences exerted by classroom climates, school climates and national educational systems.


**Keywords:** formative evaluation, feedback, co-evaluation, self-evaluation


## RESUMEN

Se ha descrito a la evaluación formativa como una de las principales prácticas pedagógicas para promover el desarrollo de los aprendizajes ya que permite que docentes y estudiantes visualicen las brechas entre el nivel de dominio alcanzado y el nivel de dominio esperado y direccionen sus acciones hacia la mejora. Pese a ello, la operativización de las estrategias que conforman dicha metodología aún es confusa y no se ha definido con precisión qué implica un uso óptimo en cuanto a su calidad. El presente estudio tiene como objetivo diseñar y validar una Escala de Estrategias para la Evaluación Formativa y una Escala de Calidad de la Retroalimentación para el Aprendizaje. Los datos de 364 docentes de primaria y secundaria han sido analizados a través de análisis factoriales exploratorios y análisis factoriales confirmatorios con modelos de ecuaciones estructurales. Luego, se ha realizado un análisis de validación cruzada para cada escala entre dos submuestras aleatorias. También, se han explorado las diferencias según el nivel educativo donde se desempeñaba el profesorado participante, primaria o secundaria y las diferencias según su género. Los resultados han señalado una adecuada bondad de ajuste para la Escala de Estrategias de Evaluación Formativa: $\chi 2$ /gl = 3.2, CFI = .91 y RMSEA = .07 y para la Escala de Calidad de la Retroalimentación para el Aprendizaje $\chi 2$ /gl = 1.8, CFI = .94 y RMSEA = .05. No se han encontrado diferencias significativas según nivel educativo ni según género. La

discusión presenta un modelo heurístico que ilustra las relaciones entre como se ha definido evaluación formativa y calidad de la retroalimentación con procesos del aprendizaje y de la enseñanza, considerando las influencias que ejercen los climas de aula, los climas escolares y los sistemas educativos nacionales.

**Palabras clave:** evaluación formativa, retroalimentación, coevaluación, autoevaluación

## INTRODUCTION

### Learning and formative evaluation

Human learning has been defined as a meta-process of psychological transformation that emerges from the interaction between a person and an enriching social environment, as a result of internal processes of explicitation and implicitization, which include the integration of conceptual and motivational content (Illeris, 2014; Well, 2017). This is consistent with the notion of proximal development that emphasizes the role of expert support in learning (Vygotsky, 1979). Both arguments highlight the role of intersubjective processes in learning, which in the case of formal education occur between students and teachers (Greene et al., 2004). In schools, one of the consequences of these interactions is the information that arises about the quality of learning, that is, feedback, a constitutive process of formative evaluation.

In general, formative evaluation has been described as a set of pedagogical strategies that allows teachers and students to describe, analyze, evaluate, and guide the progression of student learning according to previously determined objectives (Lipnevich et al., 2016; Panadero et al., 2012; Shavelson et al., 2008). In this way, feedback constitutes a bridge between learning and formative evaluation, as well as between teaching and formative evaluation, as it provides teachers and students with information about their performance (Bond et al., 2020).

### Strategies for formative evaluation: teacher feedback, co-evaluation, and self-evaluation

Research on the quality of teaching has indicated that formative evaluation has a high impact on the development of student learning (Klute et al., 2017) whether in primary, secondary or higher education (López-Pastor and Pérez-Pueyo, 2017). The literature has highlighted various formative evaluation strategies (Moos and Brookhart, 2019). In this work we have chosen to collect three strategies that have been indicated in the literature as representative of formative evaluation: teacher

feedback, co-evaluation, and self-evaluation. (Bond et al., 2020; Popham, 2013). For now, other strategies have been left out, such as, for example, those that generate situations that demonstrate learning. Teacher feedback has been defined as a teacher's evaluation of a student's performance on a task. It has been shown that its main contribution lies in the fact that, along with informing the student what actions they should develop in a better way from identifying errors, it also guides on how to achieve a better performance (Panadero and Lipnevich, 2022). Co-evaluation refers to the evaluation of a student's performance that is carried out among classmates (Panadero et al., 2023). And self-evaluation indicates the evaluation that each student makes regarding their own performance (Harris and Brown, 2022). Brown and Harris (2013) and Sánchez et. al (2017) have reported that students who have participated in self-evaluation and co-evaluation processes have obtained better learning results in subsequent tests than those who did not, demonstrating the usefulness of both practices. Of the three strategies described, teacher feedback has a central role since it arises from an expert teacher, it guides the student's practice towards improvement and allows addressing the individual differences of the students (Andrade, 2023; Hooley and Thorpe, 2017).

## Quality of feedback for learning

The academic interest in understanding teacher feedback has increased, and various studies have analyzed which factors are associated with quality feedback. For example, Ossenberg et al., (2019) have analyzed 61 publications and have identified ten attributes of feedback that improve student performance, for example: that it is detailed and that it considers student needs. Adarkwah (2021) has conducted a scoping review, where he points out that the literature classifies feedback into two types: formative feedback, which describes the progress of student performance in qualitative terms, and summative feedback, which classifies performance based on a quantification of mastery through qualifications, usually for accreditation purposes. He has concluded that quality feedback is one that precisely describes which aspects of performance to improve and that clearly indicates what to do to move forward, beyond the quantification of learning. For their part, Tay and Lam (2022) have studied the impact of different feedback strategies on 75 high school students in Singapore. They have concluded that quality feedback promotes the student's commitment to the feedback received if it visualizes learning that will occur in the future and considers the emotional implication that it will have. Finally, Panadero and Lipnevich (2022) have analyzed 14 feedback models in a systematic review. From this work they have developed the MISCA model composed of five factors: Message, Implementation, Student, Context and Agents. This model emphasizes that quality feedback takes into account the characteristics of the students, who,

in addition to receiving feedback, are also producers of feedback for themselves in self-evaluation and producers of feedback for their classmates in co-evaluation. Regarding the context, it has been indicated that quality feedback requires that each educational center promotes positive school climates and classroom climates oriented to learning (Heritage, 2010), since effective feedback favorably affects student performance if it manages to positively affect students. their motivational states (Rowe, 2017). Likewise, it has been pointed out that national educational systems can favor the development of formative evaluation, and within it feedback, or hinder its implementation (Bond et al., 2020; van der Kleij et al., 2018; van der Kleij and DeLuca, 2023).

### Justification of the research

Although the works described offer a valuable contribution to the understanding of formative evaluation and the quality of feedback, the number of studies on the subject tends to be greater in higher education than in primary and secondary education (Alqassab et al., *2023;* Sánchez et al., 2017). On the other hand, the operationalization of feedback processes is still confusing and requires greater conceptual organization (Hortigüela et al., 2019; Van der Kleij et al., 2018). Probably, the above is related to the fact that in some educational contexts feedback is used for several objectives, whether formative, summative (grading with grades) or mixed, which in some way constitutes an obstacle to monitoring the use of this strategy and to know its impact on learning (Adarkwah, 2021). This work seeks to respond to the challenges described. The main objective of this research is to design two teacher self-report scales with a clear theoretical structure and empirical validity. A scale to evaluate the implementation of Formative Evaluation Strategies and another scale to evaluate the Quality of Feedback for Learning that teachers offer to students in classes. The development of both scales can be useful for educational centers to initiate or develop processes of understanding the processes of formative evaluation and feedback of learning or for the development of advisory programs in formative evaluation for teachers in training and in practice. (Matthews et al., 2023; Pat-El et al., 2013; Shavelson et al., 2008).

## METHOD

### Participants

The group of participants is made up of 364 teachers who teach in the Tarapacá Region, Chile. Teachers teach various subjects in courses from 1st to 12th grade. The

average age is 38 years. 242 women (66.5%) and 122 men (33.5%) participated. They belong to a total of 13 schools, 9 private schools with public financing and 4 public schools. 278 teachers (76.4%) work in private schools with public financing and 86 work in public schools (23.6%). In 2021, the total amount of schoolteachers in the Tarapacá Region has corresponded to 4,607, 1.8% of the total school teachers in Chile and they serve approximately 85,200 students (MINEDUC Study Center, 2022). Most of the participating teachers belong to schools that are oriented towards teaching innovation and educational improvement, an aspect formally recognized by Chilean public educational policies. Although this may constitute a limitation, it has been considered appropriate to validate these scales with teachers oriented towards educational improvement to adequately explore the validity of complex constructs that are related to said improvement.

## Instrument design

To provide theoretical content for the design of the Formative Evaluation Strategies Scale (E3F), the conceptualization of Popham (2013) and Shavelson et al. (2008). This framework has guided the construction of items distributed in three subscales: Teacher feedback, Co-evaluation, and Self-evaluation. The authors of the present work, considering this theoretical framework, have developed a set of items that describe actions typical of such strategies, that consider evaluation procedures such as the use of rubrics, guidelines, or tutorials (Andrade, 2023), and that emphasize the role of personal reflection or that showed the opposite version of traditional evaluation systems.

For the design of the Feedback for Learning Quality Scale (ECRA , acronym in the spanish language, Escala Calidad de la Retroalimentación para el aprendizaje). Four reviews of the literature have been selected, according to the following criteria: they are theoretical systematizations or systematic reviews, they were published in the last five years, and they consider the impact of the teaching feedback processes in the student. These articles have been described in the introductory section. In Table 1 a synthesis has been made from the contents of these four works in dimensions that are consistent with the way in which current literature indicates the objectives of feedback: performance, motivation, and self-regulation of learning (Lipnevich and Panadero, 2021). In this synthesis it is possible to observe the configuration of three dimensions: an instructional dimension that emphasizes the transmission of the message and the content of the feedback, an interactional dimension that highlights the value of motivation and empathizing with the actions and affections of the students. and a self-regulatory dimension aimed at promoting students to reflect on their own involvement and performance in learning tasks.

Taken together, both conceptual frameworks offer a solid theoretical structure and support the design of an item base for each scale. In the case of the Formative Evaluation Strategies Scale, an initial base of 12 items was built, four items for three subscales. Regarding the Quality of Feedback for Learning Scale, an initial base of 22 items distributed in three subscales was developed, seven items representative of the instructional dimension, seven items representative of the interactional dimension and eight items representative of the self-regulatory dimension. Both scales contain direct and inverse items, have a Likert-type response format with ranges from 1 to 5 and express *completely disagree* to *completely agree.*

## Procedure

Initially, the management teams of each school have been contacted and they have been proposed to participate in the research project. The teaching staff participated voluntarily and approved an informed consent that accounted for the anonymization and confidentiality of the data. These procedures have been approved by the Research Ethics Committee of the Autonomous University of Madrid, report CEI-125-2566. Data collection has been the same in each of the educational centers. The teachers have met in a place equipped with computers and have individually answered the scales on a virtual platform.

## Data analysis and validation procedures

Firstly, exploratory factor analyzes (EFA) have been carried out that have guided the configuration of the subscales of both scales. The EFAs have been carried out on random subsamples corresponding to half of the total sample size (N = 182). For this, the principal components method with Equamax rotation has been used. For the identification of factors, factor loading values equal to or greater than .40 have been considered. Subsequently, the indicators of the KMO test, Bartlet's sphericity test and the total explained variance of each scale and its subscales were obtained (Lloret-Segura, 2014). This has allowed the selection of items and the factorial configuration for both scales.

Secondly, in the remaining random subsamples of half of the total sample size (N = 182) the items selected for both scales have been analyzed using a confirmatory factor analysis (CFA) with structural equation models (SEM) based on original designs, using the maximum likelihood (ML) method, and following the considerations of Ruiz et. al (2020). Then, in each scale, the mean of each item, its standard deviation, the standard error, and the *t*- test statistic have been obtained for a sample that has compared the mean obtained with the central value of the range of responses.

Next, for the data set of each scale, Mardía's multivariate normality index has been obtained, using the asymmetry and kurtosis indicators, their critical value and range. Once the final models were defined, the Cronbach's alpha and McDonald's omega statistics were obtained. Consequently, the goodness of fit and the explained variance of each instrument have been determined. Afterwards, a cross-validation analysis was carried out for each scale between two random subsamples of half the total sample size. Finally, the differences have been analyzed according to the educational level where the participating teachers work, primary or secondary, and differences according to gender. The statistical package IBM SPSS and Amos version 28 were used for data analysis.

## RESULTS

### Exploratory factor analyzes

An EFA has been carried out with the 12 items of the Formative Evaluation Strategies Scale, with the objective of identifying the contribution of each item to its unifactorial configuration. This analysis has presented a total explained variance of 27.9%, a KMO value = .76 and a *p value* of the Bartlet test = $p$ <0.01. There it has been pointed out that two items have not reached the factorial weight of .40, E3F_06 = .33 and E3F_12 = .38 and have been eliminated from the set. Subsequently, a new EFA was carried out, without these two items, and it showed a total explained variance of 31.6%, a KMO value = .76 and a *p value* of the Bartlet test = p<0.01. Then, in a random subsample composed of half of the cases of the total sample size (N = 182), with the 10 items selected, a new EFA has been carried out considering the configuration of three factors, a total explained variance of 57.3%. a KMO value = .71 and a Bartlet test *p value* = $p$ < 0.01. In this EFA, item E3F_01 has been eliminated since it is the inverse item of item E3F_07 and both appear in the same factor. In this way, with the 9 selected items, a new EFA has been carried out under the same conditions as the previous one. This EFA has shown a total explained variance of 59.1%, a KMO value = .69 and a Bartlet test *p value* = $p$ < 0.01. Factor 1 contains statements that refer to teacher feedback and self-evaluation. Factor 2 has items that refer to co-evaluation. And Factor 3 fundamentally refers to the use of rubrics. The result was a factor that explained 31% of the variance that explained teacher feedback and self-evaluation and a factor that explained 46.9% of the variance that was related to co-evaluation. For the CFA, along with the factor weights, the assignment of the item has been considered according to the theoretical model that was used to construct the questionnaire. This has been done to decide the assignment in items E3F_04, E3F_07 and ECRA_09R and E3F_05R.

**Table 1**
*Exploratory factor analysis of the Formative Evaluation Strategies Scale*

| Item | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| E3F_04: In each evaluation I use some procedure - rubric, guideline, tutoring - that informs the student of the level of mastery they achieved in completing a task. | | | .60 |
| E3F_03R: I don't usually inform my students about what they can do to learn better. | .77 | | |
| E3F_10R: I find it very difficult to inform each of my students about what mistakes they made in the evaluations. | .65 | | |
| E3F_08R: I prevent my students from commenting on their classmates' work during classes. | | .83 | |
| E3F_07: I ask my students to evaluate the work of their classmates' using rubrics or guidelines. | | .52 | .59 |
| E3F_11R: I avoid asking my students to analyze the performance of their classmates. | | .84 | |
| E3F_09R: I prefer to prevent my students from evaluating themselves. | .63 | | .39 |
| E3F_02: I ask my students to evaluate their own performance using rubrics or guidelines. | | | .73 |
| E3F_05R: I do not use self-evaluation because my students rate themselves higher than is appropriate. | .54 | | .47 |
| Explained variance | 31.2% | 15.6% | 12.2% |

*Note*. *N* = 182.

Regarding the Feedback Quality Scale, an EFA has also been carried out. Initially, in a random subsample with half of the cases (*N* = 182), an EFA has been carried out with the items of each subscale separately, based on the synthesis of Table 1. For this, acceptable values of factor loading greater than .50. Of the 7 items that corresponded to the instructional dimension, three of them have not reached the indicated factorial weight and have been eliminated from this set (ECRA_09R = .28, ECRA_19R = .46 and ECRA_20R = .46). The EFA of this subscale has a KMO value = .66, a Bartlet's p value = p <0.01 and a total explained variance of 32.5%. Of the 7 items that corresponded to the interactional dimension, two of them have not reached the indicated factorial weight and have been eliminated (ECRA_10R = .13 and ECRA_15R = .47). The EFA of this subscale presents a KMO value = .65, a Bartlet's *p value = p* <0.01 and a total explained variance of 31.4%. Finally, of the 8

items in the self-regulatory dimension, two of them have not reached the indicated factorial weight and have been eliminated (ECRA_12 = .42, ECRA_21 = .48). The EFA of this subscale presents a KMO value = .76, a Bartlet's *p value* = *p* <0.01 and a total explained variance of 37.5. With a total of 15 items, a new CFA was carried out with a KMO value = .81, a Bartlet's *p value* = *p* <0.01, and a total explained variance of 48.7%. In this analysis, items ECRA_22R, ECRA_13 and ECRA_2R have been eliminated because they are inverse items of items ECRA_16, ECRA_17 and ECRA_3, correspondingly. The factor loadings of the items in the three EFA factors of the total sample are very similar to those shown in this subsample.

**Table 1**

*Synthesis of variables of the quality of teacher feedback for learning based on the comparison of previous research*

| Literature reviews on feedback from teachers to students | | | | Dimensions |
|---|---|---|---|---|
| **Ossenberg et. al (2019)** | **Adarkwah (2021)** | **Tay and Lam (2022)** | **Baker and Lipnevich (2022)** | **Result of variable synthesis** |
| It is part of a process | It is detailed | Provides specific and general feedback on performance on a task | Message: The information contained in the feedback message must be clear and useful | Instructional dimension<br>Performance oriented |
| It is based on criteria | It is clear and simple | | | |
| Use multiple sources of evidence | | | Context: the pedagogical approach and the mode of delivery of feedback favor involvement | |
| It is frequent | | | | |
| Involves skillful interaction | Guides student action towards improvement | Guides and motivates practice towards improvement | | |
| Involves the student in more ways than one | Use an appropriate tone of voice | Pay attention to the emotional response of the student | Student: consider individual differences, motivational beliefs, prior knowledge, gender, cultural differences, self-efficacy, etc. | Interactional dimension<br>Motivation oriented |
| Adapts to the needs of the student | It is appropriate to the demands of the student | | | |
| It is welcomed by the student | | | | |
| Is it bidirectional or reciprocal | | | | |
| Focuses on the future | | Pay attention to future learning evaluation processes | Implementation: connects the instructional context with the internal processing of the student (self-regulation) | Self-regulatory dimension<br>Self-regulation oriented |
| It is timely | It is timely | It is timely | | |
| | | | Agents: promotes the integration of the teacher, peers, and the learner | |

*Note.* Own elaboration based on the works indicated.

**Table 2**

*Exploratory factor analysis of the Quality of Feedback for Learning Scale*

| Item | Factor 1 | Factor 2 | Factor 3 |
|---|---|---|---|
| ECRA_01: I show students how to do or execute the learning tasks that I propose to them. | .59 | | |
| ECRA_07: My comments or observations are clear, easy to understand and contain instructions on what to do to achieve the learning objectives. | .70 | | |
| ECRA_08: When I inform a student about the result of their homework, I use a fraternal, respectful, and cordial tone. | .45 | .49 | |
| ECRA_18: My comments on student performance are specific and address important actions that can be improved. | .66 | | |
| ECRA_05: I consider the individual characteristics of my students (social context, personality characteristics, special needs) to comment on their work based on their particularities. | .46 | .36 | |
| ECRA_04: I almost always put myself in my students' shoes and try to think about how they approach the learning tasks that I propose to them. | .40 | .58 | |
| ECRA_03: I am willing to receive comments from my students about the results of their evaluations. | | .79 | |
| ECRA_06R: I avoid thinking about how my students feel in the learning tasks that I propose to them. | | .56 | .51 |
| ECRA_11R: I spend little effort asking the expectations that students have about the learning tasks that I propose. | | | .82 |
| ECRA_14R: I almost never ask the student what he thinks about his own performance in executing the tasks. | | | .75 |
| ECRA_16: I don't let much time go by to give my students observations or comments on their work. | .32 | | .31 |
| ECRA_17: I tell students what procedures or tasks they performed correctly so that they recognize their achievements. | .54 | | .44 |
| Explained variance | 33.2% | 10.4% | 9.0% |

*Note* . *N* = 182.

With 12 items, a new EFA has been carried out aimed at knowing the trifactoriality. The method of principal components and Equamax rotation has been used. For the CFA, along with the factor weights, the assignment of the item has been considered according to the theoretical model that was used to construct the questionnaire. This has been done to decide the assignment, especially in items ECRA_6, ECRA_16 and ECRA_17. This structure is confirmed with the EFA of the total sample. This analysis presents a KMO value = .83, a Bartlet's *p value* = p<0.01 and a total explained variance of 52.7%. Table 2 shows the items of Factor 1, of the instructional dimension and that make up the focused instruction subscale, the items of Factor 2 that correspond to the interactional dimension and that make up the empathic interaction subscale, and the items of Factor 3 that They come from the self-regulatory dimension and make up the self-regulation subscale of student performance.

## Confirmatory factor analysis: goodness of fit, explained variance and descriptive analysis

Based on the EFA described in Table 1, a CFA has been carried out with the 9 items selected from the Formative Evaluation Strategies Scale in a random subsample with the other half of the cases in the total sample (*N* = 182). An original model was obtained that presents indicators close to the following reference criteria for the goodness of fit of SEM models: $\chi 2$ /df < 3, CFI ≥ .95 and RMSEA ≤ .05. The goodness-of-fit indicators presented by the SEM model of said scale are acceptable: $\chi 2$ /df = 1.3, CFI = .97 and RMSEA = .04; CI (.05 - .11). The resulting structural model is presented in Figure 1 and contains three latent variables that represent the subscales: Teacher feedback, composed of three items; Co-evaluation, composed of three items, Self-evaluation, composed of three items and a second-order latent variable.

Regarding the Quality of Feedback for Learning Scale, based on the previous EFA, a CFA has also been carried out on a random subsample with this half of the cases of the total sample (*N* = 182). This analysis has considered the 12 items presented in Table 2. In this way, an original model was obtained that satisfactorily responds to the goodness of fit indicators described in the preceding paragraph: $\chi 2$ /df = 1.2, CFI = .96 and RMSEA = .03; CI (.00 - .08). The resulting structural model is presented in Figure 2. It contains three latent variables that represent the subscales: Focused instruction, which emerges from the instructional dimension and contains five items, Empathic interaction, which emerges from the interactional dimension and contains three items, and Self-regulation of student performance, which emerges from the self-regulatory dimension, and which contains four items; and a second-order latent variable.

Figure 1 presents the standardized weights and the coefficients of determination or $R^2$ of each item and each latent variable of the three subscales of the Formative Evaluation Strategies Scale. Most of the items exceed the standardized weight indicator of .40, except for items E3F_03R and E3F_10R. The $R^2$ values in the subscales are high, being lower in the case of the Co-evaluation strategy with an indicator of .73. The means and values of the *t test* present in Table 3 indicate values that are significantly above the central value of the range of responses.

**Figure 1**

*Structural model of the Formative Evaluation Strategies Scale*

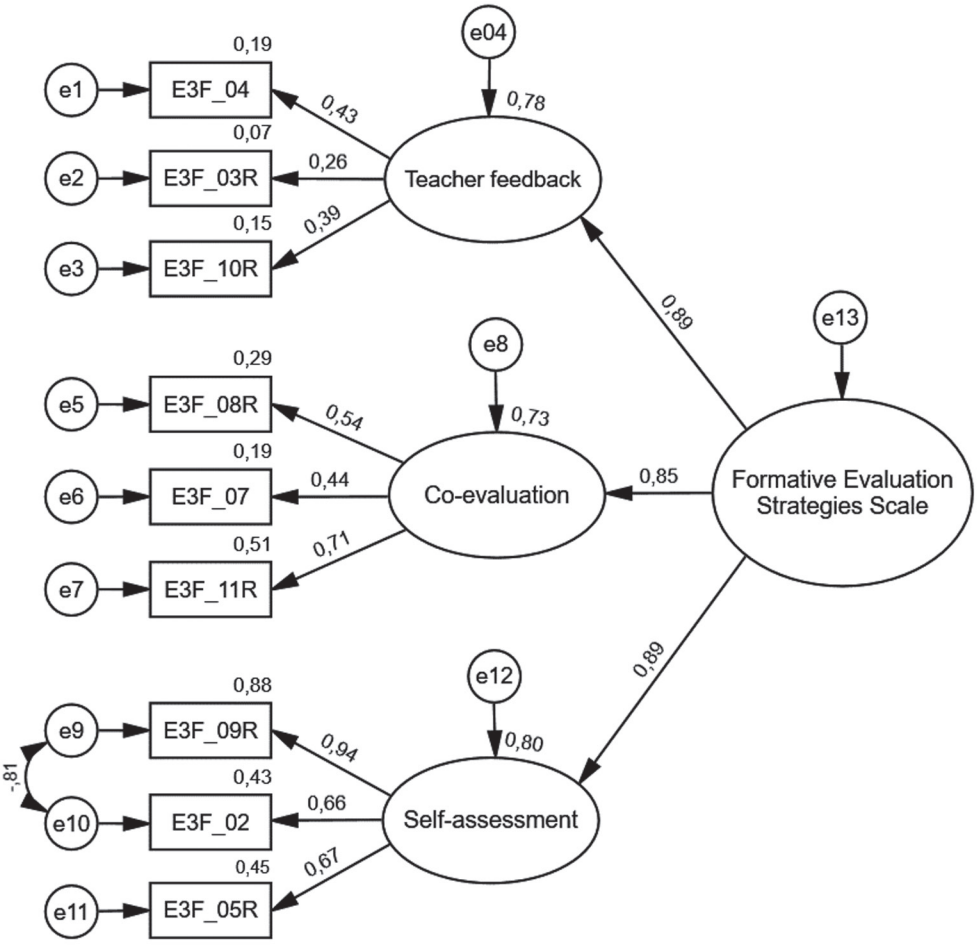


**Table 3**

*Means, standard deviations, standard error, and Student's t-test statistics for a sample of the Formative Evaluation Strategies Scale*

| Item | *M* | SD | SE | *t* | *gl* | *p* |
|---|---|---|---|---|---|---|
| E3F_04 | 4.3 | .76 | 0.05 | 23.8 | 181 | <.001 |
| E3F_03R | 4.4 | .83 | 0.06 | 22.4 | 181 | <.001 |
| E3F_10R | 3.8 | 1.0 | 0.08 | 10.9 | 181 | <.001 |
| E3F_08R | 3.4 | 1.1 | 0.08 | 5.12 | 181 | <.001 |
| E3F_07 | 3.1 | 1.2 | 0.09 | 2.02 | 181 | .045 |
| E3F_11R | 3.4 | 1.1 | 0.08 | 5.57 | 181 | <.001 |
| E3F_09R | 3.9 | 1.0 | 0.07 | 11.7 | 181 | <.001 |
| E3F_02 | 3.9 | 1.0 | 0.07 | 12.0 | 181 | <.001 |
| E3F_05R | 3.8 | 1.0 | 0.08 | 10.6 | 181 | <.001 |
| E3F_RT | 4.2 | .61 | 0.04 | 26.4 | 181 | <.001 |
| E3F_CV | 3.6 | .91 | 0.06 | 9.02 | 181 | <.001 |
| E3F_AV | 3.8 | .93 | 0.06 | 12.8 | 181 | <.001 |
| E3F Total | 3.9 | .58 | 0.04 | 20.9 | 181 | <.001 |

*Note.* $N$ = 182 . $M$ = Mean; SD= Standard Deviation; SE= Standard Error; $t$ = t value; $p$ = *Student's t* test for one sample with test value = 3 and significance level .05; E3F_RT = Teacher feedback subscale; E3F_CV = Co-evaluation subscale; E3F_AV = Self-evaluation subscale; E3F = Formative Evaluation Strategies Scale.

**Figure 2**
*Structural model of the Quality of Feedback for Learning Scale*

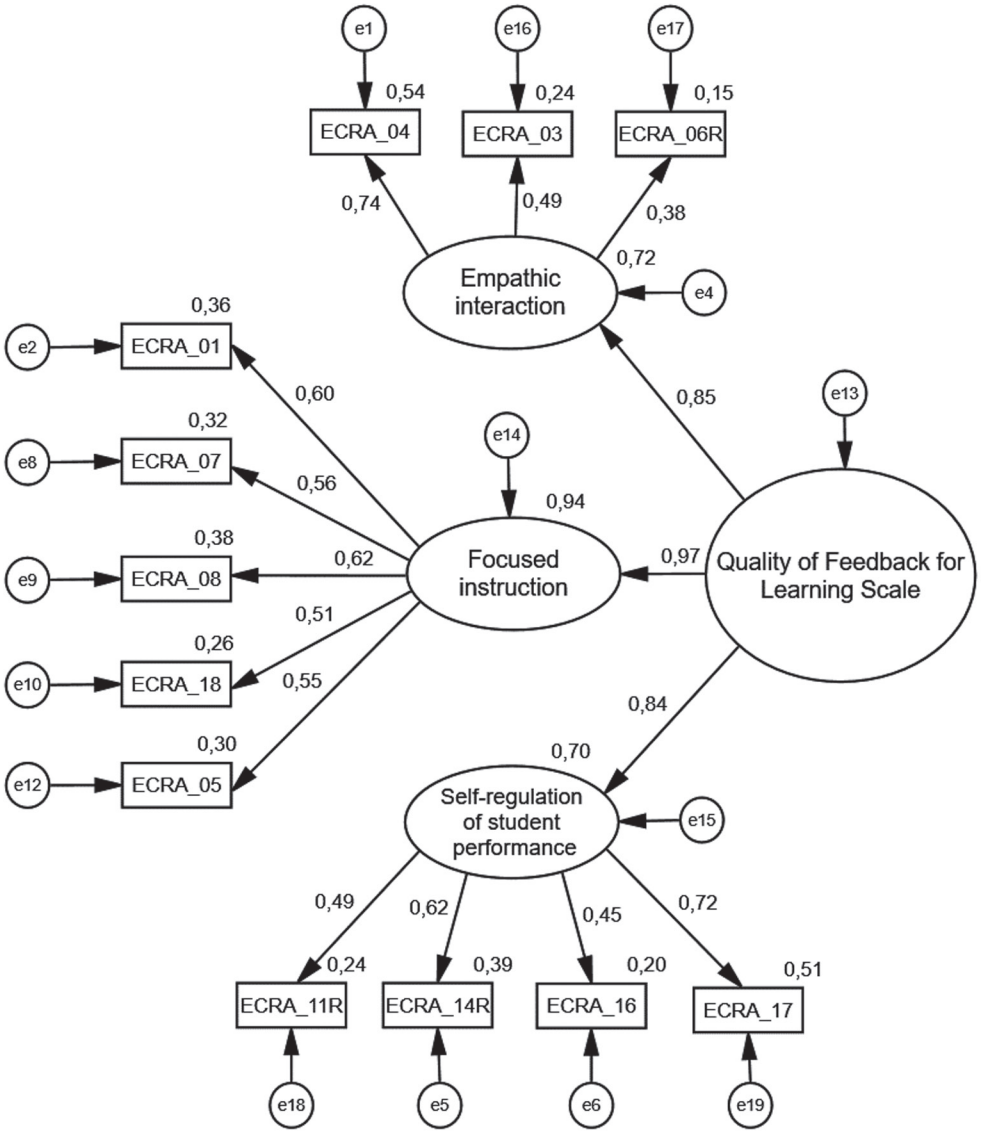


**Table 4**

*Means, standard deviations, standard error, and one-sample Student t-test statistics for the Quality of Feedback for Learning Scale*

| Item | *M* | SD | SE | *t* | *gl* | *p* |
|---|---|---|---|---|---|---|
| ECRA_01 | 4.7 | .49 | .03 | 60.0 | 181 | <.001 |
| ECRA_07 | 4.3 | .73 | .03 | 37.8 | 181 | <.001 |
| ECRA_08 | 4.7 | .55 | .03 | 65.3 | 181 | <.001 |
| ECRA_18 | 4.2 | .64 | .04 | 31.7 | 181 | <.001 |
| ECRA_05 | 4.5 | .97 | .03 | 43.8 | 181 | <.001 |
| ECRA_04 | 4.5 | 1.0 | .03 | 52.6 | 181 | <.001 |
| ECRA_03 | 4.6 | .64 | .03 | 49.1 | 181 | <.001 |
| ECRA_06R | 4.3 | 1.0 | .05 | 24.9 | 181 | <.001 |
| ECRA_11R | 3.9 | .96 | .05 | 16.1 | 181 | <.001 |
| ECRA_14R | 4.1 | .62 | .05 | 20.8 | 181 | <.001 |
| ECRA_16 | 4.0 | .42 | .05 | 19.1 | 181 | <.001 |
| ECRA_17 | 4.5 | .47 | .03 | 46.0 | 181 | <.001 |
| ECRA_IF | 4.5 | .64 | .02 | 67.4 | 181 | <.001 |
| ECRA_IE | 4.5 | .41 | .02 | 60.6 | 181 | <.001 |
| ECRA_ADE | 4.1 | .49 | .03 | 32.6 | 181 | <.001 |
| ECRA Total | 4.3 | .73 | .02 | 62.5 | 181 | <.001 |

*Note.* $N$ = 182. $M$ = Mean; SD= Standard deviation; SE = standard error; $t$ = t value; $p$ = *Student's t* test for one sample with test value = 3 and significance level .05; ECRA_IF = Focused instruction subscale; ECRA_IE = Empathic interaction subscale; ECRA_ADE = Self-regulation of student performance subscale; ECRA = Quality of Feedback for Learning Scale.

Figure 2 presents the standardized weights and the coefficients of determination or $R^2$ of each item and each latent variable of the three subscales of the Quality of Feedback for Learning Scale. It is observed that most of the items exceed the standardized measurement weight indicator of .40, except for item ECRA_06 = .38. The $R^2$ values in the subscales are high, being lower in the case of the self-regulation of student performance subscale with an indicator of .70. The means and the values of the Student's *t test* for one sample presented in Table 4 indicate that the scores are significantly above the central value of the range of responses.

## Normality analysis

Mardía's (1974) multivariate normality indicators show that the distribution of the E3F is close to normal and that the ECRA is not normal. In the case of asymmetry, the value of the statistic is greater than its critical value. Regarding kurtosis, the value of the statistic is not within the critical range established by Mardía according to the sample size (Wulandari et al., 2021). Since SEM with ML requires normality, the *p value* of the Bollen-Stine (BS) index was obtained with two thousand bootstraps (Cheung and Lau, 2008; Fan, 2003). A *p- value* of .06 for the BS index has been obtained. The *p* indicators of the BS Index correct the detected abnormality by exceeding the *p value* of .05.

## Reliability analysis

Cronbach's Alpha values indicate acceptable or optimal internal consistency reliability since they are equal to or greater than .70. The Formative Evaluation Strategies Scale has obtained a Cronbach's alpha of .71 and the Quality of Feedback for Learning Scale has obtained a Cronbach's alpha of .78. The values of McDonald's omega indicator are identical to those of Cronbach's alpha in both scales. The correlations between scales and subscales show the absence of collinearity and a significant relationship between the subscales of each scale (Table 5).

**Table 5**

*Multiple correlation statistics between scales and subscales*

| Variable | E3F | E3F_RT | E3F_CV | E3F_AV | ECRA | ECRA_IF | ECRA_IE |
|---|---|---|---|---|---|---|---|
| E3F_RT | .69* | | | | | | |
| E3F_CV | .78* | .25* | | | | | |
| E3F_AV | .74* | .38* | .36* | | | | |
| ECRA | .42* | .51* | .19* | .29* | | | |
| ECRA_IF | .31* | .44* | .10* | .20* | .79* | | |
| ECRA_IE | .31* | 30* | .17* | .25* | .70* | .38* | |
| ECRA_ADE | .37* | 4.5* | .18* | .24* | .86* | .50* | .44* |

*Note*. *N* = 364. *The correlation is significant at the two-sided .01 level; E3F = Formative Evaluation Strategies Scale; E3F_RT = Teacher feedback subscale; E3F_CV = Co-evaluation subscale; E3F_AV = Self-evaluation subscale; ECRA = Learning Feedback Quality Scale ; ECRA_IF scale = Focused instruction subscale; ECRA_IE = Empathic interaction subscale; ECRA_AD = Self-regulation of student performance subscale.

## Cross validation

By segmenting the total sample into two random subsamples of equal size, goodness-of-fit indicators close to satisfactory criteria are obtained on both scales (Table 6).

**Table 6**
*Confirmatory factor analyzes with random subsamples*

| E3F | | | | | ECRA | | | | |
|---|---|---|---|---|---|---|---|---|---|
| **Model** | **χ2 /df** | **CFI** | **RMSEA** | **C.I.** | **Model** | **χ2 /df** | **CFI** | **RMSEA** | **C.I.** |
| E3F-1 (*N* =182) | 2.5 | .91 | .08 | (.05 - .11) | ECRA-1 (*N* =182) | 1.2 | .96 | .03 | (.00 -.06) |
| E3F-2 (*N* =182) | 2.2 | .90 | .08 | (.05 - .01) | ECRA-2 (*N* =182) | 1.8 | .90 | .06 | (.04 -.09) |
| E3F (*N* =364) | 3.2 | .91 | .07 | (.05 - .09) | ECRA (*N* =364) | 1.8 | .94 | .05 | (.03 -.06) |

*Note.* E3F = Formative Evaluation Strategies Scale; ECRA = Quality of Feedback for Learning Scale; CI = Confidence interval

## Analysis of differences according to educational level

Based on the analysis of the ANOVA test, the existence of significant differences between the means of both scales and the three educational levels where the participating teachers work have been explored, primary 1st to 4th, primary 5th to 8th and secondary 9th to 12th. The results indicate that there is no significant difference according to educational level, in both scales (Tables 7 and 8).

**Table 7**
*Statistics of the ANOVA test for the comparison of means according to educational level*

| Scale | Level | *M* | OF | *N* | F | *p* |
|---|---|---|---|---|---|---|
| Formative Evaluation Strategies Scale | Primary 1st to 4th | 3.8 | .60 | 81 | .97 | .37 |
| | Primary 5th to 8th | 3.9 | .59 | 110 | | |
| | Secondary 9th to 12th | 3.9 | .59 | 173 | | |
| | Total | 3.8 | .59 | 364 | | |

*Note.* M = Mean; SD = Standard deviation; N = Number of participants; F = ANOVA statistic; p = p value.

**Table 8**
*Statistics of the ANOVA test for the comparison of means according to educational level*

| Scale | Level | *M* | SD | *N* | F | *p* |
|---|---|---|---|---|---|---|
| Quality of Feedback for Learning Scale | Primary 1st to 4th | 4.3 | .41 | 81 | .124 | .94 |
| | Primary 5th to 8th | 4.3 | .42 | 110 | | |
| | Secondary 9th to 12th | 4.3 | .40 | 173 | | |
| | Total | 4.3 | .41 | 364 | | |

*Note.* M = Mean; SD = Standard deviation; N = Number of participants; F = ANOVA statistic; p = p value.

## Analysis of differences according to gender

To check the impact of the gender of the teachers on the means of both scales, analyzes were carried out with the Student's t test for independent samples. The results have indicated that in both scales the difference in means between male and female teachers is not significant (Tables 9 and 10).

**Table 9**
*Statistics of the Student t test for the differences in means according to gender*

| Scale | Gender | *M* | OF | *N* | F | *p* |
|---|---|---|---|---|---|---|
| Formative Evaluation Strategies Scale | Women | 3.9 | .58 | 242 | .27 | .60 |
| | Men | 3.7 | .61 | 122 | | |
| | Total | 3.8 | .59 | 364 | | |

**Table 10**
*Student t-test statistics for the difference in means by gender*

| Subscale | N | *M* | OF | *N* | F | *p* |
|---|---|---|---|---|---|---|
| Quality of Feedback for Learning Scale | Women | 4.3 | .38 | 242 | 1.7 | .18 |
| | Men | 4.2 | .44 | 122 | | |
| | Total | 4.3 | .41 | 364 | | |

## DISCUSSION

The average scores of both scales indicate that the group of teachers reports carrying out a formative evaluation and feedback of acceptable quality. A predictable

result, considering that the participating teachers work in schools that are oriented towards teaching innovation. Both instruments present adequate goodness of fit, reliability and correlation between them. The absence of differences in both scales, between the levels of education (primary and secondary) and by gender, allows us to conclude that the instruments manage to investigate the constructs in a transversal way.

The Formative Evaluation Strategies Scale presents three subscales that are representative of the construct. Feedback, co-evaluation, self-evaluation, and the use of tools such as rubrics are distinguished in this scale. This is an important aspect since these strategies will be formative if they manage to guide the progress of student learning beyond grades (Andrade, 2023; López-Pastor and Pérez-Pueyo, 2017). Consequently, these results are consistent with an understanding of formative evaluation that has highlighted its ability to describe, analyze, evaluate and direct learning progress, while allowing teachers and students to identify gaps between the level of mastery achieved and the level of expected mastery. It demonstrates the student's performance from multiple sources, turns one's own learning and the learning of one's classmates into an object of personal reflection and guides performance toward permanent improvement; either informally or spontaneously, planned for interaction in the classroom or in a more formal way or integrated into the curriculum ( Lipnevich et al., 2016; Panadero et al., 2012; Shavelson et al., 2008).

The Quality of Feedback for Learning Scale identifies three subscales that refer to dimensions of the quality of the feedback that teachers offer to students. The *focused instruction subscale* indicates that providing quality feedback implies offering comments that are clear, easy to understand and respectful, considering the individual characteristics of the student and indicating what specific actions or elements can be addressed to improve student performance on the task. The *empathic interaction subscale* emphasizes the need to recognize the impact that feedback could have on the students' motivation, both in their actions and their affects, and the relevance of listening to their impressions about the teaching process. Finally, the self-regulation subscale of student performance refers to how quality feedback encourages the student to identify their level of commitment to improvement and to reflect on their own performance, provides comments temporally close to the task to enhance the perceived usefulness, and visualizes the correct or optimal performance of the students in the task. These results agree with aspects that the literature highlights are basic to defining what to provide feedback and how to do it (Panadero and Lipnevich. 2022). Although the Quality of Feedback for Learning Scale presents these subscales as an expression of the quality of teaching feedback, in the teaching exercise, teachers could evaluate which of these factors they can emphasize, depending on the progress of the learning student at mastery levels or depending on the teaching situation. Thus, as evidenced in the

results of this work, quality feedback is specific, indicates what and how to improve, is respectful of the individual characteristics of the student, is empathetic with the student experience and ensures that students increase their self-regulation. by inviting you to reflect on your own performance in learning tasks (Adarkwah, 2021; Ossenberg et al., 2019; Rowe, 2017; Tay and Lam, 2022).

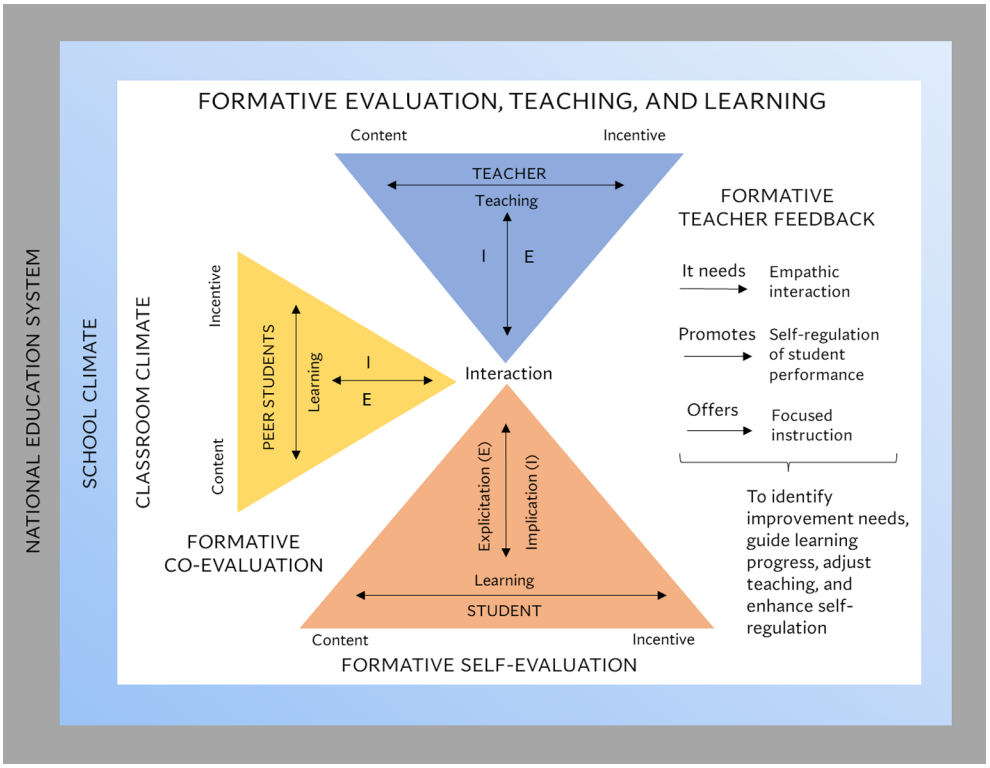
The theoretical scope of this research allows us to visualize the connection between the variables studied. From that interaction, concepts (content) and motivations (incentives) are mobilized that nourish and guide performance and its meaning, both among students and teachers (Illeris, 2014). The symbolic contents, data, or information about one's own performance in learning activities are internalized through implicitization mechanisms and externalized through processes of explicitation or transmission that are inherent to corporality (Pozo, 2017). In this process, the formative evaluation strategies organized by teachers direct, encourage, and nourish - from the context - progress in the domain of student learning, since they make feedback a transformative communicative interaction, whether in the form of dialogue intrapersonal (internal feedback) or from interpersonal dialogue; by its contents (focus on the task), by its form (empathy) and by its mechanism (self-regulation). All of this is favored by adequate school climates and classroom climates (Heritage, 2010; Matthews et al., 2023; Pat-El et al., 2013). As van der Kleij and DeLuca (2023) point out, public educational policies can favor the implementation of formative evaluation by promoting teacher training programs that focus on improving school autonomy, but they can hinder its implementation by generating very extensive curricula or not caring about from teacher burnout or the improvement of school infrastructure (van der Kleij et al., 2018).

## Limitations

The teachers who participated in this research belong to the Chilean educational context so that the extrapolation of these results to other contexts must be done with precautions and taking into consideration sociocultural similarities. The measurement process is transversal and includes the self-report of teachers who mostly belong to schools that are formally oriented towards quality and educational improvement. Only data from teachers are collected and analyzed, which invites us to explore the impact of these constructs on student measurements (Adarkwah, 2021).

**Figure 3**

*Explanatory diagram of the connection between learning, teaching, and formative evaluation*




Note. Own elaboration. Heuristic model that connects the results of this research and the contributions of Illeris (2014) and Pozo (2017).

## Future developments

It is proposed to study the predictive validity of both scales with measurements that incorporate student processes, longitudinally, multilevel, in different sociocultural contexts and with a greater number of participants. Likewise, it is suggested to continue researching these scales in digital formative evaluation processes (Hooley and Thorpe, 2017) and in higher education. Likewise, it will be favorable to add other formative evaluation strategies, for example, those related to the production of evidence of learning and the shared construction of evaluation criteria or mastery levels with the students. This will allow us to highlight the value of the interpersonal and pedagogical interaction present in formative evaluation,

as one of the main drivers of human learning, given its motivational, conceptual, behavioral and identity role (Illeris, 2014; Pozo, 2017; Vygotsky, 1979).

## CONCLUSIONS

This research contributes to advancing on the path towards a better-defined formative evaluation methodology, particularly in primary and secondary school educational contexts and from teacher self-report. Two brief scales have been designed and validated that identify elements that are central, both in formative evaluation and in the quality of teaching feedback. Furthermore, both instruments allow self-reporting of teachers' actions or beliefs that can be related to formative evaluation and to more spontaneous or informal feedback and to that which presents higher levels of structure (Bond et al., 2020). Both scales can be useful for new research, for the implementation of advisory programs for teachers in training and in practice (Matthews et al., 2023; Pat-El et al., 2013) and for the development of skills or competencies in the student (Shavelson et al., 2008).

## FINANCING

This work has been funded by the National Research and Development Agency (ANID). Scholarship program. Doctorate Scholarships Chile 2019 – 72200107.

## ACKNOWLEDGEMENTS

Knowledge Generation Project: "Motivation. Evaluation and Self-regulation V". PID2022-138175NB-100. Ministry of Science and Innovation. Spain.

Research Center for Inclusive Education. SCIA-ANID CIE160009 Program. Chile.

## REFERENCES

Adarkwah, M. A. (2021). The power of assessment feedback in teaching and learning: a narrative review and synthesis of the literature. *SN Social Science*, *1*(75). 1-44. https://doi.org/10.1007/s43545-021-00086-w

Andrade, H. L. (2023). What Is Next for Rubrics?: A Reflection on Where We Are and Where to Go From Here In C. Gonsalves, & J. Pearson (Eds.), *Improving Learning Through Assessment Rubrics: Student Awareness of What and How*

*They Learn* (pp. 314-326). IGI Global. https://doi.org/10.4018/978-1-6684-6086-3.ch017

Alqassab, M., Strijbos, J. W., & Panadero, E. (2023). A Systematic Review of Peer Assessment Design Elements. *Educational Psychology Review, 35*(18), 1-36. https://doi.org/10.1007/s10648-023-09723-7

Bond, E., Woolcott, G., & Markopoulos, C. (2020). Why aren't teachers using formative assessment? What can be done about it? *Assessment Matters, 14*, 112–136. https://doi.org/10.18296/am.0046

Marrón, G., & Harris, L. (2013). Student self-Assessment. In J. McMillan (Ed.), *SAGE Handbook of Research on Classroom Assessment* (pp. 367-393). SABIO. https://doi.org/10.4135/9781452218649

Centro de estudios MINEDUC. (2022). *Apuntes 22: Variación en la información estadística de los docentes de la educación en desempeño, año 2021*. Centro de Estudios Ministerio de Educación de Chile (MINEDUC). https://bit.ly/3VDJQQd

Cheung, G. W., & Lau, R.S. (2008). Testing Mediation and Suppression Effects of Latent Variables: Bootstrapping with Structural Equation Models. *Organizational Research Methods*, *11*(2), 296–325. https://doi.org/10.1177/1094428107300343

Fan, X. (2003). Using commonly available software for bootstrapping in both substantive and measurement analyses. *Educational and Psychological Measurement*, *63*(1), 24-50. https://doi.org/10.1177/0013164402239315

Greene, B. A., Miller, R. B., Crowson, S. M., Duque, B. L., & Akey, K. L. (2004). Predicting high school students' cognitive engagement and achievement: Contributions of classroom perceptions and motivation. *Contemporary Educational Psychology*, *29*, 462–482. https://doi.org/10.1016/j.cedpsych.2004.01.006

Harris, L., & Brown. G. (2022). *Self-Assessment*. Rutledge. https://doi.org/10.4324/9781138609877-REE1-1

Heritage, M. (2010). *Formative assessment: Making it happen in the classroom*. https://dx.doi.org/10.4135/9781452219493

Hooley, D. S., & Thorpe. J. (2017). The effects of formative reading assessments closely linked to classroom texts on high school reading comprehension. *Educational Technology Research and Development,* (65), 1215–1238. https://doi.org/10.1007/s11423-017-9514-5

Hortigüela, D., Pérez-Pueyo, A., & González-Calvo, G. (2019). Pero… ¿a qué nos referimos realmente con la evaluación formativa y compartida?: Confusiones habituales y reflexiones prácticas. *Revista Iberoamericana de Evaluación Educativa*, *12*(1). https://doi.org/10.15366/riee2019.12.1.001

Illeris, K. (2014). Transformative Learning and Identity. *Journal of Transformative Education*, *12*(2). 148–163. https://doi.org/10.1177/1541344614548423

Klute, M., Apthorp, H., Harlacher, J., & Reale, M. (2017). *Formative assessment and elementary school student academic achievement: A review of the evidence* (REL 2017–259). U.S. Department of Education. Institute of Education Sciences. National Center for Education Evaluation and Regional Assistance. Regional Educational Laboratory Central.

Lipnevich, A., & Panadero. E. (2021). A Review of Feedback Models and Theories: Descriptions. Definitions and Conclusions. *Frontiers in Education*, *6*, Artículo 720195. https://bit.ly/3vH3foO

Lipnevich, A., Berg, A., & Smith, J. (2016). Toward a model of student response to feedback. In G. T. Marrón., & l. R. Harris (Eds.), *Handbook of Human and Social Conditions in Assessment* (pp. 159–185). Rutledge. https://doi.org/10.4324/9781315749136

Lloret-Segura, S., Ferreres-Traver, A., Hernández-Baeza, A., & Tomás-Marco, I. (2014). El análisis factorial exploratorio de los ítems: una guía práctica. Revisada y actualizada. *Anales de Psicología*, *30*(3), 1151-1169. https://dx.doi.org/10.6018/analesps.30.3.199361

López-Pastor, V., & Pérez-Pueyo. A. (2017). *Evaluación formativa y compartida en educación: experiencias de éxito en todas las etapas educativas*. Universidad de León. https://bit.ly/4abtkeZ

Mardia, K. V. (1974). Applications of Some Measures of Multivariate Skewness and Kurtosis in Testing Normality and Robustness Studies. *The Indian Journal of Statistics*, *36*(2), 115-128. https://bit.ly/49eBVfE

Matthews, K., Sherwood, C., Enright. E., & Cook-Sather, R. (2023). What do students and teachers talk about when they talk together about feedback and assessment? Expanding notions of feedback literacy through pedagogical partnership. *Assessment y Evaluation in Higher Education, 49*(1), 26-38. 10.1080/02602938.2023.2170977

Moos, C., & Brookhart. S. (2019). *Advancing In Every Classroom: A Guide For Instructional Leaders*. ASCD. https://bit.ly/4cAfLai

Ossenberg, C., Henderson, A., & Mitchell, M. (2019). What attributes guide best practice for effective feedback? A scoping reviews. *Advances in Health Sciences Education*, (24), 383–401. https://doi.org/10.1007/s10459-018-9854-x

Panadero, E., & Lipnevich. R. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, (35). https://doi.org/10.1016/j.edurev.2021.100416

Panadero, E., Alonso Tapia, J., & Huertas, J. A. (2012). Rubrics and self-assessment scripts effects on self-regulation. Learning and self-efficacy in secondary education. *Learning and Individual Differences*, *22*(6), 806-813. https://doi.org/10.1016/j.lindif.2012.04.007

Panadero, E., Alqassab, M., Fernández-Ruiz, J., & Ocampo, J. (2023) A systematic review on peer assessment: intrapersonal and interpersonal factors. *Assessment y Evaluation in Higher Education*, *8*, 1053-1075. 10.1080/02602938.2023.2164884

Pat-El, R. J., Tillema, H., Segers, M., & Vedder, P. (2013). Validation of Assessment for Learning Questionnaires for teachers and students. *The British journal of educational psychology*, *83*(1), 98-113. https://doi.org/10.1111/j.2044-8279.2011.02057.x

Popham., J. (2013). *Evaluación Transformativa: el poder transformador de la evaluación formativa*. Narcea.

Pozo, J. I. (2017). Aprender más allá del cuerpo: de las representaciones encarnadas a la explicitación mediada por representaciones externas. *Infancia y Aprendizaje*, *40*(2), 219-276. 10.1080/02103702.2017.1306942

Rowe, A. D. (2017). Feelings about feedback: the role of emotions in assessment for learning. In D., Carless, S. Bridges, C. Chan, & R. Glofcheski (Eds), *Scaling up assessment for learning in higher education. The enabling power of assessment (vol 5*). Springer.https://doi.org/10.1007/978-981-10-3045-1_11

Ruiz, M., Pardo, A., & San Martín. R. (2010). Modelos de ecuaciones estructurales. *Papeles del Psicólogo, 31*(1), 34-35. https://bit.ly/3TANiIL

Sánchez, C. E., Atkinson, K. M., Koenka, A. C., Moshontz, H., & Cooper, H. (2017). Self-grading and peer-grading for formative and summative assessments in 3rd through 12th grade classrooms: A meta-analysis. *Journal of Educational Psychology, 109*(8), 1049-1066. https://doi.org/10.1037/edu0000190

Shavelson, R. J, Young, D. B, Ayala, C. C., Brandon, P. R., Furtak, E. M., Ruiz-Primo, M. A., …, & Yin, Y. (2008). On the impact of curriculum-embedded formative assessment on learning: A collaboration between curriculum and assessment developers. *Applied Measurement in Education*, *21*(4), 295–314. https://doi.org/10.1080/08957340802347647

Tay H. Y. & Lam, K. W. L. (2022). Students' engagement across a typology of teacher feedback practices. *Educational Research for Policy and Practice*, (21), 427–445. https://doi.org/10.1007/s10671-022-09315-2

van der Kleij, F. M., & DeLuca. C. (2023). Implementation of assessment for learning. In R. Tierney, F. Rizvi, & K. Ercikan (Eds.), *International Encyclopedia of Education* (pp. 147-154). https://doi.org/10.1016/B978-0-12-818630-5.09028-X

Van der Kleij, F. M., Cumming, J., & Looney, A. (2018). Policy expectations and support for teacher formative assessment in Australian education reform. *Assessment in Education: Principles. Policy y Practice*, *25*(6), 620-637. 10.1080/0969594X.2017.1374924

Vygotsky, L. S. (1979). Interacción entre aprendizaje y desarrollo. In M., Cole, J. Vera, S. Scribner, & E. Souberman (Eds.), *El desarrollo de los procesos psicológicos superiores* (pp. 123-140). Crítica-Grijalbo.

Wulandari, D., Sutrisno, S., & Nirwana, M. B. (2021). Mardia's Skewness and Kurtosis for Assessing Normality Assumption in Multivariate Regression. *Enthusiastic: International Journal of Applied Statistics and Data Science*, *1*(1), 1 – 6. https://doi.org/10.20885/enthusiastic.vol1.iss1.art1